Abel Symposia 13

Elena Celledoni
Giulia Di Nunno
Kurusch Ebrahimi-Fard
Hans Zanna Munthe-Kaas *Editors*

Computation and Combinatorics in Dynamics, Stochastics and Control

The Abel Symposium, Rosendal, Norway, August 2016

Springer

ABEL SYMPOSIA

Edited by the Norwegian Mathematical Society

More information about this series at http://www.springer.com/series/7462

Participants in 2016 Abel Symposium. (Photo: H. Munthe-Kaas)

Elena Celledoni • Giulia Di Nunno •
Kurusch Ebrahimi-Fard • Hans Zanna Munthe-Kaas
Editors

Computation and Combinatorics in Dynamics, Stochastics and Control

The Abel Symposium, Rosendal, Norway, August 2016

ABEL
PRISEN

Springer

Editors
Elena Celledoni
Department of Mathematical Sciences
Norwegian University of Science
and Technology
Trondheim, Norway

Giulia Di Nunno
Department of Mathematics
University of Oslo
Oslo, Norway

Kurusch Ebrahimi-Fard
Department of Mathematical Sciences
Norwegian University of Science
and Technology
Trondheim, Norway

Hans Zanna Munthe-Kaas
Department of Mathematics
University of Bergen
Bergen, Norway

ISSN 2193-2808 ISSN 2197-8549 (electronic)
Abel Symposia
ISBN 978-3-030-01592-3 ISBN 978-3-030-01593-0 (eBook)
https://doi.org/10.1007/978-3-030-01593-0

Library of Congress Control Number: 2018966592

Mathematics Subject Classification (2010): 15A52, 16W30, 17D25, 35R60, 37E20, 60H15, 76M35, 93C10

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The Norwegian government established the Abel Prize in mathematics in 2002, and the first prize was awarded in 2003. In addition to honoring the great Norwegian mathematician Niels Henrik Abel by awarding an international prize for outstanding scientific work in the field of mathematics, the prize shall contribute toward raising the status of mathematics in society and stimulate the interest for science among school children and students. In keeping with this objective, the Niels Henrik Abel Board has decided to finance annual Abel Symposia. The topic of the symposia may be selected broadly in the area of pure and applied mathematics. The symposia should be at the highest international level and serve to build bridges between the national and international research communities. The Norwegian Mathematical Society is responsible for the events. It has also been decided that the contributions from these symposia should be presented in a series of proceedings, and Springer Verlag has enthusiastically agreed to publish the series. The Niels Henrik Abel Board is confident that the series will be a valuable contribution to the mathematical literature.

Chair of the Niels Henrik Abel Board — Kristian Ranestad

Preface

In recent years we have witnessed a remarkable convergence between individual mathematical disciplines that approach deterministic and stochastic dynamical systems from mathematical analysis, computational mathematics, and control theoretical perspectives. One of the prime examples is the theory of *rough paths*, pioneered by Terry Lyons (Oxford). Massimiliano Gubinelli (Paris/Bonn) subsequently developed the notions of controlled and branched rough paths. This line of work culminated in the 2014 Fields Medal being awarded to Martin Hairer (Warwick/London) for his far-reaching work on *regularity structures*, which led him to breakthrough discoveries in the theory of stochastic partial differential equations.

Rough paths theory has strong connections to the analysis of *geometric integration algorithms* for deterministic flows, where the need to understand structure preservation has led to the development of new analytical tools based on modern algebra and combinatorics. Recent developments in these fields provide a common mathematical framework for attacking many different problems related to differential geometry, analysis and algorithms for stochastic and deterministic dynamics.

In the Abel Symposium 2016 (August 16–19), leading researchers in the fields of deterministic and stochastic differential equations, numerical analysis, control theory, algebra, and random processes met at the picturesque Barony in Rosendal near Bergen for a lively exchange of research ideas and presentation of the current state of the art in these fields. The current Abel Symposia volume may serve as a point of departure for exploring these related but diverse fields of research, as well as an indicator of important current and future developments in modern mathematics.

Trondheim, Norway — Elena Celledoni
Oslo, Norway — Giulia Di Nunno
Trondheim, Norway — Kurusch Ebrahimi-Fard
Bergen, Norway — Hans Zanna Munthe-Kaas

Contents

Facilitated Exclusion Process

Jinho Baik, Guillaume Barraquand, Ivan Corwin, and Toufic Suidan

Abstract We study the Facilitated TASEP, an interacting particle system on the one dimensional integer lattice. We prove that starting from step initial condition, the position of the rightmost particle has Tracy Widom GSE statistics on a cube root time scale, while the statistics in the bulk of the rarefaction fan are GUE. This uses a mapping with last-passage percolation in a half-quadrant which is exactly solvable through Pfaffian Schur processes. Our results further probe the question of how first particles fluctuate for exclusion processes with downward jump discontinuities in their limiting density profiles. Through the Facilitated TASEP and a previously studied MADM exclusion process we deduce that cube-root time fluctuations seem to be a common feature of such systems. However, the statistics which arise are shown to be model dependent (here they are GSE, whereas for the MADM exclusion process they are GUE). We also discuss a two-dimensional crossover between GUE, GOE and GSE distribution by studying the multipoint distribution of the first particles when the rate of the first one varies. In terms of half-space last passage percolation, this corresponds to last passage times close to the boundary when the size of the boundary weights is simultaneously scaled close to the critical point.

1 Introduction

Exclusion processes on $\mathbb{Z}$ are expected, under mild hypotheses, to belong to the KPZ universality class [6, 11]. As a consequence, one expects that if particles start densely packed from the negative integers – the step initial condition – the positions

J. Baik
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
e-mail: baik@umich.edu

G. Barraquand · I. Corwin (✉)
Department of Mathematics, Columbia University, New York, NY, USA
e-mail: barraquand@math.columbia.edu

T. Suidan

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_1

of particles in the bulk of the rarefaction fan will fluctuate on a cube-root time scale with GUE Tracy-Widom statistics in the large time limit. The motivation for this paper is to consider the fluctuations of the location of the rightmost particle and probe its universality over different exclusion processes.

In the totally asymmetric simple exclusion process (TASEP) the first particle jumps by 1 after an exponentially distributed waiting time of mean 1, independently of everything else. Hence its location satisfies a classical Central Limit Theorem when time goes to infinity (i.e. square-root time fluctuation with limiting Gaussian statistics). This is true for any totally asymmetric exclusion process starting from step initial condition. However, in the asymmetric simple exclusion process (ASEP), the trajectory of the first particle is affected by the behaviour of the next particles. This results in a different limit theorem. Tracy and Widom showed [19, Theorem 2] that the fluctuations still occur on the $t^{1/2}$ scale, but the limiting distribution is different and depends on the strength of the asymmetry (see also [13] where the same distribution arises for the first particle's position in a certain zero-range process). In [3], another partially asymmetric process called the MADM exclusion process was studied. The first particle there fluctuates on a $t^{1/3}$ scale with Tracy-Widom GUE limit distribution, as if it was in the bulk of the rarefaction fan. An explanation for why the situation is so contrasted with ASEP (and other model where the first particle has the same limit behaviour) is that the MADM, when started from step initial condition, develops a downward jump discontinuity of its density profile around the first particle (see Figure 3 in [3]).

In this paper, we test the universality of the fluctuations of the first particle in the presence of a jump discontinuity – does the $t^{1/3}$ scale and GUE statistics survive over other models? We solve this question for the Facilitated TASEP. Our results show that the GUE distribution does not seem to survive in general, though we do still see the $t^{1/3}$ scale.

1.1 The Facilitated TASEP

The Facilitated Totally Asymmetric Simple Exclusion Process (abbreviated FTASEP in the following) was introduced in [4] and further studied in [9, 10]. This is an interacting particle system on $\mathbb{Z}$, satisfying the exclusion rule, which means that each site is occupied by at most 1 particle. A particle sitting at site x jumps to the right by 1 after an exponentially distributed waiting time of mean 1, provided that the target site (i.e. $x + 1$) is empty and that the site $x - 1$ is occupied. Informally, the dynamics are very similar with TASEP, with the only modification being particles need to wait to have a left neighbour (facilitation) before moving (See Fig. 1). It was introduced as a simplistic model for motion in glasses: particles move faster in less crowded areas (modelled by the exclusion rule), but need a stimulus to move (modelled by the facilitation rule). We focus here on the step initial condition: at time 0, the particles occupy all negative sites, and the non-negative sites are empty.

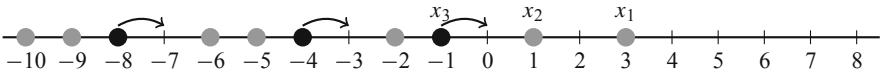


Fig. 1 The particles in black jump by 1 at rate 1 whereas particles in gray cannot

Since the dynamics preserve the order between particles, we can describe the configuration of the system by their ordered positions

$$\cdots < x_2 < x_1 < \infty.$$

Let us collect some (physics) results from [9] which studies the hydrodynamic behaviour – but not the fluctuations. Assume that the system is at equilibrium with an average density of particles ρ. A family of translation invariant stationary measures indexed by the average density – conjecturally unique – is described in the end of Sect. 3.1. Then the flux, i.e. the average number of particles crossing a given bond per unit of time, is given by (see [9, Eq. (3)] and (8) in the present paper)

$$j(\rho) = \frac{(1-\rho)(2\rho-1)}{\rho}. \tag{1}$$

This is only valid when $\rho > 1/2$. When $\rho < 1/2$, [9] argues that the system eventually reaches a static state that consists of immobile single-particle clusters. One expects that the limiting density profile, informally given by

$$\rho(x,t) := \lim_{T\to\infty} \mathbb{P}\big(\exists \text{ particle at site } xT \text{ at time } tT\big),$$

exists and is a weak solution subject to the entropy condition of the conservation equation

$$\frac{\partial}{\partial t}\rho(x,t) + \frac{\partial}{\partial x} j(\rho(x,t)) = 0. \tag{2}$$

Solving this equation subject to the initial condition $\rho(x,t) = \mathbb{1}_{\{x<0\}}$ yields the density profile (depicted in Fig. 2)

$$\forall t > 0,\ \rho(xt,t) = \begin{cases} 1 & \text{if } x < -1, \\ \frac{1}{\sqrt{2+x}} & \text{if } -1 \leqslant x \leqslant 1/4 \\ 0 & \text{if } x > 1/4. \end{cases}$$

See also [9, Eq. (5)]. It is clear that there must be a jump discontinuity in the macroscopic density profile since in FTASEP particles can travel only in regions where the density is larger than $1/2$.

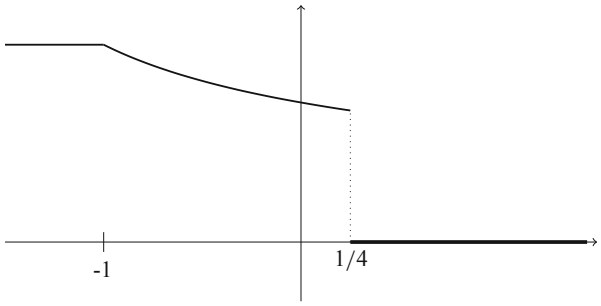


Fig. 2 Limiting density profile, i.e. graph of the function $x \mapsto \rho(x, 1)$

In general, the property of the flux which is responsible for the jump discontinuity is the fact that $j(\rho)/\rho$, i.e. the drift of a tagged particle, is not decreasing as a function of ρ. The density around the first particle will be precisely the value ρ_0 that maximizes the drift. Let us explain why. On one hand, the characteristics of PDEs such as (2) are straight lines ([7, §3.3.1.]), which means in our case that for any density $\bar{\rho}$ occurring in the rarefaction fan in the limit profile, there exists a constant $\pi(\bar{\rho})$ such that

$$\rho\big(\pi(\bar{\rho})t, t\big) = \bar{\rho}. \tag{3}$$

$\pi(\bar{\rho})$ is the macroscopic position of particles around which the density is $\bar{\rho}$. Differentiating (3) with respect to t and using the conservation equation (2) yields $\pi(\rho) = \frac{\partial j(\rho)}{\partial \rho}$. If we call ρ_0 the density around the first particle, then the macroscopic position of the first particle should be $\pi(\rho_0)$. On the other hand, the first particle has a constant drift, which is[1] $j(\rho_0)/\rho_0$. Combining these observations yields

$$\left.\frac{\partial j(\rho)}{\partial \rho}\right|_{\rho=\rho_0} = \frac{j(\rho_0)}{\rho_0} \quad \text{i.e.} \quad \left.\frac{\mathrm{d}}{\mathrm{d}\rho}\frac{j(\rho)}{\rho}\right|_{\rho=\rho_0} = 0.$$

This implies that a discontinuity of the density profile at the first particle can occur only if the drift is not strictly decreasing as a function of ρ, and it suggests that ρ_0 is indeed the maximizer of the drift (see also [3, Section 4] for a different justification). In the example of the FTASEP, the maximum of

$$\frac{j(\rho)}{\rho} = \frac{(1-\rho)(2\rho-1)}{\rho^2}$$

[1] Assuming local equilibrium – which is not expected to be satisfied around the first particle but close to it– the drift is given by $j(\rho)/\rho$ when the density is ρ.

is such that $\rho_0 = 2/3$ and $\pi(\rho_0) = 1/4$. In particular, this means that $x_1(t)/t$ should converge to $1/4$ when t goes to infinity.

The fluctuations of $x_1(t)$ around $t/4$ are not GUE distributed as for the MADM exclusion process [3, Theorem 1.3], but rather follow the GSE Tracy-Widom distribution in the large time limit.

Theorem 1 *For FTASEP with step initial data,*

$$\mathbb{P}\left(\frac{x_1(t) - \frac{t}{4}}{2^{-4/3}t^{1/3}} \geqslant x\right) \xrightarrow[t\to\infty]{} F_{\mathrm{GSE}}(-x),$$

where the GSE Tracy-Widom distribution function F_{GSE} is defined in Definition 7.

In the bulk of the rarefaction fan, however, the locations of particles fluctuate as the KPZ scaling theory predicts [12, 18].

Theorem 2 *For FTASEP with step initial data, and for any $r \in (0, 1)$,*

$$\mathbb{P}\left(\frac{x_{\lfloor rt\rfloor}(t) - t\frac{1-6r+r^2}{4}}{\varsigma t^{1/3}} \geqslant x\right) \xrightarrow[t\to\infty]{} F_{\mathrm{GUE}}(-x),$$

where $\varsigma = 2^{-4/3}\frac{(1+r)^{5/3}}{(1-r)^{1/3}}$ and the GUE Tracy-Widom distribution function F_{GUE} is defined in Sect. 4.

We now consider a slightly more general process depending on a parameter $\alpha > 0$ that we denote FTASEP(α), where the first particle jumps at rate α instead of 1. We already know the nature of fluctuations of $x_1(t)$ when $\alpha = 1$. It is natural to expect that fluctuations are still GSE Tracy-Widom distributed on the $t^{1/3}$ scale for $\alpha > 1$. However, if α is very small, one expects that the first particle jumps according to a Poisson point process with intensity α and thus $x_1(t)$ has Gaussian fluctuations on the $t^{1/2}$ scale. It turns out that the threshold between these regimes happen when $\alpha = 1/2$.

Theorem 3 *Let $\mathbf{x}(t) = \{x_n(t)\}_{n\geqslant 1}$ be the particles positions in the FTASEP(α) started from step initial condition, when the first particle jumps at rate α. Then,*

1. *For $\alpha > 1/2$,*

$$\mathbb{P}\left(\frac{x_1(t) - \frac{t}{4}}{2^{-4/3}t^{1/3}} \geqslant x\right) \xrightarrow[t\to\infty]{} F_{\mathrm{GSE}}(-x).$$

2. *For $\alpha = 1/2$,*

$$\mathbb{P}\left(\frac{x_1(t) - \frac{t}{4}}{2^{-4/3}t^{1/3}} \geqslant x\right) \xrightarrow[t\to\infty]{} F_{\mathrm{GOE}}(-x).$$

3. *For* $\alpha < 1/2$,

$$\mathbb{P}\left(\frac{x_1(t) - t\alpha(1-\alpha)}{\varsigma t^{1/2}} \geqslant x\right) \xrightarrow[t\to\infty]{} G(-x),$$

where $G(x)$ *is the standard Gaussian distribution function and* $\varsigma = \frac{1-2\alpha}{\sqrt{\alpha(1-\alpha)}}$.

It is also possible to characterize the joint distribution of several particles. An interesting case arises when we scale α close to the critical point and we look at particles indexed by $\eta t^{2/3}$ for different values of $\eta \geqslant 0$. More precisely, we scale

$$\alpha = \frac{1 + 2^{4/3}\varpi\tau^{-1/3}}{2},$$

where $\varpi \in \mathbb{R}$ is a free parameter and for any $\eta \geqslant 0$ consider the rescaled particle position at time t

$$X_t(\eta) := \frac{x_{2^{1/3}\eta t^{2/3}}(t) - \frac{t}{4} + \eta\rho_0^{-1}2^{1/3}t^{2/3} - \eta^2 2^{-4/3}}{t^{1/3}2^{-4/3}}, \tag{4}$$

where $\rho_0 = 2/3$ (This is the density near the first particles in FTASEP(1)).

Theorem 4 *For any* $p_1, \ldots, p_k \in \mathbb{R}$, *and* $0 \leqslant \eta_1 < \cdots < \eta_k$

$$\lim_{t\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{k} \{X_t(\eta_i) \geqslant p_i\}\right) = \mathrm{Pf}(J - \mathsf{K}^{\mathrm{cross}})_{\mathbb{L}^2(\mathbb{D}_k(-p_1,\ldots,-p_k))},$$

where the right hand side is the Fredholm Pfaffian (see Definition 5*) of some kernel* $\mathsf{K}^{\mathrm{cross}}$ *(depending on* ϖ *and the* η_i*) introduced in* [1, *Section* 2.5] *(see also Sect.* 5 *of the present paper) on the domain* $\mathbb{D}_k(-p_1, \ldots, -p_k)$ *where*

$$\mathbb{D}_k(x_1, \ldots, x_k) = \{(i, x) \in \{1, \ldots, k\} \times \mathbb{R} : x \geqslant x_i\}.$$

For the FTASEP, that is when $\alpha = 1$ we have

Theorem 5 *For any* $p_1, \ldots, p_k \in \mathbb{R}$, *and* $0 < \eta_1 < \cdots < \eta_k$

$$\lim_{t\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{k} \{X_t(\eta_i) \geqslant p_i\}\right) = \mathrm{Pf}(J - \mathsf{K}^{\mathrm{SU}})_{\mathbb{L}^2(\mathbb{D}_k(-p_1,\ldots,-p_k))},$$

where the right hand side is the Fredholm Pfaffian of some kernel K^{SU} *(depending on the* η_i*) introduced in* [1, *Section* 2.5] *(see also Sect.* 5*).*

1.2 Half-Space Last Passage Percolation

Our route to prove Theorems 1, 2, 3, 4 and 5 in Sect. 3 uses a mapping with Last Passage Percolation (LPP) on a half-quadrant.

Definition 1 (Half-space exponential weight LPP) Let $\left(w_{n,m}\right)_{n\geqslant m\geqslant 0}$ be a sequence of i.i.d. exponential random variables with rate 1 (see Definition 2) when $n \geqslant m+1$ and with rate α when $n = m$. We define the exponential last passage percolation time on the half-quadrant, denoted $H(n,m)$, by the recurrence for $n \geqslant m$,

$$H(n,m) = w_{n,m} + \begin{cases} \max\left\{H(n-1,m); H(n,m-1)\right\} & \text{if } n \geqslant m+1, \\ H(n,m-1) & \text{if } n = m \end{cases}$$

with the boundary condition $H(n,0) = 0$.

We show in Proposition 2 that FTASEP is equivalent to a TASEP on the positive integers with a source of particles at the origin. We call the latter model half-line TASEP. The mapping between the two processes is the following: we match the gaps between consecutive particles in the FTASEP with the occupation variables in the half-line TASEP. Otherwise said, we study how the holes travel to the left in the FTASEP and prove that if one shrinks all distances between consecutive holes by one, the dynamics of holes follow those of the half-line TASEP (see the proof of Proposition 2, in particular Fig. 6). In the case of full-space TASEP it is well-known that the height function of TASEP has the same law as the border of the percolation cluster of the LPP model with exponential weights (in a quadrant). This mapping remains true for half-line TASEP and LPP on the half-quadrant (Lemma 2, see Fig. 3).

The advantage of this mapping between FTASEP and half-space last-passage percolation is that we can now use limit theorems proved for the latter (see [1] and references therein), which we recall below.

Theorem 6 ([1, Theorem 1.4]) *The last passage time on the diagonal $H(n,n)$ satisfies the following limit theorems, depending on the rate α of the weights on the diagonal.*

1. *For $\alpha > 1/2$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H(n,n) - 4n}{2^{4/3}n^{1/3}} < x\right) = F_{\mathrm{GSE}}(x)\,.$$

2. *For $\alpha = 1/2$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H(n,n) - 4n}{2^{4/3}n^{1/3}} < x\right) = F_{\mathrm{GOE}}(x)\,,$$

where the GOE Tracy-Widom distribution function F_{GOE} is defined in Lemma 6.

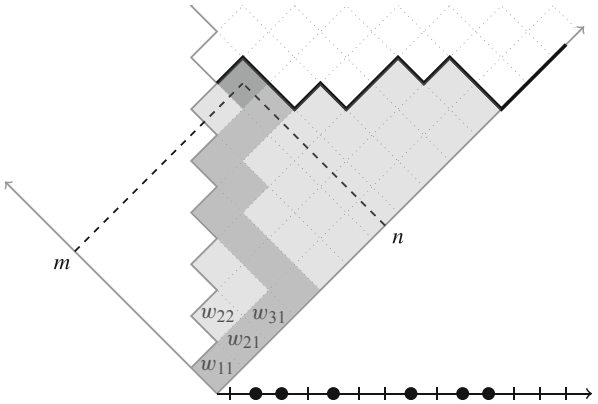


Fig. 3 LPP on the half-quadrant. One admissible path from $(1, 1)$ to (n, m) is shown in dark gray. $H(n, m)$ is the maximum over such paths of the sum of the weights w_{ij} along the path. The light gray area corresponds to the percolation cluster at some fixed time, and its border (shown in black) is associated with the particle system depicted on the horizontal line

3. *For* $\alpha < 1/2$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H(n,n) - \frac{n}{\alpha(1-\alpha)}}{\sigma n^{1/2}} < x\right) = G(x),$$

where $G(x)$ *is the probability distribution function of the standard Gaussian, and*

$$\sigma^2 = \frac{1-2\alpha}{\alpha^2(1-\alpha)^2}.$$

Away from the diagonal, the limit theorem satisfied by $H(n, m)$ happens to be exactly the same as in the unsymmetrized or full-space model.

Theorem 7 ([1, Theorem 1.5]) *For any* $\kappa \in (0, 1)$ *and* $\alpha > \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$, *we have that when* $m = \kappa n + sn^{2/3-\epsilon}$, *for any* $s \in \mathbb{R}$ *and* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H(n,m) - (1+\sqrt{\kappa})^2 n}{\sigma n^{1/3}} < x\right) = F_{\mathrm{GUE}}(x),$$

where

$$\sigma = \frac{(1+\sqrt{\kappa})^{4/3}}{\sqrt{\kappa}^{1/3}}.$$

In [1], we also explained how to obtain a two dimensional crossover between all the above cases by tuning the parameters α and κ close to their critical value in the scale $n^{-1/3}$ (see Fig. 4). The proofs of the following results were omitted in [1] (they were stated as Theorem 1.8 and 1.9 in [1]) and we include them in Sect. 5. Let us define

$$H_n(\eta) = \frac{H\big(n + n^{2/3}\xi\eta, n - n^{2/3}\xi\eta\big) - 4n + n^{1/3}\xi^2\eta^2}{\sigma n^{1/3}},$$

where $\eta \geqslant 0$, $\sigma = 2^{4/3}$ and $\xi = 2^{2/3}$. We scale α as

$$\alpha = \frac{1 + 2\sigma^{-1}\varpi n^{-1/3}}{2}$$

where $\varpi \in \mathbb{R}$ is a free parameter.

Theorem 8 *For $0 \leqslant \eta_1 < \cdots < \eta_k$, $\varpi \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{k} \big\{H_n(\eta_i) < h_i\big\}\right) = \mathrm{Pf}\big(J - \mathsf{K}^{\mathrm{cross}}\big)_{\mathbb{L}^2(\mathbb{D}_k(h_1,\ldots,h_k))},$$

where $\mathsf{K}^{\mathrm{cross}}$ is defined in Sect. 5.

We refer to [1, Sections 1.5 and 2.5] for comments and explanations about this kernel and its various degenerations. The phase diagram of one-point fluctuations is represented on Fig. 4.

In the case when $\alpha > 1/2$ is fixed, the joint distribution of passage-times is governed by the so-called symplectic-unitary transition [8].

Theorem 9 *For $\alpha > 1/2$ and $0 < \eta_1 < \cdots < \eta_k$, we have that*

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{k} \big\{H_n(\eta_i) < h_i\big\}\right) = \mathrm{Pf}\big(J - \mathsf{K}^{\mathrm{SU}}\big)_{\mathbb{L}^2(\mathbb{D}_k(h_1,\ldots,h_k))},$$

where K^{SU} is a certain matrix kernel introduced in [1] *(See also Sect. 5).*

Theorem 9 corresponds to the $\varpi \to +\infty$ degeneration of Theorem 8.

Outline of the Paper

In Sect. 2, we provide the precise definitions of all probability distributions arising in this paper. In Sect. 3, we explain the mapping between FTASEP and TASEP on a half-space with a source, or equivalently exponential LPP on a half-space. We prove the limit theorems for the fluctuations of particles positions in FTASEP(α) using the

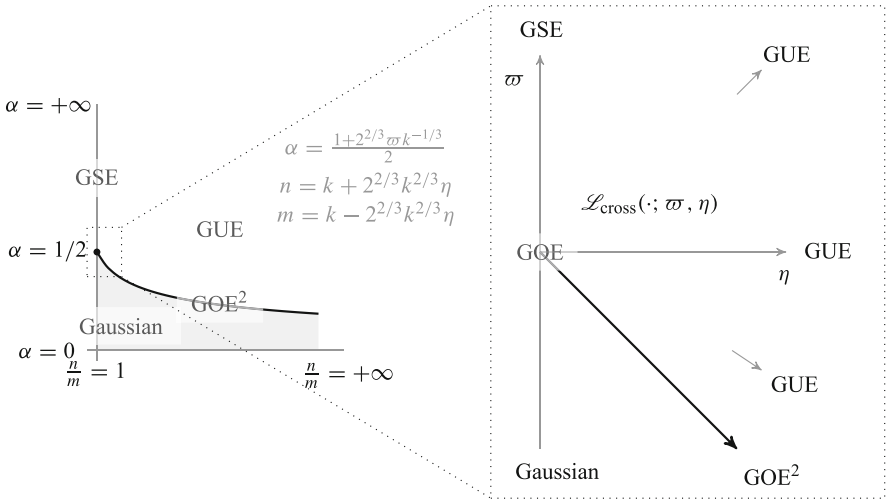


Fig. 4 Phase diagram of the fluctuations of $H(n, m)$ as $n \to \infty$ when α and the ratio n/m varies. The gray area corresponds to a region of the parameter space where the fluctuations are on the scale $n^{1/2}$ and Gaussian. The bounding GOE^2 curve asymptotes to zero as n/m goes to $+\infty$. The crossover distribution $\mathscr{L}_{\mathrm{cross}}(\cdot; \varpi, \eta)$ is defined in [1, Definition 2.9] and describes the fluctuations in the vicinity of $n/m = 1$ and $\alpha = 1/2$

asymptotic results for half-space LPP. In Sect. 4, we recall the k-point distribution along space-like paths in half-space LPP with exponential weights (Proposition 3), derived in [1]. In Sect. 5, we provide a rigorous derivation of Theorem 8 and 9 from Proposition 3. This boils down to an asymptotic analysis of the correlation kernel that was omitted in [1].

2 Definitions of Distribution Functions

In this section, we provide definitions of the probability distributions arising in the paper.

Definition 2 The exponential distribution with rate $\alpha \in (0, +\infty)$, denoted $\mathscr{E}(\alpha)$, is the probability distribution on $\mathbb{R}_{>0}$ such that if $X \sim \mathscr{E}(\alpha)$,

$$\forall x \in \mathbb{R}_{>0},\ \mathbb{P}(X > x) = e^{-\alpha x}.$$

Let us introduce a convenient notation that we use throughout the paper to specify integration contours in the complex plane.

Definition 3 Let $\mathscr{C}_a^\varphi$ be the union of two semi-infinite rays departing $a \in \mathbb{C}$ with angles φ and $-\varphi$. We assume that the contour is oriented from $a + \infty e^{-i\varphi}$ to $a + \infty e^{+i\varphi}$.

We recall that for an integral operator $\mathscr{K}$ defined by a kernel $\mathsf{K} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, its Fredholm determinant $\det(I + \mathscr{K})_{\mathbb{L}^2(\mathbb{X},\mu)}$ is given by the series expansion

$$\det(I + \mathscr{K})_{\mathbb{L}^2(\mathbb{X},\mu)} = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \int_{\mathbb{X}} \dots \int_{\mathbb{X}} \det\left(\mathsf{K}(x_i, x_j)\right)_{i,j=1}^{k} \mathrm{d}\mu^{\otimes k}(x_1 \dots x_k),$$

whenever it converges. Note that we will omit the measure μ in the notations and write simply $\mathbb{L}^2(\mathbb{X})$ when the uniform or the Lebesgue measure is considered. With a slight abuse of notations, we will also write $\det(I + \mathsf{K})_{\mathbb{L}^2(\mathbb{X})}$ instead of $\det(I + \mathscr{K})_{\mathbb{L}^2(\mathbb{X})}$.

Definition 4 The GUE Tracy-Widom distribution, denoted $\mathscr{L}_{\mathrm{GUE}}$ is a probability distribution on $\mathbb{R}$ such that if $X \sim \mathscr{L}_{\mathrm{GUE}}$,

$$\mathbb{P}(X \leqslant x) = F_{\mathrm{GUE}}(x) = \det(I - \mathsf{K}_{\mathrm{Ai}})_{\mathbb{L}^2(x,+\infty)}$$

where K_{Ai} is the Airy kernel,

$$\mathsf{K}_{\mathrm{Ai}}(u, v) = \int_{\mathscr{C}_{-1}^{2\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{e^{z^3/3 - zu}}{e^{w^3/3 - wv}} \frac{1}{z - w}. \tag{5}$$

In order to define the GOE and GSE distribution in a form which is convenient for later purposes, we introduce the concept of Fredholm Pfaffian.

Definition 5 ([16, Section 8]) For a 2×2-matrix valued skew-symmetric kernel,

$$\mathsf{K}(x, y) = \begin{pmatrix} \mathsf{K}_{11}(x, y) & \mathsf{K}_{12}(x, y) \\ \mathsf{K}_{21}(x, y) & \mathsf{K}_{22}(x, y) \end{pmatrix}, \quad x, y \in \mathbb{X},$$

we define its Fredholm Pfaffian by the series expansion

$$\mathrm{Pf}\big(J + \mathsf{K}\big)_{\mathbb{L}^2(\mathbb{X},\mu)} = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \int_{\mathbb{X}} \dots \int_{\mathbb{X}} \mathrm{Pf}\Big(K(x_i, x_j)\Big)_{i,j=1}^{k} \mathrm{d}\mu^{\otimes k}(x_1 \dots x_k), \tag{6}$$

provided the series converges, and we recall that for an skew-symmetric $2k \times 2k$ matrix A, its Pfaffian is defined by

$$\mathrm{Pf}(A) = \frac{1}{2^k k!} \sum_{\sigma \in \mathscr{S}_{2k}} \mathrm{sign}(\sigma) a_{\sigma(1)\sigma(2)} a_{\sigma(3)\sigma(4)} \dots a_{\sigma(2k-1)\sigma(2k)}. \tag{7}$$

The kernel J is defined by

$$J(x,y) = \begin{cases} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

In Sect. 5, we will need to control the convergence of Fredholm Pfaffian series expansions. This can be done using Hadamard's bound.

Lemma 1 ([1, Lemma 2.5]) *Let $\mathsf{K}(x,y)$ be a 2×2 matrix valued skew symmetric kernel. Assume that there exist constants $C>0$ and constants $a>b\geqslant 0$ such that*

$$|\mathsf{K}_{11}(x,y)| < Ce^{-ax-ay}, \quad |\mathsf{K}_{12}(x,y)| = |\mathsf{K}_{21}(y,x)| < Ce^{-ax+by}, \quad |\mathsf{K}_{22}(x,y)| < Ce^{bx+by}.$$

Then, for all $k\in\mathbb{Z}_{>0}$,

$$\left|\mathrm{Pf}\big[\mathsf{K}(x_i,x_j)\big]_{i,j=1}^{k}\right| < (2k)^{k/2}C^k\prod_{i=1}^{k}e^{-(a-b)x_i}.$$

Definition 6 The GOE Tracy-Widom distribution, denoted $\mathscr{L}_{\mathrm{GOE}}$, is a continuous probability distribution on $\mathbb{R}$ whose cumulative distribution function $F_{\mathrm{GOE}}(x)$ (i.e. $\mathbb{P}(X\leqslant x)$ where $X\sim\mathscr{L}_{\mathrm{GOE}}$) is given by

$$F_{\mathrm{GOE}}(x) = \mathrm{Pf}\big(J-\mathsf{K}^{\mathrm{GOE}}\big)_{\mathbb{L}^2(x,\infty)},$$

where $\mathsf{K}^{\mathrm{GOE}}$ is the 2×2 matrix valued kernel defined by

$$\mathsf{K}_{11}^{\mathrm{GOE}}(x,y) = \int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{z-w}{z+w}e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{K}_{12}^{\mathrm{GOE}}(x,y) = -\mathsf{K}_{21}^{\mathrm{GOE}}(x,y) = \int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_{-1/2}^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{w-z}{2w(z+w)}e^{z^3/3+w^3/3-xz-yw},$$

$$\begin{aligned}\mathsf{K}_{22}^{\mathrm{GOE}}(x,y) &= \int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{z-w}{4zw(z+w)}e^{z^3/3+w^3/3-xz-yw} \\ &\quad+\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\frac{e^{z^3/3-zx}}{4z}-\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\frac{e^{z^3/3-zy}}{4z}-\frac{\mathrm{sgn}(x-y)}{4},\end{aligned}$$

where $\mathrm{sgn}(x)=\mathbb{1}_{x>0}-\mathbb{1}_{x<0}$.

Definition 7 The GSE Tracy-Widom distribution, denoted $\mathscr{L}_{\mathrm{GSE}}$, is a continuous probability distribution on $\mathbb{R}$ whose cumulative distribution function F_{GOE} is given by

$$F_{\mathrm{GSE}}(x) = \mathrm{Pf}\big(J - \mathsf{K}^{\mathrm{GSE}}\big)_{\mathbb{L}^2(x,\infty)},$$

where $\mathsf{K}^{\mathrm{GSE}}$ is a 2×2-matrix valued kernel defined by

$$\mathsf{K}_{11}^{\mathrm{GSE}}(x, y) = \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{4zw(z+w)} e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{K}_{12}^{\mathrm{GSE}}(x, y) = -\mathsf{K}_{21}^{\mathrm{GSE}}(x, y) = \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{4z(z+w)} e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{K}_{22}^{\mathrm{GSE}}(x, y) = \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{4(z+w)} e^{z^3/3+w^3/3-xz-yw} \mathrm{d}z\mathrm{d}w.$$

3 Facilitated Totally Asymmetric Simple Exclusion Process

3.1 Definition and Coupling

A configuration of particles on a subset X of $\mathbb{Z}$ can be described either by occupation variables, i.e. a collection $\boldsymbol{\eta} = (\eta_x)_{x\in X}$ where $\eta_x = 1$ if the site x is occupied and $\eta_x = 0$ else, or a vector of particle positions $\mathbf{x} = (x_i)_{i\in I}$ where the particles are indexed by some set I. We will use both notations.

Definition 8 The FTASEP is a continuous-time Markov process defined on the state space $\{0, 1\}^{\mathbb{Z}}$ via its Markov generator, acting on local functions $f : \{0, 1\}^{\mathbb{Z}} \to \mathbb{R}$ by

$$Lf(\boldsymbol{\eta}) = \sum_{x\in\mathbb{Z}} \eta_{x-1}\eta_x(1-\eta_{x+1})\big(f(\boldsymbol{\eta}_{x,x+1}) - f(\boldsymbol{\eta})\big),$$

where the state $\boldsymbol{\eta}_{x,x+1}$ is obtained from $\boldsymbol{\eta}$ by exchanging occupation variables at sites x and $x + 1$.

That this generator defines a Markov process corresponding to the particle dynamics described in the introduction can be justified, for instance, by checking the conditions of [14, Theorem 3.9].

We will be mostly interested in initial configurations that are right-finite, which means that there exists a right-most particle. Since the dynamics preserves the order between particles, it is convenient to alternatively describe a configuration

of particles by their positions

$$\cdots < x_2 < x_1 < \infty.$$

We also consider a more general version of the process where the first particle jumps at rate α, while all other particles jump at rate 1, and denote this process FTASEP(α). Let us define state spaces corresponding to configurations of particles in FTASEP(α) where the distance between consecutive particles is at most 2:

$$\mathbb{X}_{>0} := \Big\{(x_i)_{i\in\mathbb{Z}_{>0}} \in \mathbb{Z}^{\mathbb{Z}_{>0}} : \ \forall i \in \mathbb{Z}_{>0},\ x_i - x_{i+1} - 1 \in \{0,1\}\Big\},$$

and

$$\mathbb{X} := \Big\{(x_i)_{i\in\mathbb{Z}} \in \mathbb{Z}^{\mathbb{Z}} : \ \forall i \in \mathbb{Z},\ x_i - x_{i+1} - 1 \in \{0,1\}\Big\}.$$

Because of the facilitation rule, it is clear that the FTASEP(α) dynamics preserve both state spaces.

Definition 9 For $\alpha > 0$, the FTASEP(α) is a continuous-time Markov process defined on the state space $\mathbb{X}_{>0}$ via its Markov generator, acting on local functions $f : \mathbb{X}_{>0} \to \mathbb{R}$ by

$$\mathcal{L}_\alpha^{FTASEP} f(\mathbf{x}) = \alpha \mathbb{1}_{x_1 - x_2 = 1}\big(f(\mathbf{x}_1^+) - f(\mathbf{x})\big) + \sum_{i\geqslant 2} \mathbb{1}_{x_i - x_{i+1} = 1}\mathbb{1}_{x_{i-1}-x_i=2}\big(f(\mathbf{x}_i^+) - f(\mathbf{x})\big),$$

where we use the convention that the state $\mathbf{x}_i^+$ is obtained from $\mathbf{x}$ by incrementing by one the coordinate x_i.

Remark 1 One may similarly define FTASEP(α) on the state space $\mathbb{X}$ instead of $\mathbb{X}_{>0}$, in order to allow initial conditions without a rightmost particle.

In order to study FTASEP(α), we use a coupling with another interacting particle system: a TASEP with a source at the origin that injects particles at exponential rate α. We consider configurations of particles on $\mathbb{Z}_{>0}$ where each site can be occupied by at most one particle, and each particle jumps to the right by one at exponential rate 1, provided the target site is empty. At site 0 sits an infinite source of particles, which means that a particle always jumps to site 1 at exponential rate α when the site 1 is empty (See Fig. 5). We will denote the occupation variables in half-line TASEP by $g_i(t)$ (equals 1 if site i is occupied, 0 else).

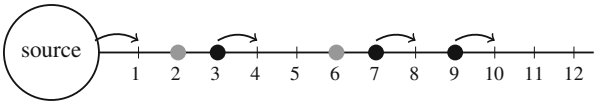


Fig. 5 Illustration of the half-line TASEP. The particles in gray cannot move because of the exclusion rule

Definition 10 The half-line TASEP with open boundary condition is a continuous-time Markov process defined on the state space $\{0,1\}^{\mathbb{Z}_{>0}}$ via its Markov generator, acting on local functions $f : \{0,1\}^{\mathbb{Z}_{>0}} \to \mathbb{R}$ by

$$\mathscr{L}_\alpha^{half} f(\mathbf{g}) = \alpha\big(f(1, g_2, g_3, \dots) - f(g_1, g_2, \dots)\big) + \sum_{x \in \mathbb{Z}_{>0}} g_x(1 - g_{x+1})\big(f(\mathbf{g}_{x,x+1}) - f(\mathbf{g})\big),$$

where the state $\mathbf{g}_{x,x+1}$ is obtained from $\mathbf{g}$ by exchanging occupation variables at sites x and $x+1$. We define the integrated current $N_x(t)$ as the number of particles on the right of site x (or at site x) at time t.

Define maps

$$\begin{aligned} \Phi_{>0} : \mathbb{X}_{>0} &\longrightarrow \{0,1\}^{\mathbb{Z}_{>0}} \\ (x_i)_{i\in\mathbb{Z}_{>0}} &\longmapsto \big(x_i - x_{i+1} - 1\big)_{i\in\mathbb{Z}_{>0}}, \end{aligned}$$

and

$$\begin{aligned} \Phi : \mathbb{X} &\longrightarrow \{0,1\}^{\mathbb{Z}} \\ (x_i)_{i\in\mathbb{Z}} &\longmapsto \big(x_i - x_{i+1} - 1\big)_{i\in\mathbb{Z}}. \end{aligned}$$

Proposition 1 *Let $\mathbf{x}(t) = (x_n(t))_{n\geqslant 1}$ be the particles positions in the FTASEP(α) started from some initial condition $\mathbf{x}(0) \in \mathbb{X}_{>0}$ (resp. $\mathbb{X}$). Then denoting $\mathbf{g}(t) = \{g_i(t)\}_{\mathbb{Z}_{>0}} = \Phi(\mathbf{x}(t))$, the dynamics of $\mathbf{g}(t)$ are those of half-line TASEP (resp. TASEP) starting from the initial configuration $\Phi_{>0}(\mathbf{x}(0))$ (resp. $\Phi(\mathbf{x}(0))$).*

Proof We explain how the mapping between the two processes works in the half-space case (which corresponds to the FTASEP(α) defined on the space of configurations $\mathbb{X}_{>0}$), since this is the case we will be most interested in this paper, and the full space case is very similar (Fig. 6).

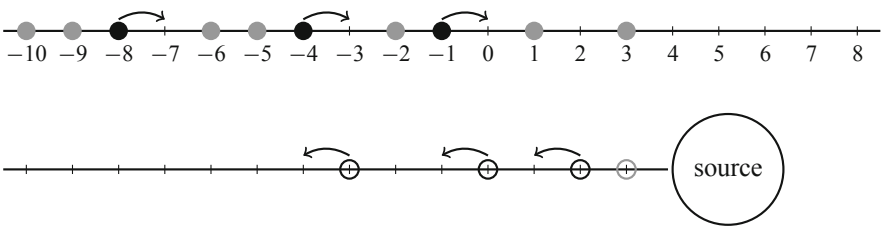


Fig. 6 Illustration of the coupling. The dynamics of particles in the bottom picture is nothing else but the dynamics of the holes in the top picture. In order to see it more precisely, consider the holes in the top picture and shrink the distances so that the distance between two consecutive holes decreases by 1; one gets exactly the bottom picture with the corresponding dynamics

Assume that particles at positions $(x_i)_{i\in\mathbb{Z}_{>0}}$ follow the FTASEP(α) dynamics, starting from some initial condition $\mathbf{x}(0) \in \mathbb{X}_{>0}$, and let us show that the $g_i = x_i - x_{i+1} - 1$ follow the dynamics of the half-line TASEP occupation variables.

If $x_1 = x_2 + 1$ (i.e. $g_1 = 0$), the first particle in FTASEP(α) jumps at rate α. After it has jumped, $x_1 = x_2 + 2$ (i.e. $g_1 = 1$). This corresponds to a particle arriving from the source to site 1 in the half-line TASEP. After this jump, $x_1 = x_2 + 2$ (i.e. $g_1 = 1$), so that the first particle cannot move in the FTASEP(α) because of the facilitation rule and no particle can jump from the source in half-line TASEP. More generally, because of the exclusion and facilitation rules, the $(i + 1)$th particle in FTASEP(α) can move only if $g_i = 1$ and $g_{i+1} = 0$ and does so at rate 1. After the move, x_{i+1} has increased by one so that $g_i = 0$ and $g_{i+1} = 1$. This exactly corresponds to the half-line TASEP dynamics.

Remark 2 Formally, Proposition 1 means that for any $\mathbf{x} \in \mathbb{X}_{>0}$ and local function $f : \{0, 1\}^{\mathbb{Z}>0} \to \mathbb{R}$,

$$\mathscr{L}_\alpha^{half} f(\Phi(\vec{x})) = \mathscr{L}_\alpha^{FTASEP}\big(f \circ \Phi\big)(\vec{x}).$$

In the following, we are mainly interested in FTASEP(α) starting from the step initial condition, or equivalently the half-line TASEP started from a configuration where all sites are initially empty.

Proposition 2 *Let $\mathbf{x}(t) = (x_n(t))\}_{n\geqslant 1}$ be the particles positions in the FTASEP(α) started from step initial condition (see Definition 9). Let $(N_x(t))_{x\in\mathbb{Z}_{>0}}$ be the currents in the half-line TASEP started from empty initial configuration (see Definition 10). Then we have the equality in law of the processes*

$$\big(x_n(t) + n\big)_{n\geqslant 1, t\geqslant 0} = \big(N_n(t)\big)_{n\geqslant 1, t\geqslant 0}.$$

Proof Because we start from step initial condition, $x_n(t) + n$ in FTASEP(α) equals the number of holes (empty sites) on the left of the nth particle. Using Proposition 1, and denoting the occupation variables in half-line TASEP by g_i, we have

$$x_n + n \overset{(d)}{=} \sum_{i\geqslant n} g_i = N_n,$$

jointly for all n as claimed.

Let us explain how Proposition 2 enables us to quickly recover the results from [9]. We later provide rigorous results substantiating many of these claims, but for the moment just proceed heuristically. Consider the case $\alpha = 1$. One expects (and we prove in the next Sect. 3.2) that the law of large numbers for the current of particles in the half-line TASEP is the same as in TASEP. Intuitively, this is because we expect that the law of large numbers is determined by a conservation PDE (of the form (2)) which is simply the restriction to a half-space of the conservation PDE governing

the hydrodynamics of TASEP on the full line. Thus,

$$\frac{N_{\kappa t}(t)}{t} \xrightarrow[t\to\infty]{a.s.} \frac{1}{4}(1-\kappa)^2.$$

Then, Proposition 2 implies that for FTASEP,

$$\frac{x_{\kappa t}}{t} \xrightarrow[t\to\infty]{a.s.} \frac{1}{4}(1-\kappa)^2 - \kappa = \frac{1-6\kappa+\kappa^2}{4}.$$

One can deduce the shape of the limiting density profile from the law of large numbers of particles positions. Let $\pi(\kappa)$ be the macroscopic position of the particle indexed by κt, i.e.

$$\pi(\kappa) = \frac{1-6\kappa+\kappa^2}{4}.$$

This yields $\kappa = 3-2\sqrt{2+\pi}$ (which can be interpreted as the limit of the integrated current in the FTASEP at site πt, rescaled by t). The density profile is obtained by differentiating κ with respect to π, and we get (as in [9, Equation (5)])

$$\rho(\pi t, t) = \frac{1}{\sqrt{2+\pi}}.$$

In light of the mapping between FTASEP and half-line TASEP from Proposition 1, it is possible to write down a family of translation invariant stationary measures in the FTASEP. They are given by choosing gaps between consecutive particles as i.i.d Bernoulli random variables. From these, we may also deduce the expression for the flux from (1). Assume that the system is at equilibrium, such that the gaps between consecutive particles are i.i.d. and distributed according to the $Bernoulli(p)$ distribution. Let us call ν_p this measure on $\{0,1\}^{\mathbb{Z}}$. Then, by the renewal theorem, the average density ρ is related to p via

$$\rho = \frac{1}{1+\mathbb{E}[\text{gap}]} = \frac{1}{1+p}.$$

The flux $j(\rho)$ is the product of the density times the drift of one particle, and since particles jump by 1, the drift is given by the probability of a jump for a tagged particle, i.e. $p(1-p)$. Indeed, considering a tagged particle in the stationary distribution, then its right neighbour has a probability p of being empty and its left neighbour has a probability $1-p$ of being occupied. This yields

$$j(\rho) = \rho(1-p)p = \frac{(1-\rho)(2\rho-1)}{\rho}. \tag{8}$$

3.2 Proofs of Limit Theorems

We use now the coupling from Proposition 2 to translate the asymptotic results about last passage percolation from Theorems 6, 7, 8 and 9 into limit theorems for the FTASEP(α).

Let $\mathbf{x}(t) = \{x_n(t)\}_{n \geqslant 1}$ be the particles positions in the FTASEP(α) started from step initial condition. Using Proposition 2, we have that for any $y \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}\big(x_n(t) \leqslant y\big) &= \mathbb{P}\big(x_n(t) \leqslant \lfloor y \rfloor\big) \\ &= \mathbb{P}\big(N_n(t) \leqslant \lfloor y \rfloor + n\big).\end{aligned}$$

In order to connect the problem with half-space last passage percolation, we use the next result.

Lemma 2 *Consider the exponential LPP model in a half-quadrant where the weights on the diagonal have parameter α, and recall the definition of last passage times $H(n, m)$ from Definition 1. Consider the half-line TASEP where the source injects particles at rate α with empty initial configuration and recall $N_x(t)$, the current at site x. Then for any $t > 0$ and $n, y \in \mathbb{Z}_{>0}$ we have that*

$$\mathbb{P}\big(N_n(t) \leqslant y\big) = \mathbb{P}\big(H(n + y - 1, y) \geqslant t\big).$$

Proof This is due to a standard mapping [17] between exclusion processes and last passage percolation, where the border of the percolation cluster can be interpreted as a height function for the exclusion process. More precisely, the processes have to be coupled in such a way that the weight w_{ij} in the LPP model is the $(i - j + 1)$th waiting time of the jth particle in the half-line TASEP – the waiting time is counted from the moment when it can jump, and by convention the first waiting time is when it jumps from the source into the system.

3.2.1 GSE ($\alpha > 1/2$) and GOE ($\alpha = 1/2$) Cases

By Theorems 6 we have that

$$H(n, n) = 4n + \sigma n^{1/3} \chi_n,$$

where $\sigma = 2^{4/3}$ and χ_n is a sequence of random variables weakly converging to the GSE (divided by $\sqrt{2}$ according to the convention chosen in Definition 7) distribution when $\alpha > 1/2$ and to the GOE distribution when $\alpha = 1/2$. Let $y \in \mathbb{R}$ be fixed and

$\varsigma > 0$ be a coefficient to specify later. For $t > 0$, we have

$$\begin{aligned}\mathbb{P}\left(x_1(t) \leqslant \frac{t}{4} + t^{1/3}\varsigma y\right) &= \mathbb{P}\left(H\left(\left\lfloor \frac{t}{4} + t^{1/3}\varsigma y\right\rfloor, \left\lfloor \frac{t}{4} + t^{1/3}\varsigma y\right\rfloor\right) \geqslant t\right) \\ &= \mathbb{P}\left(4\left(\left\lfloor \frac{t}{4} + t^{1/3}\varsigma y\right\rfloor\right) + \sigma\left\lfloor \frac{t}{4} + t^{1/3}\varsigma y\right\rfloor^{1/3}\chi_{\left\lfloor \frac{t}{4}+t^{1/3}\varsigma y\right\rfloor} \geqslant t\right) \\ &= \mathbb{P}\left(4\varsigma y t^{1/3} + \left(\sigma(t/4)^{1/3} + o(t^{1/3})\right)\chi_{\left\lfloor \frac{t}{4}+t^{1/3}\varsigma y\right\rfloor} \geqslant o(t^{1/3})\right)\end{aligned}$$

where the $o(t^{1/3})$ errors are deterministic. Thus, if we set $\varsigma = 2^{-4/3}$, we obtain that

$$\lim_{t\to\infty} \mathbb{P}\left(x_1(t) \geqslant \frac{t}{4} + t^{1/3}\varsigma y\right) = \lim_{t\to\infty} \mathbb{P}\left(\chi_{\left\lfloor \frac{t}{4}+t^{1/3}\varsigma y\right\rfloor} \leqslant -y\right) = F_{\mathrm{GSE}}(-\sqrt{2}y)$$

when $\alpha > 1/2$ and

$$\lim_{t\to\infty} \mathbb{P}\left(x_1(t) \geqslant \frac{t}{4} + t^{1/3}y\right) = F_{\mathrm{GOE}}(-y)$$

when $\alpha = 1/2$.

3.2.2 Gaussian Case

By Theorem 6 we have that

$$H(n,n) = h(\alpha)n + \sigma n^{1/2}G_n,$$

where $h(\alpha) = \frac{1}{\alpha(1-\alpha)}$, $\sigma = \frac{1-2\alpha}{\alpha^2(1-\alpha)^2}$ and G_n is a sequence of random variables weakly converging to the standard Gaussian when $\alpha < 1/2$. As in the previous case, let $y \in \mathbb{R}$ be fixed and $\varsigma > 0$ be a coefficient to specify later. For $t > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(x_1(t) \leqslant t/h(\alpha) + t^{1/2}\varsigma y\right) \\ &= \mathbb{P}\left(H\left(\left\lfloor \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\right\rfloor, \left\lfloor \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\right\rfloor\right) \geqslant t\right) \\ &= \mathbb{P}\left(h(\alpha)\left(\left\lfloor \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\right\rfloor\right) + \sigma\left\lfloor \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\right\rfloor^{1/2} G_{\left\lfloor \frac{t}{h(\alpha)}+t^{1/2}\varsigma y\right\rfloor} \geqslant t\right) \\ &= \mathbb{P}\left(h(\alpha)\varsigma y t^{1/2} + \sigma(t/h(\alpha))^{1/2} G_{\left\lfloor \frac{t}{4}+t^{1/3}\varsigma y\right\rfloor} \geqslant o(t^{1/2})\right).\end{aligned}$$

Thus, if we set $\varsigma = \frac{1-2\alpha}{\sqrt{\alpha(1-\alpha)}}$, we obtain that

$$\lim_{t\to\infty} \mathbb{P}\left(x_1(t) \geqslant \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\right) = \lim_{t\to\infty} \mathbb{P}\left(G_{\lfloor \frac{t}{h(\alpha)} + t^{1/2}\varsigma y\rfloor} \leqslant -y\right) = G(-y).$$

3.2.3 GUE Case

We have

$$\mathbb{P}\left(x_{\lfloor rt\rfloor}(t) \leqslant \pi t + \varsigma t^{1/3} y\right) =$$
$$\mathbb{P}\left(H\left(2\lfloor rt\rfloor + \lfloor \pi t + \varsigma t^{1/3} y\rfloor - 1, \lfloor rt\rfloor + \lfloor \pi t + \varsigma t^{1/3} y\rfloor\right) \geqslant t\right). \tag{9}$$

By Theorem 7 we have that for $m = \kappa n + \mathcal{O}(n^{1/3})$,

$$H(n, m) = (1 + \sqrt{\kappa})^2 n + \sigma n^{1/3} \chi_n,$$

where

$$\sigma = \frac{(1+\sqrt{\kappa})^{4/3}}{\sqrt{\kappa}^{1/3}}$$

and χ_n is a sequence of random variables weakly converging to the GUE distribution, for $\alpha > \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$. Hence

$$(9) = \mathbb{P}\left(\left(1 + \sqrt{\frac{\pi}{r+\pi}}\right)^2 n + \sigma n^{1/3}\chi_n \geqslant t\right)$$

where $n = 2\lfloor rt\rfloor + \lfloor \pi t + \varsigma t^{1/3} y\rfloor - 1$. Choosing π such that $\left(1 + \sqrt{\frac{r+\pi}{2r+\pi}}\right)^2 (2r + \pi) = 1$, i.e.

$$\pi = \frac{1 - 6r + r^2}{4},$$

we get that

$$(9) = \mathbb{P}\left(\left(1 + \sqrt{\frac{r+\pi}{2r+\pi}}\right)^2 \varsigma t^{1/3} y + \sigma\big((2r+\pi)t\big)^{1/3}\chi_n \geqslant o(t^{1/3})\right).$$

Hence, letting

$$\varsigma = \sigma(2r+\pi)^{4/3} = 2^{-4/3}\frac{(1+r)^{5/3}}{(1-r)^{1/3}},$$

yields

$$\lim_{t\to\infty} (9) = \lim_{n\to\infty} \mathbb{P}\big(\chi_n \geqslant -y + o(1)\big),$$

so that

$$\lim_{t\to\infty} \mathbb{P}\left(x_{\lfloor rt\rfloor}(t) \geqslant \frac{(1-6r+r^2)t}{4} + t^{1/3} y\varsigma\right) = F_{\mathrm{GUE}}(-y),$$

for $\alpha > \frac{1-r}{2}$ (This condition comes from the condition $\alpha > \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$ in Theorem 7).

3.2.4 Crossover Case

For the sake of clarity, we explain how the proof works in the one-point case. The multipoint case is similar. Assume

$$\alpha = \frac{1+2\sigma^{-1}\varpi n^{-1/3}}{2}.$$

Combining Lemma 2 and Proposition 2 as before,

$$\mathbb{P}\left(H_n(\eta) \leqslant p\right) = \mathbb{P}\left(x_{2n^{2/3}\xi\eta+1}(4n + (p\sigma - \xi^2\eta^2)n^{1/3}) \geqslant n - 3n^{2/3}\xi\eta\right),$$

where $\xi = 2^{2/3}$. Letting

$$t = 4n + (p\sigma - \xi^2\eta^2)n^{1/3},$$

we have that

$$\alpha = \frac{1+2^{4/3}\varpi t^{-1/3}}{2} + o(t^{-1/3}),$$

$$2n^{2/3}\xi\eta + 1 = 2^{1/3}\eta t^{2/3} + o(t^{1/3}),$$

$$n - 3n^{2/3}\xi = \frac{t}{4} - \frac{3\eta\xi}{2^{2}2^{1/3}}t^{2/3} + \frac{\eta^2\xi^2 - \sigma p}{2^{8/3}}t^{1/3} + o(t^{1/3}).$$

Hence, under this matching of parameters

$$\lim_{n\to\infty} \mathbb{P}\left(H_n(\eta) \leqslant p\right) = \lim_{t\to\infty} \mathbb{P}\left(X_t(\eta) \geqslant p\right),$$

where the rescaled position $X_n(\eta)$ is defined in (4).

Remark 3 Although we do not attempt in this paper to make an exhaustive analysis of the FTASEP with respect to varying initial conditions or parameters, such further analysis is allowed by our framework in several directions. In terms of initial condition, Proposition 1 allows to study the process starting form combinations of the wedge, flat or stationary initial data and translate to FTASEP some of the results known from TASEP [5, 15]. In terms of varying parameters, one could study the effect of varying α or the speed of the next few particles and one should observe the BBP transition [2] when considering fluctuations of x_{rt} for $r > 0$ (See Remark 1.6 in [1]).

4 Fredholm Pfaffian Formulas for k-Point Distributions

We recall in this Section a result from [1] which characterizes the joint probability distribution of passage times in the half-space exponential LPP model.

Proposition 3 ([1, Proposition 1.7]) *For any $h_1, \ldots, h_k > 0$ and integers $0 < n_1 < n_2 < \cdots < n_k$ and $m_1 > m_2 > \cdots > m_k$ such that $n_i > m_i$ for all i, we have that*

$$\mathbb{P}\big(H(n_1, m_1) < h_1, \ldots, H(n_k, m_k) < h_k\big) = \mathrm{Pf}\big(J - \mathsf{K}^{\mathrm{exp}}\big)_{\mathbb{L}^2\left(\mathbb{D}_k(h_1,\ldots,h_k)\right)},$$

where J is the matrix kernel

$$J(i, u; j, v) = \mathbb{1}_{(i,u)=(j,v)} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{10}$$

and

$$\mathbb{D}_k(g_1, \ldots, g_k) = \{(i, x) \in \{1, \ldots, k\} \times \mathbb{R} : x \geqslant g_i\}.$$

The kernel $\mathsf{K}^{\mathrm{exp}}$ was introduced in [1] in Section 4.4. It is defined on the state-space $(\{1, \ldots, k\} \times \mathbb{R})^2$ and takes values in the space of skew-symmetric 2×2 real

matrices. The entries are given by

$$\mathsf{K}^{\exp}(i,x;j,y) = \mathsf{I}^{\exp}(i,x;j,y) + \begin{cases} \mathsf{R}^{\exp}(i,x;j,y) & \text{when } \alpha > 1/2, \\ \hat{\mathsf{R}}^{\exp}(i,x;j,y) & \text{when } \alpha < 1/2, \\ \bar{\mathsf{R}}^{\exp}(i,x;j,y) & \text{when } \alpha = 1/2. \end{cases}$$

Recalling the Definition 3 for integration contours in the complex plane, we define $\mathsf{I}^{\exp}$ by the following formulas.

$$\mathsf{I}_{11}^{\exp}(i,x;j,y) := \int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{4zw(z+w)} e^{-xz-yw} \frac{(1+2z)^{n_i}(1+2w)^{n_j}}{(1-2z)^{m_i}(1-2w)^{m_j}} (2z+2\alpha-1)(2w+2\alpha-1).$$

$$\mathsf{I}_{12}^{\exp}(i,x;j,y) := \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{a_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{2z(z+w)} e^{-xz-yw} \frac{(1+2z)^{n_i}}{(1-2w)^{n_j}} \frac{(1+2w)^{m_j}}{(1-2z)^{m_i}} \frac{2\alpha-1+2z}{2\alpha-1-2w}, \tag{11}$$

where in the definition of the contours $\mathscr{C}_{a_z}^{\pi/3}$ and $\mathscr{C}_{a_w}^{\pi/3}$, the constants $a_z, a_w \in \mathbb{R}$ are chosen so that $0 < a_z < 1/2$, $a_z + a_w > 0$ and $a_w < (2\alpha-1)/2$.

$$\mathsf{I}_{22}^{\exp}(i,x;j,y) := \int_{\mathscr{C}_{b_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{b_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{z+w} e^{-xz-yw} \frac{(1+2z)^{m_i}(1+2w)^{m_j}}{(1-2z)^{n_i}(1-2w)^{n_j}} \frac{1}{2\alpha-1-2z} \frac{1}{2\alpha-1-2w}, \tag{12}$$

where in the definition of the contours $\mathscr{C}_{b_z}^{\pi/3}$ and $\mathscr{C}_{b_w}^{\pi/3}$, the constants $b_z, b_w \in \mathbb{R}$ are chosen so that $0 < b_z, b_w < (2\alpha-1)/2$ when $\alpha > 1/2$, while we impose only $b_z, b_w > 0$ when $\alpha \leqslant 1/2$.

We set $\mathsf{R}_{11}^{\exp}(i,x;j,y) = 0$, and $\mathsf{R}_{12}^{\exp}(i,x;j,y) = 0$ when $i \geqslant j$, and likewise for $\hat{R}^{\exp}$ and $\bar{R}^{\exp}$. The other entries depend on the value of α and the sign of $x-y$.

Case $\alpha > 1/2$**:** When $x > y$,

$$\mathsf{R}_{22}^{\exp}(i,x;j,y) = -\int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(1-2z)^{m_j}}{(1-2z)^{n_i}(1+2z)^{n_j}} \frac{1}{2\alpha-1-2z} \frac{1}{2\alpha-1+2z} 2ze^{-|x-y|z},$$

and when $x < y$

$$\mathsf{R}_{22}^{\exp}(i,x;j,y) = \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_j}(1-2z)^{m_i}}{(1-2z)^{n_j}(1+2z)^{n_i}} \frac{1}{2\alpha-1-2z} \frac{1}{2\alpha-1+2z} 2z e^{-|x-y|z},$$

where $(1-2\alpha)/2 < a_z < (2\alpha-1)/2$. One immediately checks that $\mathsf{R}_{22}^{\exp}$ is antisymmetric as we expect. When $i < j$ and $x > y$

$$\mathsf{R}_{12}^{\exp}(i,x;j,y) = -\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{n_i}}{(1+2z)^{n_j}} \frac{(1-2z)^{m_j}}{(1-2z)^{m_i}} e^{-|x-y|z},$$

while if $x < y$, $\mathsf{R}_{12}^{\exp}(i,x;j,y) = \mathsf{R}_{12}^{\exp}(i,y;j,x)$. Note that R_{12} is not antisymmetric nor symmetric (except when $k = 1$, i.e. for the one point distribution).

Case $\alpha < 1/2$**:** When $x > y$, we have

$$\begin{aligned}
\hat{\mathsf{R}}_{22}^{\exp}(i,x;j,y) &= \frac{-e^{\frac{1-2\alpha}{2}y}}{2} \int \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(2\alpha)^{m_j}}{(1-2z)^{n_i}(2-2\alpha)^{n_j}} \frac{e^{-xz}}{2\alpha-1+2z} \\
&\quad + \frac{e^{\frac{1-2\alpha}{2}x}}{2} \int \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_j}(2\alpha)^{m_i}}{(1-2z)^{n_j}(2-2\alpha)^{n_i}} \frac{e^{-yz}}{2\alpha-1+2z} \\
&\quad - \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(1-2z)^{m_j}}{(1-2z)^{n_i}(1+2z)^{n_j}} \frac{1}{2\alpha-1-2z} \frac{1}{2\alpha-1+2z} 2z e^{-|x-y|z} \\
&\quad - \frac{e^{(x-y)\frac{1-2\alpha}{2}}}{4} \frac{(2\alpha)^{m_i}(2-2\alpha)^{m_j}}{(2-2\alpha)^{n_i}(2\alpha)^{n_j}} + \frac{e^{(y-x)\frac{1-2\alpha}{2}}}{4} \frac{(2\alpha)^{m_j}(2-2\alpha)^{m_i}}{(2-2\alpha)^{n_j}(2\alpha)^{n_i}},
\end{aligned} \tag{13}$$

where the contours in the two first integrals pass to the right of $(1-2\alpha)/2$. When $x < y$, the sign of the third term is flipped so that $\hat{\mathsf{R}}_{22}^{\exp}(i,x;j,y) = -\hat{\mathsf{R}}_{22}^{\exp}(j,y;i,x)$. One can write slightly simpler formulas by reincorporating residues in the first two integrals: thus, when $x > y$,

$$\begin{aligned}
\hat{\mathsf{R}}_{22}^{\exp}(i,x;j,y) &= \frac{-e^{\frac{1-2\alpha}{2}y}}{2} \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(2\alpha)^{m_j}}{(1-2z)^{n_i}(2-2\alpha)^{n_j}} \frac{e^{-xz}}{2\alpha-1+2z} \\
&\quad + \frac{e^{\frac{1-2\alpha}{2}x}}{2} \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_j}(2\alpha)^{m_i}}{(1-2z)^{n_j}(2-2\alpha)^{n_i}} \frac{e^{-yz}}{2\alpha-1+2z} \\
&\quad - \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(1-2z)^{m_j}}{(1-2z)^{n_i}(1+2z)^{n_j}} \frac{1}{2\alpha-1-2z} \frac{1}{2\alpha-1+2z} 2z e^{-|x-y|z},
\end{aligned} \tag{14}$$

where $\frac{2\alpha-1}{2} < a_z < \frac{1-2\alpha}{2}$. When $i < j$, if $x > y$

$$\hat{\mathsf{R}}_{12}^{\exp}(i,x;j,y) = -\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{n_i}}{(1+2z)^{n_j}} \frac{(1-2z)^{m_j}}{(1-2z)^{m_i}} e^{-|x-y|z}, \tag{15}$$

while if $x < y$, $\hat{\mathsf{R}}_{12}^{\exp}(i,x;j,y) = \hat{\mathsf{R}}_{12}^{\exp}(i,y;j,x)$.
Case $\alpha = 1/2$**:** When $x > y$,

$$\begin{aligned}\bar{\mathsf{R}}_{22}^{\exp}(i,x;j,y) = &-\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}}{(1-2z)^{n_i}} \frac{e^{-xz}}{4z} + \int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_j}}{(1-2z)^{n_j}} \frac{e^{-yz}}{4z} \\ &+ \int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{m_i}(1-2z)^{m_j}}{(1-2z)^{n_i}(1+2z)^{n_j}} \frac{e^{-|x-y|z}}{2z} - \frac{1}{4},\end{aligned} \tag{16}$$

with a modification of the last two terms when $x < y$ so that $\bar{\mathsf{R}}_{22}^{\exp}(i,x;j,y) = -\bar{\mathsf{R}}_{22}^{\exp}(j,y;i,x)$. When $i < j$, if $x > y$

$$\bar{\mathsf{R}}_{12}^{\exp}(i,x;j,y) = -\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{(1+2z)^{n_i}}{(1+2z)^{n_j}} \frac{(1-2z)^{m_j}}{(1-2z)^{m_i}} e^{-|x-y|z},$$

while if $x < y$, $\bar{\mathsf{R}}_{12}^{\exp}(i,x;j,y) = \bar{\mathsf{R}}_{12}^{\exp}(i,y;j,x)$.

Remark 4 It may be possible to write simpler integral formulas for $\mathsf{K}^{\exp}$ by changing the contours used in the definition of $\mathsf{I}^{\exp}$ and identifying certain terms of $\mathsf{R}^{\exp}$ as residues of the integrand in $\mathsf{I}^{\exp}$. The reason why we have written the kernel $\mathsf{I}^{\exp}$ as above is mostly technical. For the asymptotic analysis of these formulas, it is convenient that all contours may be deformed so that they approach 0 without encountering any singularity, as will be explained in Sect. 5.

5 Asymptotic Analysis in the Crossover Regime

This section is devoted to the proofs of Theorems 8 and 9. We start by providing formulas for the correlation kernels $\mathsf{K}^{\mathrm{cross}}$ and K^{SU} used in the statements of both theorems.

5.1 Formulas for $\mathsf{K}^{\mathrm{cross}}$

The kernel $\mathsf{K}^{\mathrm{cross}}$ introduced in [1, Section 2.5] can be written as

$$\mathsf{K}^{\mathrm{cross}}(i,x;\,j,y)=\mathsf{I}^{\mathrm{cross}}(i,x;\,j,y)+\mathsf{R}^{\mathrm{cross}}(i,x;\,j,y),$$

where we have

$$\mathsf{I}_{11}^{\mathrm{cross}}(i,x;\,j,y)=\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\\ \frac{z+\eta_i-w-\eta_j}{z+w+\eta_i+\eta_j}\frac{z+\varpi+\eta_i}{z+\eta_i}\frac{w+\varpi+\eta_j}{w+\eta_j}e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{I}_{12}^{\mathrm{cross}}(i,x;\,j,y)=\int_{\mathscr{C}_{a_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_{a_w}^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\\ \frac{z+\eta_i-w+\eta_j}{2(z+\eta_i)(z+\eta_i+w-\eta_j)}\frac{z+\varpi+\eta_i}{-w+\varpi+\eta_j}e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{I}_{21}^{\mathrm{cross}}(i,x;\,j,y)=-\,\mathsf{I}_{12}^{\mathrm{cross}}(y,x),$$

$$\mathsf{I}_{22}^{\mathrm{cross}}(i,x;\,j,y)=\int_{\mathscr{C}_{b_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_{b_w}^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{z-\eta_i-w+\eta_j}{4(z-\eta_i+w-\eta_j)}\frac{e^{z^3/3+w^3/3-xz-yw}}{(z-\varpi-\eta_i)(w-\varpi-\eta_j)}.$$

The contours in $\mathsf{I}_{12}^{\mathrm{cross}}$ are chosen so that $a_z>-\eta_i$, $a_z+a_w>\eta_j-\eta_i$ and $a_w<\varpi+\eta_j$. The contours in $\mathsf{I}_{22}^{\mathrm{cross}}$ are chosen so that $b_z>\eta_i$, $b_z>\eta_i+\varpi$ and $b_w>\eta_j$, $b_w>\eta_j+\varpi$.

We have $\mathsf{R}_{11}^{\mathrm{cross}}(i,x;\,j,y)=0$, and $\mathsf{R}_{12}^{\mathrm{cross}}(i,x;\,j,y)=0$ when $i\geqslant j$. When $i<j$,

$$\mathsf{R}_{12}^{\mathrm{cross}}(i,x;\,j,y)=\frac{-\exp\left(\frac{-(\eta_i-\eta_j)^4+6(x+y)(\eta_i-\eta_j)^2+3(x-y)^2}{12(\eta_i-\eta_j)}\right)}{\sqrt{4\pi(\eta_j-\eta_i)}},$$

which may also be written as

$$\mathsf{R}_{12}^{\mathrm{cross}}(i,x;\,j,y)=-\int_{-\infty}^{+\infty}\mathrm{d}\lambda e^{-\lambda(\eta_i-\eta_j)}\mathrm{Ai}(x_i+\lambda)\mathrm{Ai}(x_j+\lambda).$$

The kernel $\mathsf{R}_{22}^{\mathrm{cross}}$ is antisymmetric, and when $x-\eta_i>y-\eta_j$ we have

$$\mathsf{R}_{22}^{\mathrm{cross}}(i,x;\,j,y)=\\ \frac{-1}{4}\int_{\mathscr{C}_{cz}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\frac{\exp\left((z+\eta_i)^3/3+(\varpi+\eta_j)^3/3-x(z+\eta_i)-y(\varpi+\eta_j)\right)}{\varpi+z}$$

$$+\frac{1}{4}\int_{\mathscr{C}_{c_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\frac{\exp\left((z+\eta_j)^3/3+(\varpi+\eta_i)^3/3-y(z+\eta_j)-x(\varpi+\eta_i)\right)}{\varpi+z}$$

$$-\frac{1}{2}\int_{\mathscr{C}_{d_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\frac{z\exp\left((z+\eta_i)^3/3+(-z+\eta_j)^3/3-x(z+\eta_i)-y(-z+\eta_j)\right)}{(\varpi+z)(\varpi-z)},$$

where the contours are chosen so that $c_z < -\varpi$ and d_z is between $-\varpi$ and ϖ.

5.2 *Formulas for* K^{SU}

The kernel K^{SU} introduced in [1, Section 2.5] decomposes as

$$\mathsf{K}^{\mathrm{SU}}(i,x;j,y)=\mathsf{I}^{\mathrm{SU}}(i,x;j,y)+\mathsf{R}^{\mathrm{SU}}(i,x;j,y),$$

where we have

$$\mathsf{I}_{11}^{\mathrm{SU}}(i,x;j,y)=\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_1^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{(z+\eta_i-w-\eta_j)e^{z^3/3+w^3/3-xz-yw}}{4(z+\eta_i)(w+\eta_j)(z+w+\eta_i+\eta_j)},$$

$$\mathsf{I}_{12}^{\mathrm{SU}}(i,x;j,y)=\int_{\mathscr{C}_{a_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_{a_w}^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{z+\eta_i-w+\eta_j}{2(z+\eta_i)(z+w+\eta_i-\eta_j)}e^{z^3/3+w^3/3-xz-yw},$$

$$\mathsf{I}_{21}^{\mathrm{SU}}(i,x;j,y)=-\mathsf{I}_{12}^{\mathrm{SU}}(j,x_j;i,y_i)$$

$$\mathsf{I}_{22}^{\mathrm{SU}}(i,x;j,y)=\int_{\mathscr{C}_{b_z}^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}_{b_w}^{\pi/3}}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\frac{z-\eta_i-w+\eta_j}{z-\eta_i+w-\eta_j}e^{z^3/3+w^3/3-xz-yw}.$$

The contours in $\mathsf{I}_{12}^{\mathrm{SU}}$ are chosen so that $a_z > -\eta_i$, $a_z+a_w > \eta_j-\eta_i$. The contours in $\mathsf{I}_{22}^{\mathrm{SU}}$ are chosen so that $b_z > \eta_i$ and $b_w > \eta_j$.

We have $\mathsf{R}_{11}^{\mathrm{SU}}(i,x;j,y)=0$, and $\mathsf{R}_{12}^{\mathrm{SU}}(i,x;j,y)=0$ when $i \geqslant j$. When $i<j$,

$$\mathsf{R}_{12}^{\mathrm{SU}}(i,x;j,y)=\mathsf{R}_{12}^{\mathrm{cross}}(i,x;j,y)=\frac{-\exp\left(\frac{-(\eta_i-\eta_j)^4+6(x+y)(\eta_i-\eta_j)^2+3(x-y)^2}{12(\eta_i-\eta_j)}\right)}{\sqrt{4\pi(\eta_j-\eta_i)}}.$$

The kernel $\mathsf{R}_{22}^{\mathrm{SU}}$ is antisymmetric, and when $x-\eta_i > y-\eta_j$ we have

$$\mathsf{R}_{22}^{\mathrm{SU}}(i,x;j,y)=$$
$$-\frac{1}{2}\int_{\mathscr{C}_0^{\pi/3}}\frac{\mathrm{d}z}{2\mathrm{i}\pi}z\exp\left((z+\eta_i)^3/3+(-z+\eta_j)^3/3-x(z+\eta_i)-y(-z+\eta_j)\right)$$

where the contours are chosen so that $a_z > -\varpi$ and b_z is between $-\varpi$ and ϖ.

5.3 *Proof of Theorem 8*

Recall that we scale α as

$$\alpha = \frac{1 + 2\sigma^{-1}\varpi n^{-1/3}}{2}.$$

The proof of Theorem 8 follows the same lines as that of Theorems 1.4 and 1.5 in Sections 5 and 6 of [1] (corresponding to Theorems 6 and 7 in the present paper). We introduce the rescaled correlation kernel

$$\mathsf{K}^{\exp,n}(i, x_i; j, x_j) :=$$
$$\begin{pmatrix} \sigma^2 n^{2/3} e^{\eta_i x_i + \eta_j x_j - \eta_i^3/3 - \eta_j^3/3} \mathsf{K}_{11}^{\exp}(i, X_i; j, X_j) & \sigma n^{1/3} e^{\eta_i x_i - \eta_j x_j - \eta_i^3/3 + \eta_j^3/3} \mathsf{K}_{12}^{\exp}(i, X_i; j, X_j) \\ \sigma n^{1/3} e^{-\eta_i x_i + \eta_j x_j + \eta_i^3/3 - \eta_j^3/3} \mathsf{K}_{21}^{\exp}(i, X_i; j, X_j) & e^{-\eta_i x_i - \eta_j x_j + \eta_i^3/3 + \eta_j^3/3} \mathsf{K}_{22}^{\exp}(i, X_i; j, X_j) \end{pmatrix},$$

where

$$X_i = 4n + n^{1/3}(\sigma x_i - \xi^2 \eta_i^2),$$

so that we have

$$\mathbb{P}\left(H_n(\eta_1) < x_1, \ldots, H_n(\eta_k) < x_k\right) = \mathrm{Pf}\big(J - \mathsf{K}^{\exp,\mathrm{n}}\big)_{\mathbb{L}^2(\mathbb{D}_k(x_1,\ldots,x_k))},$$

where the quantity $H_n(\eta)$ is defined in Sect. 1.2. We will decompose the kernel as $\mathsf{K}^{\exp,n}(i, x_i; j, x_j) = \mathsf{I}^{\exp,n}(i, x_i; j, x_j) + \mathsf{R}^{\exp,n}(i, x_i; j, x_j)$ according to the formulas in Sect. 4. The parameter α can be greater or smaller than $1/2$ depending on the sign of ϖ, so that we will need to be careful with the choice of contours.

In order to prove Theorem 8, we need to show that

$$\lim_{n\to\infty} \mathrm{Pf}\big(J - \mathsf{K}^{\exp,\mathrm{n}}\big)_{\mathbb{L}^2(\mathbb{D}_k(x_1,\ldots,x_k))} = \mathrm{Pf}\big(J - \mathsf{K}^{\mathrm{cross}}\big)_{\mathbb{L}^2(\mathbb{D}_k(x_1,\ldots,x_k))}. \tag{17}$$

We will first show that the kernel $\mathsf{K}^{\exp,\mathrm{n}}(i, x; j, y)$ converges to $\mathsf{K}^{\mathrm{cross}}(i, x; j, y)$ for fixed $(i, x; j, y)$. Then, we will prove uniform bounds on the kernel $\mathsf{K}^{\exp,\mathrm{n}}$ so that the Fredholm Pfaffian is an absolutely convergent series of integrals and hence the pointwise convergence of kernels implies the convergence of Fredholm Pfaffians.

We introduce two types of modifications of the contour $\mathscr{C}_0^{\pi/3}$. For a parameter $r > 0$, we denote by $\mathscr{C}[r]$ the contour formed by the union of an arc of circle around 0 of radius $rn^{-1/3}$, between $-\pi/3$ and $\pi/3$, and two semi-infinite rays in directions $\pm\pi/3$ that connect the extremities of the arc to ∞ (see Fig. 7, left). With

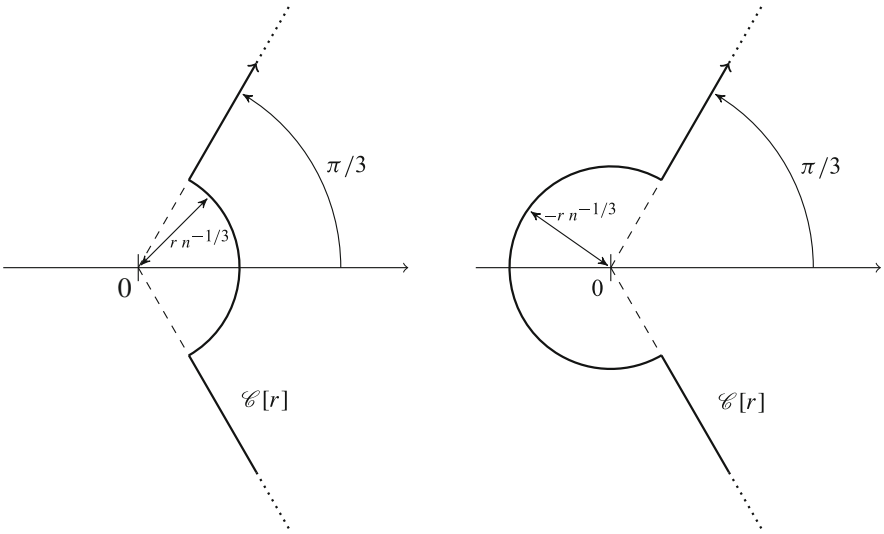


Fig. 7 The contours $\mathscr{C}[r]$ when $r > 0$ (left) and $\mathscr{C}[r]$ when $r < 0$ (right)

this definition 0 is on the left of the contour $\mathscr{C}[r]$. For a parameter $r < 0$, we denote by $\mathscr{C}[r]$ a similar contour where the arc of circle has radius $-r$ and is now between angles from $\pi/3$ to $5\pi/3$ so that 0 is on the right of $\mathscr{C}[r]$ (see Fig. 7, right).

Thanks to Cauchy's theorem, we have some freedom to deform the contours used in the definition of $\mathsf{K}^{\exp}$ in Sect. 4, as long as we do not cross any pole. Thus we can write

$$\mathsf{K}_{11}^{\exp,n}(i,x;j,y) = e^{\eta_i x+\eta_j y-\eta_i^3/3-\eta_j^3/3}\sigma^2 n^{2/3}\int_{\mathscr{C}[1]}\frac{\mathrm{d}z}{2\mathrm{i}\pi}\int_{\mathscr{C}[1]}\frac{\mathrm{d}w}{2\mathrm{i}\pi}\,\frac{z-w}{4zw(z+w)}$$
$$(2z+2\sigma^{-1}\varpi n^{-1/3})(2w+2\sigma^{-1}\varpi n^{-1/3})\exp\Big(n(f(z)+f(w))$$
$$+n^{2/3}(\xi\eta_i\log(1-4z^2)+\xi\eta_j\log(1-4w^2))$$
$$+n^{1/3}\xi^2\eta_i^2 z+n^{1/3}\xi^2\eta_j^2 w-n^{1/3}\sigma(xz+yw)\Big), \qquad (18)$$

where the function f is

$$f(z) = -4z+\log(1+2z)-\log(1-2z).$$

To take asymptotics of this expression, we use Laplace's method. The function f has a double critical point at 0. We have

$$f(z) = \frac{\sigma^3}{3}z^3+\mathscr{O}(z^4), \qquad (19)$$

where $\sigma = 2^{4/3}$ and we know from Lemma 5.9 in [1] that the contour $\mathscr{C}_0^{\pi/3}$ is steep-descent for $\mathfrak{Re}[f]$ (which shows that the main contribution to the integral comes from integration in a neighborhood of 0, see the proof of Theorem 6 in Section 5 of [1]). Let us make the change of variables $z = n^{-1/3}z/\sigma$ and likewise for w, and use Taylor expansions of all terms in the integrand. Using the same kind of estimates (to control the error made when approximating the integrand) as in Proposition 5.8 in [1], we arrive at

$$\mathsf{K}_{11}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} e^{\eta_i x+\eta_j y-\eta_i^3/3-\eta_j^3/3} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{z+w} \frac{z+\varpi}{z} \frac{w+\varpi}{w}$$
$$\exp\left(z^3/3+w^3/3-4\xi\eta_i z^2/\sigma^2-4\xi\eta_j w^2/\sigma^2+\xi^2\eta_i z/\sigma+\xi^2\eta_j w/\sigma-xz-yw\right).$$

With our choice of σ and ξ, we have that $4\xi/\sigma^2 = \xi^2/\sigma = 1$, so that after a change of variables (a simple translation where z becomes $z+\eta_i$ and w becomes $w+\eta_j$),

$$\mathsf{K}_{11}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} \mathsf{K}_{11}^{\mathrm{cross}}(i,x;j,y) =$$
$$\int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_1^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z+\eta_i-w-\eta_j}{(z+\eta_i)(w+\eta_j)} \frac{z+\varpi+\eta_i}{z+\eta_i} \frac{w+\varpi+\eta_j}{w+\eta_j} e^{z^3/3+w^3/3-xz-yw}.$$

Regarding K_{12}, we write $\mathsf{K}_{12}^{\exp,n} = \mathsf{I}_{12}^{\exp,n} + \mathsf{R}_{12}^{\exp,n}$ where

$$\mathsf{I}_{12}^{\exp,n}(i,x;j,y) = e^{\eta_i x-\eta_j y-\eta_i^3/3+\eta_j^3/3}\sigma n^{1/3} \int_{\mathscr{C}[a_z]} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}[a_w]} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{2z(z+w)}$$
$$\frac{2z+2\sigma^{-1}\varpi n^{-1/3}}{-2w+2\sigma^{-1}\varpi n^{-1/3}} \exp\Big(n(f(z)+f(w))$$
$$+n^{2/3}(\xi\eta_i\log(1-4z^2)-\xi\eta_j\log(1-4w^2))+n^{1/3}\xi^2\eta_i^2 z$$
$$+n^{1/3}\xi^2\eta_j^2 w - n^{1/3}\sigma(xz+yw)\Big), \qquad (20)$$

where the contours are chosen so that $a_z > 0$, $a_z + a_w > 0$ and $a_w < \varpi$. Applying Laplace method as for K_{11}, we arrive at

$$\mathsf{I}_{12}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} e^{\eta_i x-\eta_j y-\eta_i^3/3+\eta_j^3/3} \int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{a_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{2z(z+w)} \frac{z+\varpi}{-w+\varpi}$$
$$\exp\left(z^3/3+w^3/3-4\xi\eta_i z^2/\sigma^2+4\xi\eta_j w^2/\sigma^2+\xi^2\eta_i z/\sigma+\xi^2\eta_j w/\sigma-xz-yw\right).$$

Thus, we find that after a change of variables

$$\mathsf{I}_{12}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} \mathsf{I}_{12}^{\text{cross}}(i,x;j,y) =$$
$$\int_{\mathscr{C}_{a_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{a_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z+\eta_i-w+\eta_j}{2(z+\eta_i)(z+\eta_i+w-\eta_j)} \frac{z+\varpi+\eta_i}{-w+\varpi+\eta_j} e^{z^3/3+w^3/3-xz-yw},$$

where the contours in the last equation are now chosen so that $a_z > -\eta_i$, $a_z + a_w > \eta_j - \eta_i$ and $a_w < \varpi + \eta_j$. When $i < j$ (and consequently $\eta_i < \eta_j$), and for x, y such that $\sigma x - \xi^2\eta_i > \sigma y - \xi^2\eta_j$ (which is equivalent to $x - \eta_i > y - \eta_j$), we use Eq. (15) for R_{12} and find

$$\mathsf{R}_{12}^{\exp,\mathrm{n}}(i,x;j,y) \xrightarrow[n\to\infty]{} -\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \exp\big((z-\eta_i)^3/3-(z-\eta_j)^3/3-x(z-\eta_i)+y(z-\eta_j)\big).$$

One can check that with $x - \eta_i > y - \eta_j$, the integrand is integrable on the contour $\mathscr{C}_{1/4}^{\pi/3}$. When $x - \eta_i < y - \eta_j$ however, we have

$$\mathsf{R}_{12}^{\exp,\mathrm{n}}(i,x;j,y) \xrightarrow[n\to\infty]{} -\int_{\mathscr{C}_{1/4}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \exp\big((z+\eta_j)^3/3-(z+\eta_i)^3/3+x(z+\eta_i)-y(z+\eta_j)\big).$$

One can evaluate the integrals above, and we find that in both cases

$$\mathsf{R}_{12}^{\exp,\mathrm{n}}(i,x;j,y) \xrightarrow[n\to\infty]{} \mathsf{R}_{12}^{\text{cross}}(i,x;j,y) = \frac{-\exp\left(\frac{-(\eta_i-\eta_j)^4+6(x+y)(\eta_i-\eta_j)^2+3(x-y)^2}{12(\eta_i-\eta_j)}\right)}{\sqrt{4\pi(\eta_j-\eta_i)}}.$$

As for K_{22}, we again decompose the kernel as $\mathsf{K}_{22}^{\exp,n} = \mathsf{I}_{22}^{\exp,n} + \mathsf{R}_{22}^{\exp,n}$. For $\mathsf{I}_{22}^{\exp,n}$, we chose contours that pass to the right of all poles except $1/2$, as in the case $\alpha = 1/2$ of Sect. 4. We can write

$$\mathsf{I}_{22}^{\exp,n}(i,x;j,y) = e^{-\eta_i x-\eta_j y+\eta_i^3/3+\eta_j^3/3} \int_{\mathscr{C}[b_z]} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}[b_w]} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{z+w}$$
$$\frac{1}{(-2z+2\sigma^{-1}\varpi n^{-1/3})(-2w+2\sigma^{-1}\varpi n^{-1/3})} \exp\Big(n(f(z)+f(w))$$
$$+n^{2/3}(-\xi\eta_i\log(1-4z^2)-\xi\eta_j\log(1-4w^2))$$
$$+n^{1/3}\xi^2\eta_i^2 z+n^{1/3}\xi^2\eta_j^2 w-n^{1/3}\sigma(xz+yw)\Big), \qquad (21)$$

where b_z and b_w are positive and greater that ϖ. Again, by Laplace's method we obtain

$$\mathsf{I}_{22}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} e^{-\eta_i x-\eta_j y+\eta_i^3/3+\eta_j^3/3} \int_{\mathscr{C}_{b_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{b_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-w}{z+w}$$
$$\frac{\exp\left(z^3/3+w^3/3+4\xi\eta_i z^2/\sigma^2+4\xi\eta_j w^2/\sigma^2+\xi^2\eta_i z/\sigma+\xi^2\eta_j w/\sigma-xz-yw\right)}{(2z-2\varpi)(2w-2\varpi)}.$$

Thus,

$$\mathsf{I}_{22}^{\exp,n}(i,x;j,y) \xrightarrow[n\to\infty]{} \mathsf{I}_{12}^{\mathrm{cross}}(i,x;j,y) =$$
$$\int_{\mathscr{C}_{b_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \int_{\mathscr{C}_{b_w}^{\pi/3}} \frac{\mathrm{d}w}{2\mathrm{i}\pi} \frac{z-\eta_i-w+\eta_j}{4(z-\eta_i+w-\eta_j)} \frac{e^{z^3/3+w^3/3-xz-yw}}{(z-\varpi-\eta_i)(w-\varpi-\eta_j)},$$

where the contours are chosen so that $b_z > \eta_i$, $b_z > \eta_i + \varpi$ and $b_w > \eta_j$, $b_w > \eta_j + \varpi$. For R_{22} we use (13). Note that the form of the expression does not depend on whether ϖ is positive or negative, because of our choice of contours for $\mathsf{I}_{22}^{\exp,n}$ in (21). We find for $x_i - \eta_i > x_j - \eta_j$

$$\mathsf{R}_{22}^{\exp,\,\mathrm{n}}(i,x;j,y) \xrightarrow[n\to\infty]{}$$
$$\frac{-1}{4}\int_{\mathscr{C}_{c_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{\exp\left((z+\eta_i)^3/3+(\varpi+\eta_j)^3/3-x(z+\eta_i)-y(\varpi+\eta_j)\right)}{\varpi+z}$$
$$+\frac{1}{4}\int_{\mathscr{C}_{c_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{\exp\left((z+\eta_j)^3/3+(\varpi+\eta_i)^3/3-y(z+\eta_j)-x(\varpi+\eta_i)\right)}{\varpi+z}$$
$$-\frac{1}{2}\int_{\mathscr{C}_{d_z}^{\pi/3}} \frac{\mathrm{d}z}{2\mathrm{i}\pi} \frac{z\exp\left((z+\eta_i)^3/3+(-z+\eta_j)^3/3-x(z+\eta_i)-y(-z+\eta_j)\right)}{(\varpi+z)(\varpi-z)}$$
$$-\frac{1}{4}\exp\left((-\varpi+\eta_j)^3/3+(\varpi+\eta_i)^3/3-y(-\varpi+\eta_j)-x(\varpi+\eta_i)\right)$$
$$+\frac{1}{4}\exp\left((-\varpi+\eta_i)^3/3+(\varpi+\eta_j)^3/3-x(-\varpi+\eta_i)-y(\varpi+\eta_j)\right),$$

where the contours are chosen so that $c_z > -\varpi$ and d_z is between $-\varpi$ and ϖ. When $x-\eta_i < y-\eta_j$, $\mathsf{R}_{22}^{\exp,\mathrm{n}}$ is determined by antisymmetry.

At this point, we have shown that when $\alpha = 1/2$ and for any set of points $\{i_r, x_{i_r}; j_s, x_{j_s}\}_{1\leqslant r,s\leqslant k} \in \{1,\dots,k\}\times\mathbb{R}$,

$$\mathrm{Pf}\Big(\mathsf{K}^{\exp,\mathrm{n}}\big(i_r,x_{i_r};j_s,x_{j_s}\big)\Big)_{r,s=1}^{k} \xrightarrow[q\to 1]{} \mathrm{Pf}\Big(\mathsf{K}^{\mathrm{cross}}\big(i_r,x_{i_r};j_s,x_{j_s}\big)\Big)_{r,s=1}^{k}.$$

In order to conclude that the Fredholm Pfaffian likewise has the desired limit, one needs a control on the entries of the kernel $\mathsf{K}^{\exp,\mathrm{n}}$, in order to apply dominated convergence.

Lemma 3 *Let $a \in \mathbb{R}$ and $0 \geqslant \eta_1 < \cdots < \eta_k$ be fixed. There exist positive constants C, c, m for $n > m$ and $x, y > a$,*

$$\begin{aligned}
\left|\mathsf{K}_{11}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C\exp\big(-cx-cy\big),\\
\left|\mathsf{K}_{12}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C\exp\big(-cx\big),\\
\left|\mathsf{K}_{22}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C.
\end{aligned}$$

Proof The proof is very similar to that of Lemmas 5.11 and 6.4 in [1]. Indeed, using the same approach as in the proof of these lemmas, we obtain that

$$\begin{aligned}
\left|\mathsf{I}_{11}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C\exp\big(-cx-cy\big),\\
\left|\mathsf{I}_{12}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C\exp\big(-cx\big),\\
\left|\mathsf{I}_{22}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C\exp\big(-cx-cy\big),
\end{aligned}$$

and

$$\begin{aligned}
\left|\mathsf{R}_{11}^{\exp,\mathrm{n}}(i,x;j,y)\right| &= 0,\\
\left|\mathsf{R}_{12}^{\exp,\mathrm{n}}(i,x;j,y)\right| &\leqslant C\mathbb{1}_{i<j}\exp\big((x+y)(\eta_i-\eta_j)\big),\\
\left|\mathsf{R}_{22}^{\exp,\mathrm{n}}(i,x;j,y)\right| &< C.
\end{aligned}$$

Recall that when $i < j$, $\eta_i - \eta_j < 0$, so that the bounds on $\mathsf{I}^{\exp,\mathrm{n}}$ and $\mathsf{R}^{\exp,\mathrm{n}}$ combine together to the statement of Lemma 3.

The bounds from Lemma 3 are such that the hypotheses in Lemma 1 are satisfied. We conclude, applying dominated convergence in the Pfaffian series expansion, that

$$\lim_{n\to\infty}\mathbb{P}\left(\bigcap_{i=1}^{k}\big\{H_n(\eta_i) < x_i\big\}\right) = \mathrm{Pf}\big(J-\mathsf{K}^{\mathrm{cross}}\big)_{\mathbb{L}^2(\mathbb{D}_k(x_1,\ldots,x_k))}.$$

5.4 *Proof of Theorem 9*

The proof is very similar as that of Theorem 8. We use a similar rescaling of the kernel: we define the rescaled kernel

$$\mathsf{K}^{\exp,n}(i,x_i;j,x_j) :=$$

$$\begin{pmatrix} \varpi^{-2}\sigma^2 n^{2/3} e^{\eta_i x_i+\eta_j x_j-\eta_i^3/3-\eta_j^3/3}\mathsf{K}_{11}^{\exp}(i,X_i;j,X_j) & \varpi^{-1}\sigma n^{1/3} e^{\eta_i x_i-\eta_j x_j-\eta_i^3/3+\eta_j^3/3}\mathsf{K}_{12}^{\exp}(i,X_i;j,X_j) \\ \varpi^{-1}\sigma n^{1/3} e^{-\eta_i x_i+\eta_j x_j+\eta_i^3/3-\eta_j^3/3}\mathsf{K}_{21}^{\exp}(i,X_i;j,X_j) & \varpi^{2} e^{-\eta_i x_i-\eta_j x_j+\eta_i^3/3+\eta_j^3/3}\mathsf{K}_{22}^{\exp}(i,X_i;j,X_j) \end{pmatrix},$$

Then, we decompose the kernel as $\mathsf{K}^{\exp,n}(i,x_i;j,x_j) = \mathsf{I}^{\exp,n}(i,x_i;j,x_j) + \mathsf{R}^{\exp,n}(i,x_i;j,x_j)$ using the formulas (and choice of contours) of Sect. 4 in the case $\alpha > 1/2$. Thus, the formulas are slightly simpler than in the proof of Theorem 8. To show that the kernel $\mathsf{K}^{\exp,n}$ converges pointwise to K^{SU}, we follow the same steps as in the proof of Theorem 8 as if $\varpi = +\infty$ and contours $\mathscr{C}[a_z]$, $\mathscr{C}[a_w]$, $\mathscr{C}[b_z]$, $\mathscr{C}[b_w]$ are chosen to be consistent with the constraints on contours in the case $\alpha > 1/2$ of Sect. 4. Finally, the kernel satisfies the same uniform bounds as in Lemma 3, so that we conclude the proof by dominated convergence as above.

Acknowledgements G.B and I.C. would like to thank Sidney Redner for drawing their attention to the Facilitated TASEP. Part of this work was done during the stay of J.B, G.B and I.C. at the Kavli Institute of Theoretical Physics and supported by the National Science Foundation under Grant No. NSF PHY11-25915. J.B. was supported in part by NSF grant DMS-1361782, DMS- 1664692 and DMS-1664531, and the Simons Fellows program. G.B. was partially supported by the Laboratoire de Probabilités et Modèles Aléatoires UMR CNRS 7599, Université Paris-Diderot–Paris 7 and the Packard Foundation through I.C.'s Packard Fellowship. I.C. was partially supported by the NSF through DMS-1208998 and DMS-1664650, the Clay Mathematics Institute through a Clay Research Fellowship, the Institute Henri Poincaré through the Poincaré Chair, and the Packard Foundation through a Packard Fellowship for Science and Engineering.

References

1. Baik, J., Barraquand, G., Corwin, I., Suidan, T.: Pfaffian Schur processes and last passage percolation in a half-quadrant. To appear in Ann. Probab. (2017). arXiv:1606.00525
2. Baik, J., Ben Arous, G., Péché, S.: Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann. Probab. **33** 1643–1697 (2005)
3. Barraquand, G., Corwin, I.: The q-Hahn asymmetric exclusion process. Ann. Appl. Probab. **26**(4), 2304–2356 (2016)
4. Basu, U., Mohanty, P.K.: Active absorbing-state phase transition beyond directed percolation: a class of exactly solvable models. Phys. Rev. E **79**, 041143 (2009). https://doi.org/10.1103/PhysRevE.79.041143
5. Ben Arous, G., Corwin, I.: Current fluctuations for TASEP: a proof of the Prähofer–Spohn conjecture. Ann. Probab. **39**(1), 104–138 (2011)
6. Corwin, I.: The Kardar–Parisi–Zhang equation and universality class. Random Matrices Theory Appl. **1**(1), 1130001 (2012)

7. Evans, L.C.: Partial differential equations. In: Graduate Studies in Mathematics, vol. 19. AMS, Providence (1998)
8. Forrester, P.J., Nagao, T., Honner, G.: Correlations for the orthogonal-unitary and symplectic-unitary transitions at the hard and soft edges. Nucl. Phys. B **553**(3), 601–643 (1999)
9. Gabel, A., Krapivsky, P.L., Redner, S.: Facilitated asymmetric exclusion. Phys. Rev. Lett. **105**(21), 210603 (2010)
10. Gabel, A., Redner, S.: Cooperativity-driven singularities in asymmetric exclusion. J. Stat. Mech. **2011**(6), P06008 (2011)
11. Halpin-Healy, T., Takeuchi, K.A.: A KPZ cocktail-shaken, not stirred.... J. Stat. Phys. **160**(4), 794–814 (2015)
12. Krug, J., Meakin, P., Halpin-Healy, T.: Amplitude universality for driven interfaces and directed polymers in random media. Phys. Rev. A **45**, 638–653 (1992). https://doi.org/10.1103/PhysRevA.45.638
13. Lee, E., Wang, D.: Distributions of a particle's position and their asymptotics in the q-deformed totally asymmetric zero range process with site dependent jumping rates. arXiv preprint:1703.08839 (2017)
14. Liggett, T.M.: Interacting Particle Systems. Springer, Berlin (2005)
15. Prähofer, M., Spohn, H.: Current fluctuations for the totally asymmetric simple exclusion process. In: In and Out of Equilibrium (Mambucaba, 2000). Progress in Probability, vol. 51, pp. 185–204. Birkhäuser, Boston (2002)
16. Rains, E.M.: Correlation functions for symmetrized increasing subsequences. arXiv preprint math/0006097 (2000)
17. Rost, H.: Non-equilibrium behaviour of a many particle process: density profile and local equilibria. Probab. Theory Rel. Fields **58**(1), 41–53 (1981)
18. Spohn, H.: KPZ scaling theory and the semi-discrete directed polymer model. MSRI Proceedings. arXiv:1201.0645 (2012)
19. Tracy, C.A., Widom, H.: Asymptotics in ASEP with step initial condition. Commun. Math. Phys. **290**(1), 129–154 (2009)

Stochastic Functional Differential Equations and Sensitivity to Their Initial Path

D. R. Baños, G. Di Nunno, H. H. Haferkorn, and F. Proske

Abstract We consider systems with memory represented by stochastic functional differential equations. Substantially, these are stochastic differential equations with coefficients depending on the past history of the process itself. Such coefficients are hence defined on a functional space. Models with memory appear in many applications ranging from biology to finance. Here we consider the results of some evaluations based on these models (e.g. the prices of some financial products) and the risks connected to the choice of these models. In particular we focus on the impact of the initial condition on the evaluations. This problem is known as the analysis of sensitivity to the initial condition and, in the terminology of finance, it is referred to as the Delta. In this work the initial condition is represented by the relevant past history of the stochastic functional differential equation. This naturally leads to the redesign of the definition of Delta. We suggest to define it as a functional directional derivative, this is a natural choice. For this we study a representation formula which allows for its computation without requiring that the evaluation functional is differentiable. This feature is particularly relevant for applications. Our formula is achieved by studying an appropriate relationship between Malliavin derivative and functional directional derivative. For this we introduce the technique of *randomisation of the initial condition*.

D. R. Baños · H. H. Haferkorn · F. Proske
Department of Mathematics, University of Oslo, Oslo, Norway
e-mail: davidru@math.uio.no; hanneshh@math.uio.no; proske@math.uio.no

G. Di Nunno (✉)
Department of Mathematics, University of Oslo, Oslo, Norway

Norwegian School of Economics and Business Administration, Bergen, Norway
e-mail: giulian@math.uio.no

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_2

1 Introduction

Several phenomena in nature show evidence of both a stochastic behaviour and a dependence on the past history when evaluating the present state. Examples of models taking into account both features come from biology in the different areas of population dynamics, see e.g. [8, 26], or gene expression, see e.g. [27], or epidemiology, see e.g. [11]. We find several stochastic models dealing with delay and memory also in the different areas of economics and finance. The delayed response in the prices of both commodities and financial assets is studied for example in [1, 2, 5, 6, 12, 13, 23–25, 36, 37]. The very market inefficiency and also the fact that traders persistently use past prices as a guide to decision making induces memory effects that may be held responsible for market bubbles and crashes. See e.g. [3, 22].

In this work we consider a general stochastic dynamic model incorporating delay or memory effects. Indeed we consider stochastic functional differential equations (SFDE), which are substantially stochastic differential equations with coefficients depending on the past history of the dynamic itself. These SFDEs have already been studied in the pioneering works of [28, 29, 38] in the Brownian framework. The theory has later been developed including models for jumps in [9]. From another perspective models with memory have been studied via the so-called functional Itô calculus as introduced in [17] and then developed steadily in e.g. [14, 15]. For a comparison of the two approaches we refer to e.g. [16, 18], see also [9, Appendix] for a short survey on the different notions of derivative. In the deterministic framework functional differential equations are widely studied. See, e.g. [21].

By model risk we generically mean all risks entailed in the choice of a model in view of prediction or forecast. One aspect of model risk management is the study of the sensitivity of a model to the estimates of its parameters. In this paper we are interested in the sensitivity to the initial condition. In the terminology of mathematical finance this is referred to as the Delta. However, in the present setting of SFDEs, the very concept of Delta has to be defined as new, being the initial condition an initial path and not only a single initial point as in the standard stochastic differential equations. It is the first time that the sensitivity to the initial path is tackled, though it appears naturally whenever working in presence of memory effects.

As illustration, on the probability space $(\Omega, \mathscr{F}, P)$, let us consider the SFDE:

$$\begin{cases} dx(t) = f(t, x(t), x_t)dt + g(t, x(t), x_t)dW(t), \quad t \in (0, T] \\ (x(0), x_0) = \eta \end{cases}$$

where by $x(t)$ we mean the evaluation at time t of the solution process and by x_t we mean the segment of past that is relevant for the evaluation at t. Let us also consider the evaluation $p(\eta)$ at $t = 0$ of some value $\Phi({}^{\eta}x(T), {}^{\eta}x_T)$ at $t = T$ of a functional

Φ of the model. Such evaluation is represented as the expectation:

$$p(\eta) = E\left[\Phi({}^{\eta}x(T), {}^{\eta}x_T)\right]. \tag{1}$$

We have marked explicitly the dependence on the initial path η by an anticipated superindex.

Evaluations of this type are typical in the pricing of financial derivatives, which are financial contracts with payoff Ψ written on an underlying asset with price dynamics S given by an SFDE of the type above. Indeed in this case the classical non arbitrage pricing rule provides a fair price in the form

$$p_{risk-neutral}(\eta) = E_{{}^{\eta}Q}\left[\frac{\Psi({}^{\eta}S(T), {}^{\eta}S_T)}{N(T)}\right] = E\left[{}^{\eta}Z(T)\frac{\Psi({}^{\eta}S(T), {}^{\eta}S_T)}{N(T)}\right],$$

where ${}^{\eta}Z(T) = \frac{d^{\eta}Q}{dP}$ is the Radon-Nykodim derivative of the risk-neutral probability measure ${}^{\eta}Q$ and $N(T)$ is a chosen numéraire used for discounting. We observe that such pricing measure ${}^{\eta}Q$ depends on η by construction.

Analogously, in the so-called benchmark approach to pricing (see e.g. [32]), a non-arbitrage fair price is given in the form

$$p_{benchmark}(\eta) = E\left[\frac{\Psi({}^{\eta}S(T), {}^{\eta}S_T)}{{}^{\eta}G(T)}\right],$$

where ${}^{\eta}G(T)$ is the value of an appropriate benchmark process, used in discounting and guaranteeing that the very P is an appropriate pricing measure. Here we note that the benchmark depends on the initial path η of the underlying price dynamics. Both pricing approaches can be represented as (1) and from now on we shall generically call *payoff* the functional Φ, borrowing the terminology from finance.

Then, in the present notations, the study of the sensitivity to the initial condition consists in the study of some derivative of $p(\eta)$:

$$\frac{\partial}{\partial\eta}p(\eta) = \frac{\partial}{\partial\eta}E\left[\Phi({}^{\eta}x(T), {}^{\eta}x_T)\right].$$

and its possible representations.

In this work we interpret the derivative above as a functional directional derivative and we study formulae for its representations. Our approach takes inspiration from the seminal papers [19, 20]. Here Malliavin calculus is used to obtain a nice formula, where the derivative is itself represented as an expectation of the product of the functional Φ and some random variable, called Malliavin weight.

We remark immediately that the presence of memory has effects well beyond the expected and the formulae we obtain will not be, unfortunately, so elegant. The representation formulae we finally obtain do not formally present or require the Fréchet differentiability of Φ. This is particularly relevant for applications e.g. to pricing. To obtain our formulae we shall study the relationship between functional Fréchet derviatives and Malliavin derivatives. However, this relationship has to be

carefully constructed. Our technique is based on what we call *the randomisation of the initial path condition*, which is based on the use of an independent Brownian noise to "shake" the past.

The paper is organised as follows. In Sect. 2 we provide a detailed background of SFDEs. The first part of Sect. 3 is dedicated to the study of the sensitivity to the initial path condition and the technique of randomisation. We obtain a general representation formula for the sensitivity. Here we see that there is a balance between the generality of the functional Φ allowed and the regularity on the coefficients of the dynamics of the underlying. The second part of Sect. 3 presents further detailed results in the case of a suitable randomisation choice. The Appendix contains some technical proof, given with the aim of a self-contained reading.

2 Stochastic Functional Differential Equations

In this section we present a general setup for stochastic functional differential equations (SFDEs). Our framework is inspired by and generalises [5, 6, 25].

2.1 The Model

On the complete probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]}, P)$ where the filtration satisfies the usual assumptions and is such that $\mathscr{F} = \mathscr{F}_T$, we consider $W = \{W(t,\omega);\ \omega \in \Omega, t \in [0,T]\}$ an m-dimensional standard $(\mathscr{F}_t)_{t\in[0,T]}$-Brownian motion. Here let $T \in [0,\infty)$.

We are interested in stochastic processes $x : [-r, T] \times \Omega \to \mathbb{R}^d$, $r \geqslant 0$, with finite second order moments and a.s. continuous sample paths. So, one can look at x as a random variable $x : \Omega \to \mathscr{C}([-r,T],\mathbb{R}^d)$ in $L^2(\Omega, \mathscr{C}([-r,T],\mathbb{R}^d))$ where $\mathscr{C}([-r,T],\mathbb{R}^d)$ is the space of continuous functions from $[-r,T]$ to $\mathbb{R}^d$. In fact, we can look at x as

$$x : \Omega \to \mathscr{C}([-r,T],\mathbb{R}^d) \hookrightarrow L^2([-r,T],\mathbb{R}^d) \hookrightarrow \mathbb{R}^d \times L^2([-r,T],\mathbb{R}^d)$$

where the notation $\hookrightarrow$ stands for *continuously embedded in*, which holds since the domains are compact. The parameter r here above introduced represents the time gap linked to the delay or memory effect.

From now on, for any $u \in [0,T]$, we write $M_2([-r,u],\mathbb{R}^d) := \mathbb{R}^d \times L^2([-r,u],\mathbb{R}^d)$ for the so-called Delfour-Mitter space endowed with the norm

$$\|(v,\theta)\|_{M_2} = \left(|v|^2 + \|\theta\|_2^2\right)^{1/2}, \quad (v,\theta) \in M_2([-r,u],\mathbb{R}^d), \tag{2}$$

where $\|\cdot\|_2$ stands for the L^2-norm and $|\cdot|$ for the Euclidean norm in $\mathbb{R}^d$. For short we denote $M_2 := M_2([-r,0],\mathbb{R}^d)$.

The interest of using such space comes from two facts. On the one hand, the space M_2 endowed with the norm (2) has a Hilbert structure which allows for a Fourier representation of its elements. On the other hand, as we will see later on, the point 0 plays an important role and therefore we need to distinguish between two processes in $L^2([-r,0],\mathbb{R}^d)$ that have different images at the point 0. In general the spaces $M_2([-r,u],\mathbb{R}^d)$ are also natural to use since they coincide with the corresponding spaces of continuous functions $\mathscr{C}([-r,u],\mathbb{R}^d)$ completed with respect to the norm (2), by taking the natural injection $i(\varphi(\cdot)) = (\varphi(u), \varphi(\cdot)1_{[-r,u)})$ for a $\varphi \in \mathscr{C}([-r,u],\mathbb{R}^d)$ and by closing it.

Furthermore, by the continuous embedding above, we can consider the random process $x : \Omega \times [-r,u] \longrightarrow \mathbb{R}^d$ as a random variable

$$x : \Omega \longrightarrow M_2([-r,u],\mathbb{R}^d)$$

in $L^2(\Omega, M_2([-r,u],\mathbb{R}^d))$, that is

$$\|x\|_{L^2(\Omega,M_2([-r,u],\mathbb{R}^d))} = \left(\int_\Omega \|x(\omega)\|^2_{M_2([-r,u],\mathbb{R}^d)} P(d\omega)\right)^{1/2} < \infty.$$

For later use, we write $L^2_A(\Omega, M_2([-r,u],\mathbb{R}^d))$ for the subspace of $L^2(\Omega, M_2([-r,u],\mathbb{R}^d))$ of elements that admit an $(\mathscr{F}_t)_{t\in[0,u]}$-adapted modification.

To deal with memory and delay we use the concept of segment of x. Given a process x, the delay gap r, and a specified time $t \in [0,T]$, the *segment of* x in the past time interval $[t-r,t]$ is denoted by $x_t(\omega,\cdot) : [-r,0] \to \mathbb{R}^d$ and it is defined as

$$x_t(\omega,s) := x(\omega, t+s), \qquad s \in [-r,0].$$

So $x_t(\omega,\cdot)$ is the segment of the ω-trajectory of the process x, and contains all the information of the past down to time $t-r$. In particular, the segment of x_0 relative to time $t=0$ is the initial path and carries the information about the process from before $t=0$.

Assume that, for each $\omega \in \Omega$, $x(\cdot,\omega) \in L^2([-r,T],\mathbb{R}^d)$. Then $x_t(\omega)$ can be seen as an element of $L^2([-r,0],\mathbb{R}^d)$ for each $\omega \in \Omega$ and $t \in [0,T]$. Indeed the couple $(x(t), x_t)$ is a $\mathscr{F}_t$-measurable random variable with values in M_2, i.e. $(x(t,\omega), x_t(\omega,\cdot)) \in M_2$, given $\omega \in \Omega$.

Let us consider an $\mathscr{F}_0$-measurable random variable $\eta \in L^2(\Omega, M_2)$. To shorten notation we write $\mathbb{M}_2 := L^2(\Omega, M_2)$. A stochastic functional differential equation (SFDE), is written as

$$\begin{cases} dx(t) = f(t,x(t),x_t)dt + g(t,x(t),x_t)dW(t), \quad t \in [0,T] \\ (x(0),x_0) = \eta \in \mathbb{M}_2 \end{cases} \tag{3}$$

where

$$f : [0, T] \times M_2 \to \mathbb{R}^d \quad \text{and} \quad g : [0, T] \times M_2 \to L(\mathbb{R}^m, \mathbb{R}^d).$$

2.2 Existence and Uniqueness of Solutions

Under suitable hypotheses on the functionals f and g, one obtains existence and uniqueness of the strong solution (in the sense of L^2) of the SFDE (3). The solution is a process $x \in L^2(\Omega, M_2([-r, T], \mathbb{R}^d))$ admitting an $(\mathscr{F}_t)_{t\in[0,T]}$-adapted modification, that is, $x \in L^2_A(\Omega, M_2([-r, T], \mathbb{R}^d))$.

We say that two processes $x^1, x^2 \in L^2(\Omega, M_2([-r, T], \mathbb{R}^d))$ are L^2-unique, or unique in the L^2-sense if $\|x_1 - x_2\|_{L^2(\Omega, M_2([-r,T],\mathbb{R}^d))} = 0$.

Hypotheses (EU):

(EU1) (Local Lipschitzianity) The drift and the diffusion functionals f and g are Lipschitz on bounded sets in the second variable uniformly with respect to the first, i.e., for each integer $n \geqslant 0$, there is a Lipschitz constant L_n independent of $t \in [0, T]$ such that,

$$|f(t, \varphi_1) - f(t, \varphi_2)|_{\mathbb{R}^d} + \|g(t, \varphi_1) - g(t, \varphi_2)\|_{L(\mathbb{R}^m,\mathbb{R}^d)} \leqslant L_n \|\varphi_1 - \varphi_2\|_{M_2}$$

for all $t \in [0, T]$ and functions $\varphi_1, \varphi_2 \in M_2$ such that $\|\varphi_1\|_{M_2} \leqslant n$, $\|\varphi_2\|_{M_2} \leqslant n$.

(EU2) (Linear growths) There exists a constant $C > 0$ such that,

$$|f(t, \psi)|_{\mathbb{R}^d} + \|g(t, \psi)\|_{L(\mathbb{R}^m,\mathbb{R}^d)} \leqslant C \left(1 + \|\psi\|_{M_2}\right)$$

for all $t \in [0, T]$ and $\psi \in M_2$.

The following result belongs to [28, Theorem 2.1]. Its proof is based on an approach similar to the one in the classical deterministic case based on successive Picard approximations.

Theorem 1 (Existence and Uniqueness) *Given Hypotheses* (EU) *on the coefficients f and g and the initial condition $\eta \in \mathbb{M}_2$, the SFDE (3) has a (strong) solution ${}^{\eta}x \in L^2_A(\Omega, M_2([-r, T], \mathbb{R}^d))$ which is unique in the sense of L^2. The solution (or better its adapted representative) is a process ${}^{\eta}x : \Omega \times [-r, T] \to \mathbb{R}^d$ such that*

(1) ${}^{\eta}x(t) = \eta(t)$, $t \in [-r, 0]$.
(2) ${}^{\eta}x(\omega) \in M_2([-r, T], \mathbb{R}^d)$ ω-*a.s.*
(3) *For every $t \in [0, T]$, ${}^{\eta}x(t) : \Omega \to \mathbb{R}^d$ is $\mathscr{F}_t$-measurable.*

From the above we see that it makes sense to write

$$ {}^{\eta}x(t) = \begin{cases} \eta(0) + \int_0^t f(u,\ {}^{\eta}x(u),\ {}^{\eta}x_u)du + \int_0^t g(u,\ {}^{\eta}x(u),\ {}^{\eta}x_u)dW(u),\ t \in [0,T] \\ \eta(t),\ t \in [-r,0]. \end{cases} $$

Observe that the above integrals are well defined. In fact, the process

$$ (\omega, t) \mapsto ({}^{\eta}x(t,\omega),\ {}^{\eta}x_t(\omega)) $$

belongs to $\mathbb{M}_2$ and is adapted since x is pathcontinuous and adapted and its composition with the deterministic coefficients f and g is then adapted as well. Note that ${}^{\eta}x$ represents the solution starting off at time 0 with initial condition $\eta \in \mathbb{M}_2$.

One could consider the same dynamics but starting off at a later time, let us say, $s \in (0,T]$, with initial condition $\eta \in \mathbb{M}_2$. Namely, we could consider:

$$ \begin{cases} dx(t) = f(t,x(t),x_t)dt + g(t,x(t),x_t)dW(t), \quad t \in (s,T] \\ x(t) = \eta(t-s), \quad t \in [s-r,s]. \end{cases} \tag{4} $$

Again, under (EU) the SFDE (4) has the solution,

$$ {}^{\eta}x^s(t) = \begin{cases} \eta(0) + \int_s^t f(u,\ {}^{\eta}x^s(u),\ {}^{\eta}x^s_u)du + \int_s^t g(u,\ {}^{\eta}x^s(u),\ {}^{\eta}x^s_u)dW(u), \quad t \in [s,T] \\ \eta(t-s), \quad t \in [s-r,s] \end{cases} \tag{5} $$

The right-hand side superindex in ${}^{\eta}x^s$ denotes the starting time. We will omit the superindex when starting at 0, ${}^{\eta}x^0 = {}^{\eta}x$. The interest of defining the solution to (4) starting at any time s comes from the semigroup property of the flow of the solution which we present in the next subsection. For this reason we introduce the notation

$$ X^s_t(\eta,\omega) := X(s,t,\eta,\omega) := ({}^{\eta}x^s(t,\omega), {}^{\eta}x^s_t(\omega)), \qquad \omega \in \Omega,\ s \leqslant t. \tag{6} $$

In relation to (4) we also define the following evaluation operator:

$$ \rho_0 : M_2 \to \mathbb{R}^d, \quad \rho_0\varphi := v \quad \text{for any } \varphi = (v,\theta) \in M_2. $$

We observe here that the random variable ${}^{\eta}x^s(t)$ is an evaluation at 0 of the process $X^s_t(\eta), t \in [s,T]$.

2.3 Differentiability of the Solution

We recall that our goal is the study of the influence of the initial path η on the functionals of the solution of (3). For this we need to ensure the existence of an at-least-once differentiable stochastic flow for (3). Hereafter we discuss the differentiability conditions on the coefficients of the dynamics to ensure such property on the flow.

In general, suppose we have E and F Banach spaces, $U \subseteq E$ an open set and $k \in \mathbb{N}$. We write $L^k(E, F)$ for the space of continuous k-multilinear operators $A : E^k \to F$ endowed with the uniform norm

$$\|A\|_{L^k(E,F)} := \sup\{\|A(v_1, \ldots, v_k)\|_F,\ \|v_i\|_E \leqslant 1,\ i = 1, \ldots, k\}.$$

Then an operator $f : U \to F$ is said to be of class $\mathscr{C}^{k,\delta}$ if it is C^k and $D^k f : U \to L^k(E, F)$ is δ-Hölder continuous on bounded sets in U. Moreover, $f : U \to F$ is said to be of class $\mathscr{C}_b^{k,\delta}$ if it is C^k, $D^k f : U \to L^k(E, F)$ is δ-Hölder continuous on U, and all its derivatives $D^j f$, $1 \leqslant j \leqslant k$ are globally bounded on U. The derivative D is taken in the Fréchet sense.

First of all we consider SFDEs in the special case when

$$g(t, (\varphi(0), \varphi(\cdot))) = g(t, \varphi(0)), \quad \varphi = (\varphi(0), \varphi(\cdot)) \in \mathbb{M}_2$$

that is, g is actually a function $[0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$.

For completeness we give the definition of stochastic flow.

Definition 1 Denote by $S([0, T]) := \{s, t \in [0, T] : 0 \leqslant s < t < T\}$. Let E be a Banach space. A stochastic $\mathscr{C}^{k,\delta}$-semiflow on E is a measurable mapping $X : S([0, T]) \times E \times \Omega \to E$ satisfying the following properties:

(i) For a.e. $\omega \in \Omega$, the map $X(\cdot, \cdot, \cdot, \omega) : S([0, T]) \times E \to E$ is continuous.
(ii) For fixed $(s, t) \in S([0, T])$ the map $X(s, t, \cdot, \omega) : E \to E$ is $\mathscr{C}^{k,\delta}$ for a.e. $\omega \in \Omega$.
(iii) For $0 \leqslant s \leqslant u \leqslant t$, a.e. $\omega \in \Omega$ and $x \in E$, the property $X(s, t, \eta, \omega) = X(u, t, X(s, u, \eta, \omega), \omega)$ holds.
(iv) For all $(t, \eta) \in [0, T] \times E$ and a.e. $\omega \in \Omega$, one has $X(t, t, \eta, \omega) = \eta$.

In our setup, we consider the space $E = M_2$.

Hypotheses (FlowS):

(FlowS1) The function $f : [0, T] \times M_2 \to \mathbb{R}^d$ is jointly continuous; the map $M_2 \ni \varphi \mapsto f(t, \varphi)$ is Lipschitz on bounded sets in M_2 and $\mathscr{C}^{1,\delta}$ uniformly in t (i.e. the δ-Hölder constant is uniformly bounded in $t \in [0, T]$) for some $\delta \in (0, 1]$.
(FlowS2) The function $g : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is jointly continuous; the map $\mathbb{R}^d \ni v \mapsto g(t, v)$ is $\mathscr{C}_b^{2,\delta}$ uniformly in t.

(FlowS3) **One** of the following conditions is satisfied:

(a) There exist $C > 0$ and $\gamma \in [0, 1)$ such that

$$|f(t, \varphi)| \leqslant C(1 + \|\varphi\|_{M_2}^{\gamma})$$

for all $t \in [0, T]$ and all $\varphi \in M_2$

(b) For all $t \in [0, T]$ and $\varphi \in M_2$, one has $f(t, \varphi, \omega) = f(t, \varphi(0), \omega)$. Moreover, it exists $r_0 \in (0, r)$ such that

$$f(t, \varphi, \omega) = f(t, \tilde{\varphi}, \omega)$$

for all $t \in [0, T]$ and all $\tilde{\varphi}$ such that $\varphi(\cdot)1_{[-r,-r_0]}(\cdot) = \tilde{\varphi}(\cdot)1_{[-r,-r_0]}(\cdot)$.

(c) For all $\omega \in \Omega$,

$$\sup_{t\in[0,T]} \|(D\psi(t, v, \omega))^{-1}\|_{M_2} < \infty,$$

where $\psi(t, v)$ is defined by the stochastic differential equation

$$\begin{cases} d\psi(t, v) = g(t, \psi(t, v))dW(t), \\ \psi(0, v) = v. \end{cases}$$

Moreover, there exists a constant C such that

$$|f(t, \varphi)| \leqslant C(1 + \|\varphi\|_{M_2})$$

for all $t \in [0, T]$ and $\varphi \in M_2$.

Then, [29, Theorem 3.1] states the following theorem.

Theorem 2 *Under Hypotheses* (EU) *and* (FlowS)*, $X_t^s(\eta, \omega)$ defined in* (6) *is a $\mathscr{C}^{1,\varepsilon}$-semiflow for every $\varepsilon \in (0, \delta)$.*

Next, we can consider a more general diffusion coefficient g following the approach introduced in [29, Section 5]. Let us assume that the function g is of type:

$$g(t, (x(t), x_t)) = \bar{g}(t, x(t), a + \int_0^t h(s, (x(s), x_s))ds),$$

for some constant a and some functions $\bar{g}$ and h satisfying some regularity conditions that will be specified later. This case can be transformed into a system of the previous type where the diffusion coefficient does not explicitly depend on the segment. In fact, defining $y(t) := (y^{(1)}(t), y^{(2)}(t))^{\top}$ where $y^{(1)}(t) := x(t)$, $t \in [-r, T]$, $y^{(2)}(t) := a + \int_0^t h(s, (x(s), x_s))ds$, $t \in [0, T]$ and $y^{(2)}(t) := 0$ on

$[-r, 0]$, we have the following dynamics for y:

$$\begin{cases} dy(t) = F(t, y(t), y_t)dt + G(t, y(t))dW(t), \\ y(0) = (\eta(0), a)^\top,\ y_0 = (\eta, 0)^\top, \end{cases} \tag{7}$$

where

$$F(t, y(t), y_t) = \begin{pmatrix} f(t, y^{(1)}(t), y_t^{(1)}) \\ h(t, y^{(1)}(t), y_t^{(1)}) \end{pmatrix},$$
$$G(t, y(t)) = \begin{pmatrix} \bar{g}(t, y^{(1)}(t), y^{(2)}(t)) \\ 0 \end{pmatrix}. \tag{8}$$

The transformed system (7) is now an SFDE of type (3) where the diffusion coefficient does not explicitly depend on the segment. That is the differentiability of the flow can be studied under the corresponding Hypotheses (FlowS). Hereafter, we specify the conditions on $\bar{g}$ and h so that Hypotheses (EU) and (FlowS) are satisfied by the transformed system (7). Since the conditions (FlowS3)(a) and (b) are both too restrictive for (7), we will make sure that (FlowS3)(c) is satisfied. Under these conditions we can guarantee the differentiability of the solutions to the SFDE (4) for the above class of diffusion coefficient g.

Hypotheses (Flow):

(Flow1) f satisfies (FlowS1) and there exists a constant C such that

$$|f(t, \varphi)| \leqslant C(1 + \|\varphi\|_{M_2})$$

for all $t \in [0, T]$ and $\varphi \in M_2$.

(Flow2) $g(t, \varphi)$ is of the following form

$$g(t, \varphi) = \bar{g}(t, v, \tilde{g}(\theta)), \quad t \in [0, T], \quad \varphi = (v, \theta) \in M_2$$

where $\bar{g}$ satisfies the following conditions:

(a) The function $\bar{g} : [0, T] \times \mathbb{R}^{d+k} \to \mathbb{R}^{d \times m}$ is jointly continuous; the map $\mathbb{R}^{d+k} \ni y \mapsto \bar{g}(t, y)$ is $\mathscr{C}_b^{2,\delta}$ uniformly in t.

(b) For each $v \in \mathbb{R}^{d+k}$, let $\{\Psi(t, v)\}_{t \in [0,T]}$ solve the stochastic differential equation

$$\Psi(t, v) = v + \begin{pmatrix} \int_0^t \bar{g}(s, \Psi(s, v))dW(s) \\ 0 \end{pmatrix},$$

where 0 denotes the null-vector in $\mathbb{R}^k$. Then $\Psi(t, v)$ is Fréchet differentiable with respect to v and the Jacobi-matrix $D\Psi(t, v)$ is invertible and fulfils, for all $\omega \in \Omega$,

$$\sup_{\substack{t\in[0,T]\\ v\in\mathbb{R}^{d+k}}} \|D\Psi^{-1}(t, v, \omega)\| < \infty, \text{ where } \|\cdot\| \text{ denotes any matrix norm.}$$

and, $\tilde{g} : L^2([-r, 0], \mathbb{R}^d) \to \mathbb{R}^k$ satisfies the following conditions:

(c) It exists a jointly continuous function $h : [0, T] \times M_2 \to \mathbb{R}^k$ s.t. for each $\tilde{\varphi} \in L^2([-r, T], \mathbb{R}^d)$,

$$\tilde{g}(\tilde{\varphi}_t) = \tilde{g}(\tilde{\varphi}_0) + \int_0^t h(s, (\tilde{\varphi}(s), \tilde{\varphi}_s))ds,$$

where $\tilde{\varphi}_t \in L^2([-r, 0], \mathbb{R}^d)$ is the segment at t of a representative of $\tilde{\varphi}$.

(d) $M_2 \ni \varphi \mapsto h(t, \varphi)$ is Lipschitz on bounded sets in M_2, uniformly with respect to $t \in [0, T]$ and $\mathscr{C}^{\lambda,\delta}$ uniformly in t.

Corollary 1 *Under Hypotheses* (Flow)*, the stochastic flow* $X_t^s(\eta) = X(s, t, \eta, \omega)$*,* $\omega \in \Omega$*,* $t \geqslant s$ *to* (4) *is a* $\mathscr{C}^{\lambda,\varepsilon}$*-semiflow for every* $\varepsilon \in (0, \delta)$*. In particular,* $\varphi \mapsto X(s, t, \varphi, \omega)$ *is* C^1 *in the Fréchet sense.*

3 Sensitivity Analysis to the Initial Path Condition

From now on, we consider a stochastic process x which satisfies dynamics (3), where the coefficients f and g are such that conditions (EU) and (Flow) are satisfied.

Our final goal is to study the sensitivity of evaluations of type

$$p(\eta) = E\left[\Phi(X_T^0(\eta))\right] = E\left[\Phi({}^{\eta}x(T),\ {}^{\eta}x_T)\right],\ \eta \in \mathbb{M}_2 \tag{9}$$

to the initial path in the model ${}^{\eta}x$. Here, $\Phi : M_2 \to \mathbb{R}$ is such that $\Phi(X_T^0(\eta)) \in L^2(\Omega, \mathbb{R})$. The sensitivity will be interpreted as the directional derivative

$$\partial_h p(\eta) := \frac{d}{d\varepsilon} p(\eta + \varepsilon h)\Big|_{\varepsilon=0} = \lim_{\varepsilon\to 0} \frac{p(\eta + \varepsilon h) - p(\eta)}{\varepsilon},\ \ h \in M_2. \tag{10}$$

Hence we shall study pertubations in direction $h \in M_2$. The final aim is to give a representation of $\partial_h p(\eta)$ in which the function Φ is not directly differentiated. This is in the line with the representation of the sensitivity parameter Delta by means of weights. See, e.g. the Malliavin weight introduced in [19, 20] for the classical case of no memory. For the sake of clarity in notation in the sequel we use ∂_h for directional, D for Fréchet and $\mathscr{D}$ for Malliavin derivative. Hereafter, we impose some regularity

conditions on f and g:

Hypotheses (H):

(H1) (Global Lipschitzianity) $\varphi \mapsto f(t,\varphi)$, $\varphi \mapsto g(t,\varphi)$ are globally Lipschitz uniformly in t with Lipschitz constants L_f and L_g, i.e.

$$|f(t,\varphi_1) - f(t,\varphi_2)|_{\mathbb{R}^d} \leqslant L_f \|\varphi_1 - \varphi_2\|_{M_2}$$
$$\|g(t,\varphi_1) - g(t,\varphi_2)\|_{L(\mathbb{R}^m,\mathbb{R}^d)} \leqslant L_g \|\varphi_1 - \varphi_2\|_{M_2}$$

for all $t \in [0,T]$ and $\varphi_1, \varphi_2 \in M_2$.

(H2) (Lipschitzianity of the Fréchet derivatives) $\varphi \mapsto Df(t,\varphi)$, $\varphi \mapsto Dg(t,\varphi)$ are globally Lipschitz uniformly in t with Lipschitz constants L_{Df} and L_{Dg}, i.e.

$$\|Df(t,\varphi_1) - Df(t,\varphi_2)\| \leqslant L_{Df} \|\varphi_1 - \varphi_2\|_{M_2}$$
$$\|Dg(t,\varphi_1) - Dg(t,\varphi_2)\| \leqslant L_{Dg} \|\varphi_1 - \varphi_2\|_{M_2}$$

for all $t \in [0,T]$ and $\varphi_1, \varphi_2 \in M_2$.

The corresponding stochastic $\mathscr{C}^{1,1}$-semiflow is again denoted by X.

Before proceeding, we give a simple example of SFDE satisfying all assumptions (EU), (Flow) and (H).

Example 1 Consider the SFDE (3) where the functions f and g are given by

$$f(t,\varphi) = M(t)\varphi(0) + \int_{-r}^{0} \bar{M}(s)\varphi(s)ds,$$
$$g(t,\varphi) = \Sigma(t)\varphi(0) + \int_{-r}^{0} \bar{\Sigma}(s)\varphi(s)ds,$$

where $M : [0,T] \to \mathbb{R}^{d\times d}$, $\bar{M} : [-r,0] \to \mathbb{R}^{d\times d}$, $\Sigma : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^{d\times m})$, and $\bar{\Sigma} : [-r,0] \to L(\mathbb{R}^d, \mathbb{R}^{d\times m})$ are bounded differentiable functions, $\bar{\Sigma}(-r) = 0$ and $s \mapsto \bar{\Sigma}'(s) = \frac{d}{ds}\bar{\Sigma}(s)$ are bounded as well.

Obviously, f and g satisfy (EU) and (H) and therefore also (Flow1). In order to check conditions (Flow2), we note that

$$g(t,\varphi) = \bar{g}(t,\varphi(0),\tilde{g}(\varphi(\cdot))),$$

where

$$\bar{g}(t,y) = \Sigma(t)y^{(1)} + y^{(2)},\ y = (y^{(1)},y^{(2)})^\top,\ \text{and}\ \tilde{g}(\varphi(\cdot)) = \int_{-r}^{0} \bar{\Sigma}(s)\varphi(s)ds.$$

The function $\bar{g}$ satisfies condition (Flow2)(a) as Σ is bounded and continuous. Let us check condition (Flow2)(b) in the case $d = m = 1$. Then $\bar{g}(t,y) = \sigma(t)y^{(1)} + y^{(2)}$, where σ is a real valued, differentiable function and Ψ fulfils the

two-dimensional stochastic differential equation

$$\begin{cases} \Psi^{(1)}(t,v) = v^{(1)} + \int_0^t \left(\sigma(s)\Psi^{(1)}(s,v) + v^{(2)}\right) dW(s), \\ \Psi^{(2)}(t,v) = v^{(2)}, \end{cases}$$

which has the solution

$$\Psi^{(1)}(t,v) = \tilde{\Psi}(t)\left(v^{(1)} - \int_0^t \sigma(s)v^{(2)}\tilde{\Psi}^{-1}(s)ds + \int_0^t v^{(2)}\tilde{\Psi}^{-1}(s)dW(s)\right), \quad \Psi^{(2)}(t,v) = v^{(2)},$$

with

$$\tilde{\Psi}(t) = \exp\left\{-\frac{1}{2}\int_0^t \sigma^2(s)ds + \int_0^t \sigma(s)dW(s)\right\}.$$

Therefore, we get that

$$D\Psi(t,v) = \begin{pmatrix} 1+\tilde{\Psi}(t) & \tilde{\Psi}(t)\left(-\int_0^t \sigma(s)\tilde{\Psi}^{-1}(s)ds + \int_0^t \tilde{\Psi}^{-1}(s)dW(s)\right) \\ 0 & 1 \end{pmatrix}$$

and

$$D\Psi^{-1}(t,v) = \begin{pmatrix} \frac{1}{1+\tilde{\Psi}(t)} & -\frac{\tilde{\Psi}(t)}{1+\tilde{\Psi}(t)}\left(-\int_0^t \sigma(s)\tilde{\Psi}^{-1}(s)ds + \int_0^t \tilde{\Psi}^{-1}(s)dW(s)\right) \\ 0 & 1 \end{pmatrix}$$

Using in fact that $\tilde{\Psi}(t) > 0$ and applying the Frobenius norm $\|\cdot\|_F$, we obtain ω-a.e.

$$\begin{aligned} \|D\Psi^{-1}(t,v)\|_F &= \operatorname{tr}\left((D\Psi^{-1}(t,v))^\top D\Psi^{-1}(t,v)\right) \\ &\leqslant 2 + \tilde{\Psi}^2(t)\left(-\int_0^t \sigma(s)\tilde{\Psi}^{-1}(s)ds + \int_0^t \tilde{\Psi}^{-1}(s)dW(s)\right)^2 < \infty, \end{aligned}$$

for $t \in [0,T]$, $v \in \mathbb{R}^2$. By this Hypothesis (Flow2)(b) is fulfilled.

Moreover, a simple application of partial integration and Fubini's theorem together with the fact that $\bar{\Sigma}(-r) = 0$ shows that

$$\begin{aligned} \tilde{g}(\tilde{\varphi}_t) &= \int_{-r}^0 \bar{\Sigma}(s)\tilde{\varphi}_t(s)ds = \int_{-r}^0 \bar{\Sigma}(s)\tilde{\varphi}_0(s)ds + \int_0^t \left\{\Sigma(0)\tilde{\varphi}(u) - \int_{-r}^0 \bar{\Sigma}'(s)\tilde{\varphi}_u(s)ds\right\}du \\ &= \tilde{g}(\tilde{\varphi}_0) + \int_0^t h(t,\tilde{\varphi}(u),\tilde{\varphi}_u)du. \end{aligned}$$

It can be easlily checked that $h(t,\varphi) = \Sigma(0)\varphi(0) - \int_{-r}^{0} \bar{\Sigma}'(s)\varphi(s)ds$ satisfies the conditions given in (Flow2)(c) and (d). □

We are now ready to introduce two technical lemmas needed to prove our main results.

Lemma 1 *Assume that the solution to (4) exists and has a $C^{1,1}$-semiflow $X_t^s(\eta,\omega)$, $s \leqslant t$, $\omega \in \Omega$. Then, the following equality holds for a.e. $\omega \in \Omega$ and all directions $h \in M_2$:*

$$DX_t^s(\eta,\omega)[h] = (D^{\eta}x^s(t,\omega)[h],\, D^{\eta}x^s(t+\cdot,\omega)[h]) \in M_2.$$

Proof Note that $DX_t^s(\eta,\omega)[h] \in M_2$. Let $\{e_i\}_{i=0}^{\infty}$ be an orthonormal basis of M_2. Then,

$$\begin{aligned}
DX_t^s(\eta,\omega)[h] &= \sum_{i=0}^{\infty} \langle DX_t^s(\eta,\omega)[h], e_i\rangle_{M_2} e_i = \sum_{i=0}^{\infty} D\langle X_t^s(\eta,\omega), e_i\rangle_{M_2}[h]e_i \\
&= \sum_{i=0}^{\infty} D\left(x^s(t,\omega)e_i(0) + \int_{-r}^{0} x^s(t+u,\omega)e_i(u)du\right)[h]e_i \\
&= \sum_{i=0}^{\infty} \left(Dx^s(t,\omega)[h]e_i(0) + \int_{-r}^{0} Dx^s(t+u,\omega)[h]e_i(u)du\right)e_i \\
&= \sum_{i=0}^{\infty} \langle (D^{\eta}x^s(t,\omega)[h],\, D^{\eta}x^s(t+\cdot,\omega)[h]), e_i\rangle_{M_2} e_i \\
&= (D^{\eta}x^s(t,\omega)[h],\, D^{\eta}x^s(t+\cdot,\omega)[h]).
\end{aligned}$$

This finishes the proof. □

Lemma 2 *Let Hypotheses* (EU), (Flow) *and* (H) *be fulfilled. Then, for all $t \in [0,T]$, we have that $E[\|X_t^0(\eta)\|_{M_2}^4] < \infty$ and $E[\|DX_t^0(\eta)[h]\|_{M_2}^4] < \infty$ and the functions $t \mapsto E[\|X_t^0(\eta)\|_{M_2}^4]$ and $t \mapsto E[\|DX_t^0(\eta)[h]\|_{M_2}^4]$ are Lebesgue integrable, i.e.*

$$\int_0^T E[\|X_t^0(\eta)\|_{M_2}^4]dt < \infty, \tag{11}$$

$$\int_0^T E[\|DX_t^0(\eta)[h]\|_{M_2}^4]dt < \infty. \tag{12}$$

Proof To see this, observe that

$$\begin{aligned}\|X_s^0(\eta)\|_{M_2}^4 &= \Big(|x(s)|^2 + \int_{-r}^0 1_{(-\infty,0)}(s+u)|\eta(s+u)|^2 du + \int_{-r}^0 1_{[0,\infty)}(s+u)|x(s+u)|^2 du\Big)^2 \\ &\leqslant 3 \sup_{t\in[0,T]} |x(t)|^4 + 3\|\eta\|_{M_2}^4 + 3r^2 \sup_{t\in[0,T]} |x(t)|^4,\end{aligned}$$

and thus, for all $s \in [0, T]$

$$E[\|X_s^0(\eta)\|_{M_2}^4] \leqslant 3\|\eta\|_{M_2}^4 + 3(1+r^2)E[\sup_{t\in[0,T]} |x(t)|^4], \tag{13}$$

and

$$\int_0^T E[\|X_t^0(\eta)\|_{M_2}^4]dt \leqslant 3T\|\eta\|_{M_2}^4 + 3(1+r^2)TE[\sup_{t\in[0,T]} |x(t)|^4]. \tag{14}$$

To prove (11) it is then enough to show $E[\sup_{t\in[0,T]} |x(t)|^4] < \infty$. Therefore, consider first

$$\begin{aligned}&E[\sup_{t\in[0,T]} |x(t)|^4] \\ &= E\Big[\sup_{t\in[0,T]} \Big|\eta(0) + \int_0^t f(s, X_s^0(\eta))ds + \int_0^t g(s, X_s^0(\eta))dW(s)\Big|^4\Big] \\ &\leqslant E\Big[\sup_{t\in[0,T]} \Big(3\|\eta\|_{M_2}^2 + 3\Big(\int_0^t f(s, X_s^0(\eta))ds\Big)^2 + 3\Big(\int_0^t g(s, X_s^0(\eta))dW(s)\Big)^2\Big)^2\Big] \\ &\leqslant 27\|\eta\|_{M_2}^4 + 27T\int_0^T E[|f(s, X_s^0(\eta))|^4]ds + 27K_{BDG}E\Big[\Big(\int_0^T |g(s, X_s^0(\eta))|^2 ds\Big)^2\Big].\end{aligned}$$

Here we applied twice the fact that $(\sum_{i=1}^n a_i)^2 \leqslant n\sum_{i=1}^n |a_i|^2$ as well as Jensen's inequality, Fubini's theorem. Since the process $\int_0^{\cdot} g(s, X_s^0(\eta))dW(s)$ is a martingale (as a consequence of Theorem 1), we have also used the Burkholder-Davis-Gundy inequality (with the constant K_{BDG}).

By the linear growth condition (EU2) on f and g and (13), we have

$$\begin{aligned}|f(s, X_s^0(\eta))|^4 &\leqslant (C(1 + \|X_s^0(\eta)\|_{M_2}))^4 \leqslant 8C^4 + 8C^4\|X_s^0(\eta)\|_{M_2}^4 \\ &\leqslant 8C^4 + 24C^4\|\eta\|_{M_2}^4 + 24(1+r^2)\sup_{t\in[0,T]} |x(t)|^4,\end{aligned}$$

and the same applies to $|g(s, X_s^0(\eta))|^4$. Plugging this in the above estimates, we obtain

$$\begin{aligned}E[\sup_{t\in[0,T]} |x(t)|^4] &\leqslant 27\|\eta\|_{M_2}^4(1 + 24C^4T^2(1+K_{BDG})) + 216C^4T^2(1+K_{BDG}) \\ &\quad + 648(1+r^2)C^4T^2(1+K_{BDG})E[\sup_{t\in[0,T]} |x(t)|^4],\end{aligned}$$

which is

$$(1 - T^2k_1^2)E[\sup_{t\in[0,T]}|x(t)|^4] \leqslant k_2,$$

where

$$k_1 := \sqrt{648(1+r^2)C^4(1+K_{BDG})} \text{ and}$$
$$k_2 := 27\|\eta\|^4_{M_2}(1+24C^4T^2(1+K_{BDG})) + 216C^4T^2(1+K_{BDG}).$$

Then we distinguish two cases.

Case 1: $T < \frac{1}{k_1}$. Then $E[\sup_{t\in[0,T]}|x(t)|^4] \leqslant \frac{k_2}{(1-T^2k_1^2)}$ Hence, by (13) and (14) we have that (11) holds.

Case 2: $T \geqslant \frac{1}{k_1}$. In this case, choose $0 < T_1 < T_2 < \cdots < T_n = T$ for some finite n such that

$$T_1 < \frac{1}{k_1} \text{ and } T_i - T_{i-1} < \frac{1}{k_1}, \quad i = 2, \ldots, n.$$

By the semiflow property, we have $X^{T_1}_{T_2}(X^0_{T_1}(\eta)) = X^0_{T_2}(\eta)$, so we can solve the SFDE on $[0, T_1]$, and by Case 1 we have

$$E[\sup_{t\in[0,T_1]}|x(t)|^4] < \infty \text{ and } \int_0^{T_1} E[\|X^0_t(\eta)\|^4_{M_2}]dt < \infty.$$

Then, we use $X^0_{T_1}(\eta)$ as a new starting value and solve the equation on $[T_1, T_2]$. By the same steps as before, we obtain

$$E[\sup_{t\in[T_1,T_2]}|x(t)|^4]$$
$$\leqslant \frac{27E[\|X^0_{T_1}(\eta)\|^4_{M_2}](1+24(T_2-T_1)^2(1+K_{BDG})C^4)+216C^4(T_2-T_1)^2(1+K_{BDG})}{1-648(1+r^2)(T_2-T_1)^2(1+K_{BDG})C^4} < \infty,$$

and therefore,

$$\int_0^{T_2} E[\|X^0_t(\eta)\|^4_{M_2}]dt = \int_0^{T_1} E[\|X^0_t(\eta)\|^4_{M_2}]dt + \int_{T_1}^{T_2} E[\|X^0_t(\eta)\|^4_{M_2}]dt$$
$$\leqslant \int_0^{T_1} E[\|X^0_t(\eta)\|^4_{M_2}]dt + 3(T_2-T_1)E[\|X^0_{T_1}(\eta)\|^4_{M_2}]$$
$$+ 3(T_2-T_1)(1+r^2)E[\sup_{t\in[T_1,T_2]}|x(t)|^4] < \infty.$$

Iterating the argument, we conclude that for all $T \in (0, \infty)$, $E[\sup_{t\in[0,T]} |x(t)|^4] < \infty$ and $\int_0^T E[\|X_t^0(\eta)\|_{M_2}^4]dt < \infty$, that is (11) holds.

In order to prove (12), we define the process

$$y(t) := \begin{pmatrix} x(t) \\ Dx(t)[h] \end{pmatrix}, \ t \in [-r, T]$$

and the corresponding short-hand notation

$$\mathscr{Y}(t, \eta, h) = (X_t^0(\eta), DX_t^0(\eta)[h]) \in \mathbb{M}_2 \times \mathbb{M}_2$$

The process y satisfies the SFDE

$$\begin{aligned} y(t) &= \begin{pmatrix} \eta(0) \\ h(0) \end{pmatrix} + \int_0^t \hat{f}(s, \mathscr{Y}(s, \eta, h))ds + \int_0^t \hat{g}(s, \mathscr{Y}(s, \eta, h))dW(s), \\ y_0 &= (\eta, h) \end{aligned} \tag{15}$$

where, for $(\varphi, \psi)^\top \in M_2 \times M_2$,

$$\hat{f}(s, (\varphi, \psi)) := \begin{pmatrix} f(s, \varphi) \\ Df(s, \varphi)[\psi] \end{pmatrix}, \quad \hat{g}(s, (\varphi, \psi)) := \begin{pmatrix} g(s, \varphi) \\ Dg(s, \varphi)[\psi] \end{pmatrix}.$$

Thanks to Lemma 1, we recognize Eq. (15) as being of type (3). In fact, we can identify the $M_2 \times M_2$-valued random variable $(X_s^0(\eta), DX_s^0(\eta)[h])$ with the $M_2([-r, 0], \mathbb{R}^{2d})$-valued random variable $(y(s), y(s + \cdot))$. Using (H) it is now easy to check that $\hat{f}$ and $\hat{g}$ fulfil Hypothesis (EU), which are sufficient for the existence and uniqueness of a solution.

We can therefore argue exactly as in the proof of (11) and obtain that

$$E[\|\mathscr{Y}(t, \eta, h)\|_{M_2\times M_2}^4] < \infty \, \forall t \in [0, T] \text{ and } \int_0^T E[\|\mathscr{Y}(t, \eta, h)\|_{M_2\times M_2}^4]dt < \infty.$$

Moreover, since

$$\begin{aligned} \|\mathscr{Y}(t, \eta, h)\|_{M_2\times M_2}^4 &= \left(|y(t)|_{\mathbb{R}^{2d}}^2 + \int_{-r}^0 |y(t + u)|_{\mathbb{R}^{2d}}^2 du \right)^2 \\ &= \left(|x(t)|_{\mathbb{R}^d}^2 + |Dx(t)[h]|_{\mathbb{R}^d}^2 + \int_{-r}^0 \left(|x(t + u)|_{\mathbb{R}^d}^2 + |Dx(t + u)[h]|_{\mathbb{R}^d}^2 \right) du \right)^2 \\ &= \left(\|X_t^0(\eta)\|_{M_2}^2 + \|DX_t^0(\eta)[h]\|_{M_2}^2 \right)^2 \geqslant \|DX_t^0(\eta)[h]\|_{M_2}^4, \end{aligned}$$

we conclude that $E[\|DX_t^0(\eta)[h]\|_{M_2}^4] < \infty$ for all $t \in [0, T]$ and (12) holds. □

Our aim in the study of (10) is to give a formula for $\partial_h p(\eta)$ that avoids differentiating the function Φ. Our approach consists in randomizing the initial condition η and in finding a relationship between the Fréchet derivative $DX_T^0(\eta)$ applied to a direction $h \in \mathbb{M}_2$ and the Malliavin derivative of the X_T^0 with the randomized starting condition.

3.1 Randomization of the Initial Condition and the Malliavin Derivative

Following the approaches in, e.g. [30] or [34], we define an isonormal Gaussian process $\mathbb{B}$ on $L^2([-r, 0], \mathbb{R})$, independent of the m-dimensional Wiener process W that drives the SFDE (3). Without loss of generality, we can assume that W and $\mathbb{B}$ are defined on indepentent probability spaces $(\Omega^W, \mathscr{F}^W, P^W)$ and $(\Omega^{\mathbb{B}}, \mathscr{F}^{\mathbb{B}}, P^{\mathbb{B}})$ and that $(\Omega, \mathscr{F}, P) = (\Omega^W \times \Omega^{\mathbb{B}}, \mathscr{F}^W \otimes \mathscr{F}^{\mathbb{B}}, P^W \otimes P^{\mathbb{B}})$. From now on we shall work in $\Omega = \Omega^W \times \Omega^{\mathbb{B}}$. Hence, we correspondingly transfer the notation introduced so far to this case. However, we shall deal with the Malliavin and Skorohod calculus only with respect to $\mathbb{B}$. In fact, for the isonormal Gaussian process $\mathbb{B}$ we define the Malliavin derivative operator $\mathscr{D}$ and the Skorohod integral operator δ as performed in e.g. [30] or [34].

For immediate use, we give the link between the Malliavin derivative of a segment and the segment of Malliavin derivatives.

Lemma 3 *If $X_t^0(\eta) = ({}^\eta x(t), {}^\eta x_t) \in \mathbb{M}_2$ is Malliavin differentiable for all $t \geqslant 0$, then, for all $s \geqslant 0$, $\mathscr{D}_s {}^\eta x_t = \{\mathscr{D}_s {}^\eta x(t+u), u \in [-r, 0]\}$ and $\mathscr{D}_s X_t^0(\eta) = (\mathscr{D}_s {}^\eta x(t), \mathscr{D}_s {}^\eta x(t+\cdot)) \in \mathbb{M}_2$.*

Proof The proof follows the same lines as the proof of Lemma 1. □

Here below we discuss the chain rule for the Malliavin derivative in $\mathbb{M}_2$. This leads to the study of the interplay between Malliavin derivatives and Fréchet derivatives.

We recall that, if DX_T^0 is bounded, i.e. for all $\omega = (\omega^W, \omega^{\mathbb{B}}) \in \Omega$, $\sup_{\eta \in \mathbb{M}_2} \|DX_T^0(\eta(\omega), \omega^W)\| < \infty$, the chain rule in [34, Proposition 3.8] gives

$$\mathscr{D}_s X_T^0(\eta(\omega^W, \omega^{\mathbb{B}}), \omega^W) = DX_T^0(\eta(\omega^W, \omega^{\mathbb{B}}), \omega^W)[\mathscr{D}_s \eta(\omega^W, \omega^{\mathbb{B}})],$$

as the Malliavin derivative only acts on $\omega^{\mathbb{B}}$. We need an analoguous result also in the case when DX_T^0 is possibly unbounded. To show this, we apply $\mathscr{D}_s$ directly to the dynamics given by Eq. (3).

Theorem 3 *Let $X_\cdot^0(\eta) \in L^2(\Omega; M_2([-r, T], \mathbb{R}^d))$ be the stochastic semiflow associated to the solution of (3). Let Hypotheses* (EU), (Flow) *and* (H) *be fulfilled. Then we have*

$$\mathscr{D}_s X_T^0(\eta) = DX_T^0(\eta)[\mathscr{D}_s \eta] \quad (\omega, s) - a.e. \tag{16}$$

Proof To show this, we apply $\mathscr{D}_s$ directly to the dynamics given by Eq. (3). Doing this, we get, by definition of the operator ρ_0 and Lemma 3, for a.e. $\omega \in \Omega$

$$\rho_0(\mathscr{D}_s X_T^0(\eta)) = \mathscr{D}_s\, {}^{\eta}x(t) = \begin{cases} \mathscr{D}_s\eta(0) + \int_0^t Df(u, X_u^0(\eta))[\mathscr{D}_s X_u^0(\eta)]du \\ + \int_0^t Dg(u, X_u^0(\eta))[\mathscr{D}_s X_u^0(\eta)]dW(u), \quad t \in [0,T], \\ \mathscr{D}_s\eta(t), \quad t \in [-r, 0]. \end{cases} \tag{17}$$

Define the processes

$$y(t) := \begin{pmatrix} {}^{\eta}x(t) \\ D\,{}^{\eta}x(t)[\mathscr{D}_s\eta] \end{pmatrix}, \; z(t) := \begin{pmatrix} {}^{\eta}x(t) \\ \mathscr{D}_s\,{}^{\eta}x(t) \end{pmatrix}.$$

From the proof of Lemma 2 we know that y satisfies the SFDE

$$\begin{cases} y(t) &= \begin{pmatrix} \eta(0) \\ \mathscr{D}_s\eta(0) \end{pmatrix} + \int_0^t \hat{f}(u, y(u), y_u)du + \int_0^t \hat{g}(u, y(u), y_u)dW(u), \\ y_0 &= (\eta, \mathscr{D}_s\eta), \end{cases}$$

with the functions $\hat{f}$ and $\hat{g}$ as in the proof of Lemma 2. Moreover, by (17) and Lemma 3, it holds that z satisfies the SFDE

$$\begin{cases} z(t) &= \begin{pmatrix} \eta(0) \\ \mathscr{D}_s\eta(0) \end{pmatrix} + \int_0^t \hat{f}(u, z(u), z_u)du + \int_0^t \hat{g}(u, z(u), z_u)dW(u), \\ z_0 &= (\eta, \mathscr{D}_s\eta). \end{cases}$$

Comparing those two SFDEs, it follows that $y = z$ in $L^2(\Omega, M_2([-r,T], \mathbb{R}^d))$. Therefore,

$$E\left[\int_0^T \|y_t - z_t\|_{M_2}^2 dt\right] = E\left[\int_0^T |y(t) - z(t)|^2 + \int_{-r}^0 |y(t+u) - z(t+u)|^2 du dt\right]$$
$$\leqslant (1+r)\|y - z\|_{L^2(\Omega, M_2([-r,T],\mathbb{R}^d))} = 0,$$

which implies that $\|y_t - z_t\|_{M_2} = 0$ for a.e. $(\omega, t) \in \Omega \times [0,T]$. □

We now introduce the randomization of the initial condition. For this we consider an $\mathbb{R}$-valued functional ξ of $\mathbb{B}$, non-zero P-a.s. In particular, ξ is a random variable independent of W. Choose ξ to be Malliavin differentiable with respect to $\mathbb{B}$ with $\mathscr{D}_s\xi \neq 0$ for almost all (ω, s). Furthermore, let $\eta, h \in \mathbb{M}_2$ be random variables on Ω^W, i.e. $\eta(\omega) = \eta(\omega^W)$, $h(\omega) = h(\omega^W)$. We write $\eta, h \in \mathbb{M}_2(\Omega^W)$, where

$\mathbb{M}_2(\Omega^W)$ denotes the space of random variables in $\mathbb{M}_2$ that only depend on $\omega^W \in \Omega^W$. Here η plays the role of the "true" (i.e. not randomized) initial condition and h plays the role of the direction in which we later are going to differentiate. For simpler notation, we define $\tilde{\eta} := \eta - h$.

Corollary 2 *Let $X_\cdot^0(\tilde{\eta} + \lambda\xi h) \in L^2(\Omega; M_2([-r, T], \mathbb{R}^d))$ be the stochastic semiflow associated to the solution of (3) with initial condition $\tilde{\eta} + \lambda\xi h \in \mathbb{M}_2$, where $\lambda \in \mathbb{R}$. Let Hypotheses* (EU), (Flow) *and* (H) *be fulfilled. Then we obtain*

$$\mathscr{D}_s X_T^0(\tilde{\eta}(\omega^W) + \lambda\xi(\omega^{\mathbb{B}})h(\omega^W)) = DX_T^0(\tilde{\eta}(\omega^W) + \lambda\xi(\omega^{\mathbb{B}})h(\omega^W))[\lambda\mathscr{D}_s\xi(\omega^{\mathbb{B}})h(\omega^W)] \tag{18}$$

(ω, s)*-a.e. In short hand notation:*

$$\mathscr{D}_s X_T^0(\tilde{\eta} + \lambda\xi h) = DX_T^0(\tilde{\eta} + \lambda\xi h)[\lambda\mathscr{D}_s\xi h]. \tag{19}$$

We are now giving a derivative free representation of the expectation of the Fréchet derivative of $\Phi \circ X_T^0$ at η in direction h in terms of a Skorohod integral. This representation will later be used to get a representation for the derivative of $p(\eta)$ in direction h.

Theorem 4 *Let Hypotheses* (EU), (Flow) *and* (H) *be satisfied and let Φ be Fréchet differentiable. Furthermore, let $a \in L^2([-r, 0], \mathbb{R})$ be such that $\int_{-r}^0 a(s)ds = 1$. If $a(\cdot)\xi/\mathscr{D}_\cdot\xi$ is Skorohod integrable and if the Skorohod integral below and its evaluation at $\lambda = \frac{1}{\xi} \in \mathbb{R}$ are well defined then following relation holds*

$$E[D(\Phi \circ X_T^0)(\eta)[h]] = -E\left[\left\{\delta\Big(\Phi(X_T^0(\tilde{\eta} + \lambda\xi h))a(\cdot)\frac{\xi}{\mathscr{D}_\cdot\xi}\Big)\right\}\Big|_{\lambda=\frac{1}{\xi}}\right]. \tag{20}$$

Proof First of all we can see that, by Corollary 2, we have the relation

$$\mathscr{D}_s X_T^0(\tilde{\eta} + \lambda\xi h) = DX_T^0(\tilde{\eta} + \lambda\xi h)[\lambda\mathscr{D}_s\xi h] \quad (\omega, s) - a.e.$$

Multiplication with $\frac{\xi}{\mathscr{D}_s\xi}$ yields

$$\frac{\xi}{\mathscr{D}_s\xi}\mathscr{D}_s X_T^0(\tilde{\eta} + \lambda\xi h) = DX_T^0(\tilde{\eta} + \lambda\xi h)[h]\lambda\xi \quad (\omega, s) - a.e. \tag{21}$$

For the above, we recall that $\mathscr{D}_s\xi \neq 0$ a.e. Since the right-hand side in (21) is defined ω-wise, the evaluation at $\lambda = \frac{1}{\xi}$ yields $DX_T^0(\tilde{\eta} + h)[h]$. Summarising, we have

$$\left\{\frac{\xi}{\mathscr{D}_s\xi}\mathscr{D}_s X_T^0(\tilde{\eta} + \lambda\xi h)\right\}\Big|_{\lambda=\frac{1}{\xi}} = DX_T^0(\tilde{\eta} + \lambda\xi h)[h]\lambda\xi\Big|_{\lambda=\frac{1}{\xi}} = DX_T^0(\tilde{\eta} + h)[h] = DX_T^0(\eta)[h]$$

Multiplying with $1 = \int_{-r}^{0} a(s)ds$ and applying the chain rule, together with the fact that $D\Phi(X_T^0(\eta))$ is defined pathwise, we obtain

$$\begin{aligned} E[D(\Phi \circ X_T^0)(\eta)[h]] &= E\left[D\Phi(X_T^0(\eta))DX_T^0(\eta)[h]\right] = E\left[\int_{-r}^{0} D\Phi(X_T^0(\eta))DX_T^0(\eta)[h]a(s)ds\right] \\ &= E\left[\left\{\int_{-r}^{0} D\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))\mathscr{D}_s X_T^0(\tilde{\eta}+\lambda\xi h)a(s)\frac{\xi}{\mathscr{D}_s\xi}ds\right\}\Big|_{\lambda=\frac{1}{\xi}}\right] \\ &= E\left[\left\{\int_{-r}^{0} \mathscr{D}_s\{\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))\}a(s)\frac{\xi}{\mathscr{D}_s\xi}ds\right\}\Big|_{\lambda=\frac{1}{\xi}}\right]. \end{aligned}$$

The partial integration formula for the Skorohod integral, see e.g. [30, Prop. 1.3.3], yields

$$\begin{aligned} E[D(\Phi \circ X_T^0)(\eta)[h]] &= E\left[\left\{\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))\delta\Big(a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big) - \delta\Big(\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big)\right\}\Big|_{\lambda=\frac{1}{\xi}}\right] \\ &= E\left[\Phi(X_T^0(\eta))\delta\Big(a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big) - \left\{\delta\Big(\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big)\right\}\Big|_{\lambda=\frac{1}{\xi}}\right]. \end{aligned}$$

Observe that $\Phi(X_T^0(\eta))$ is $\mathscr{F}^W$-measurable and $\delta\Big(a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big)$ is $\mathscr{F}^{\mathbb{B}}$-measurable. The result follows from the independence of W and $\mathbb{B}$ and $E\left[\delta\Big(a(\cdot)\frac{\xi}{\mathscr{D}\xi}\Big)\right] = 0$. □

Remark 1 As for a numerically tractable approximation of the stochastic integral in the above formula we refer to [30, Section 3.1].

Proposition 1 *Define* $u(s,\lambda) := \Phi(X_T^0(\tilde{\eta}+\lambda\xi h))a(s)\frac{\xi}{\mathscr{D}_s\xi}$, $s \in [-r,0]$, $\lambda \in \mathbb{R}$. *Assume that the Skorohod integral* $\delta(u(\cdot,\lambda))$ *exists for all* $\lambda \in \mathbb{R}$. *If for all* $\Lambda > 0$ *there exists a* $C > 0$ *such that for all* $\lambda_1, \lambda_2 \in \overline{\operatorname{supp}\xi^{-1}}$, $|\lambda_1|, |\lambda_2| < \Lambda$:

$$\|u(\cdot,\lambda_1) - u(\cdot,\lambda_2)\|^2_{L^2(\Omega\times[-r,0])} + \|\mathscr{D}(u(\cdot,\lambda_1) - u(\cdot,\lambda_2))\|^2_{L^2(\Omega\times[-r,0]^2)} < C|\lambda_1 - \lambda_2|^2,$$

then the evaluation $\delta(u(\cdot,\lambda))|_{\lambda=\frac{1}{\xi}}$ *is well defined.*

Proof The Skorohod integral $\delta(u(\cdot,\lambda))$ is an element of $L^2(\Omega,\mathbb{R})$. From

$$\|\delta(u(\cdot,\lambda))\|^2_{L^2(\Omega,\mathbb{R})} \leqslant \|u(\cdot,\lambda)\|^2_{L^2(\Omega\times[-r,0],\mathbb{R})} + \|\mathscr{D}u(\cdot,\lambda)\|^2_{L^2(\Omega\times[-r,0],\mathbb{R})}$$

(see [30, eq. (1.47) Proof of Prop. 1.3.1]), under the assumptions above and by means of Kolmogorov's continuity theorem, we can see that the process

$$Z : \Omega \times \overline{\operatorname{supp}\xi^{-1}} \to L^2(\Omega,\mathbb{R}),\ \lambda \mapsto \delta(u(\cdot,\lambda))$$

has a continuous version. Applying this continuous version, the evaluation at the random variable $\frac{1}{\xi}$ is well defined:

$$\delta(u(\cdot,\lambda))(\omega)|_{\lambda=\frac{1}{\xi}} := Z(\omega,\lambda)|_{\lambda=\frac{1}{\xi}} := Z(\omega,\frac{1}{\xi}(\omega)).$$

Hence we conclude the proof. □

3.2 Representation Formula for Delta Under a Suitable Choice of the Randomization

A particularly interesting choice of randomization is $\xi = \exp(\mathbb{B}(1_{[-r,0]}))$, since in this case, $\mathscr{D}_s\xi = \xi$ for all $s \in [-r,0]$ and

$$\begin{aligned}\|\delta(u(\cdot,\lambda_1)) - \delta(u(\cdot,\lambda_2))\|^2_{L^2(\Omega)} &\\ \leqslant \|a\|^2_{L^2([-r,0])}(\|\Phi(X_T^0(\tilde{\eta}+\lambda_1\xi h)) - \Phi(X_T^0(\tilde{\eta}+\lambda_2\xi h))\|^2_{L^2(\Omega)} &\\ + \|\mathscr{D}\{\Phi(X_T^0(\tilde{\eta}+\lambda_1\xi h)) - \Phi(X_T^0(\tilde{\eta}+\lambda_2\xi h))\}\|^2_{L^2(\Omega\times[-r,0])}).&\end{aligned} \tag{22}$$

In this setup, let the following hypotheses be fulfilled:

Hypotheses (A): Φ is globally Lipschitz with Lipschitz constant L_Φ and C^1. The Fréchet derivative $D\Phi$ is globally Lipschitz with Lipschitz constant $L_{D\Phi}$.

A more general payoff function Φ will be considered in the next subsection. Recall that $p(\eta) = E[\Phi(X_T^0(\eta))]$ and the sensitivity with respect to the initial path, the Delta, in direction $h \in M_2$ is $\partial_h p(\eta) := \frac{d}{d\varepsilon}p(\eta+\varepsilon h)|_{\varepsilon=0}$.

Lemma 4 *Under Hypotheses* (EU), (Flow), (H) *and* (A), *we have*

$$\partial_h p(\eta) = E[D(\Phi\circ X_T^0)(\eta)[h]].$$

Proof By definition of the directional derivative, we have

$$\partial_h p(\eta) = \lim_{\varepsilon\to 0} E\Big[\frac{1}{\varepsilon}(\Phi(X_T^0(\eta+\varepsilon h)) - \Phi(X_T^0(\eta)))\Big] = \lim_{\varepsilon\to 0} E[F_\varepsilon],$$

where $F_\varepsilon(\omega) = \frac{1}{\varepsilon}(\Phi(X_T^0(\eta+\varepsilon h,\omega)) - \Phi(X_T^0(\eta,\omega))) \to D(\Phi\circ X_T^0(\omega))(\eta)[h]$ a.s. since the Fréchet derivative of $\Phi\circ X_T^0$ in η is defined for a.e. ω. Moreover,

$$|F_\varepsilon(\omega)| = \frac{|\Phi(X_T^0(\eta+\varepsilon h,\omega)) - \Phi(X_T^0(\eta,\omega))|}{\varepsilon} \leqslant L_\Phi \frac{\|X_T^0(\eta+\varepsilon h,\omega) - X_T^0(\eta,\omega)\|_{M_2}}{\varepsilon} =: G_\varepsilon(\omega).$$

So if we can find $G \in L^1(\Omega, P)$ s.t. $G_\varepsilon \to g$ in L^1-convergence as $\varepsilon \to 0$, we would have that $F_\varepsilon \to D(\Phi \circ X_T^0)(\eta)[h]$ in L^1-convergence by Pratt's lemma (see [33, Theorem 1]). This would conclude the proof.

Observe that, by the continuity of the norm $\|\cdot\|_{M_2}$ and the ω-wise Fréchet differentiability of X_T^0 in η, we have that

$$G_\varepsilon(\omega) \to L_\Phi \|DX_T^0(\eta, \omega)[h]\|_{M_2}, \ \omega\text{-a.e.}$$

Let $G(\omega) := L_\Phi \|DX_T^0(\eta, \omega)[h]\|_{M_2}$. By Lemma 2, $G \in L^1(\Omega, \mathbb{R})$. We apply Vitali's theorem (see [35, Theorem 16.6]) to show that the convergence $G_\varepsilon \to G$ holds in L^1. This means that we have to prove that the family $\{G_\varepsilon\}_{\varepsilon \in (-\delta, \delta)}$ for some $\delta > 0$ is uniformly integrable. To show that, we will proceed in two steps:

(1) Prove that $\|G_\varepsilon\|_{L^2(\Omega)} \leqslant K$ for some constant K uniformly in ε.
(2) Show that this implies that $\{g_\varepsilon\}_{\varepsilon \in (-\delta, \delta)}$ is uniformly integrable.

Step (1): By Lemma 2, it holds that for each fixed $\varepsilon \in (-\delta, \delta)\backslash\{0\}$, the function $s \mapsto E[(\frac{1}{\varepsilon}\|X_s^0(\eta + \varepsilon h, \omega) - X_s^0(\eta, \omega)\|_{M_2})^2]$ is integrable on $[0, T]$. Now, making use of Jensen's inequality, Fubini's theorem and the Burkholder-Davis-Gundy inequality,

$$\begin{aligned}
&E\left[\left(\frac{1}{\varepsilon}\|X_T^0(\eta + \varepsilon h) - X_T^0(\eta)\|_{M_2}\right)^2\right] \\
&= E\Bigg[\frac{1}{\varepsilon^2}\Big(\big|\varepsilon h(0) + \int_0^T \Big(f(s, X_s^0(\eta + \varepsilon h)) - f(s, X_s^0(\eta))\Big) ds \\
&\quad + \int_0^T \Big(g(s, X_s^0(\eta + \varepsilon h)) - g(s, X_s^0(\eta))\Big) dW(s)\big|^2 + \int_{-r}^0 1_{(-\infty,0)}(T + u)|\varepsilon h(u)|^2 du \\
&\quad + \int_{-r}^0 1_{[0,\infty)}(T + u)\big|\varepsilon h(0) + \int_0^{T+u} \Big(f(s, X_s^0(\eta + \varepsilon h)) - f(s, X_s^0(\eta))\Big) ds \\
&\quad + \int_0^{T+u} \Big(g(s, X_s^0(\eta + \varepsilon h)) - g(s, X_s^0(\eta))\Big) dW(s)\big|^2 du\Big)\Bigg] \\
&\leqslant 3|h(0)|^2 + \frac{3T}{\varepsilon^2}\int_0^T E[|f(s, X_s^0(\eta + \varepsilon h)) - f(s, X_s^0(\eta))|^2] ds \\
&\quad + \frac{3}{\varepsilon^2}\int_0^T E[|g(s, X_s^0(\eta + \varepsilon h)) - g(s, X_s^0(\eta))|^2] ds + \int_{-r}^0 |h(u)|^2 du + 3r|h(0)|^2 \\
&\quad + \frac{3}{\varepsilon^2}\int_{-r}^0 1_{[0,\infty)}(T + u)\int_0^{T+u} (T + u) E[|f(s, X_s^0(\eta + \varepsilon h)) - f(s, X_s^0(\eta))|^2] ds du \\
&\quad + \frac{3}{\varepsilon^2}\int_{-r}^0 1_{[0,\infty)}(T + u)\int_0^{T+u} E[|g(s, X_s^0(\eta + \varepsilon h)) - g(s, X_s^0(\eta))|^2] ds du
\end{aligned}$$

$$\leqslant 3(1+r)\|h\|_{M_2}^2 + (3+r)T\int_0^T \frac{1}{\varepsilon^2}E[|f(s,X_s^0(\eta+\varepsilon h)) - f(s,X_s^0(\eta))|^2]ds$$
$$+ (3+r)\int_0^T \frac{1}{\varepsilon^2}E[|g(s,X_s^0(\eta+\varepsilon h)) - g(s,X_s^0(\eta))|^2]ds$$
$$\leqslant 3(1+r)\|h\|_{M_2}^2 + (3+r)(L_g^2 + TL_f^2)\int_0^T E\left[\left(\frac{1}{\varepsilon}\|X_s^0(\eta+\varepsilon h) - X_s^0(\eta)\|_{M_2}\right)^2\right]ds.$$

It follows from Gronwall's inequality that

$$\|g_\varepsilon\|_{L^2(\Omega)}^2 = L_\Phi^2 E\left[\left(\frac{1}{\varepsilon}\|X_T^0(\eta+\varepsilon h) - X_T^0(\eta)\|_{M_2}\right)^2\right]$$
$$\leqslant 3L_\Phi^2(1+r)\|h\|_{M_2}^2 e^{(3+r)(TL_g^2+T^2L_f^2)} =: K^2.$$

Step (2): Fix $\delta > 0$. Then, by Hölder's inequality and Markov's inequality

$$\lim_{M\to\infty}\sup_{|\varepsilon|<\delta} E[|g_\varepsilon|1_{\{|g_\varepsilon|>M\}}] \leqslant \lim_{M\to\infty}\sup_{|\varepsilon|<\delta}\|g_\varepsilon\|_{L^2(\Omega)}\sqrt{P(|g_\varepsilon| > M)}$$
$$\leqslant \lim_{M\to\infty}\sup_{|\varepsilon|<\delta}\frac{\|g_\varepsilon\|_{L^2(\Omega)}^2}{M} \leqslant \lim_{M\to\infty}\frac{K^2}{M} = 0,$$

i.e. the family $\{g_\varepsilon\}_{\varepsilon\in(-\delta,\delta)}$ is uniformly integrable. □

With this result, we can give a derivative free representation formula for the directional derivatives of $p(\eta)$.

Theorem 5 *Let Hypotheses* (EU), (Flow), (H) *and* (A) *be fulfilled. Let* $a \in L^2([-r,0],\mathbb{R})$ *be such that* $\int_{-r}^0 a(s)ds = 1$ *and let* $\xi = \exp(\mathbb{B}(1_{[-r,0]}))$. *Then the directional derivatives of p have representation*

$$\partial_h p(\eta) = -E\left[\left\{\delta\Big(\Phi(X_T^0(\tilde{\eta}+\lambda\xi h))a(\cdot)\Big)\right\}\Big|_{\lambda=\frac{1}{\xi}}\right]. \tag{23}$$

We remark that different choices of function a may lead to different statistical properties of the estimator under the expectation sign in (23).

To prove the theorem, we need the following lemma:

Lemma 5 *Assume* (H) *and* (A) *and* $\xi = \exp(\mathbb{B}(1_{[-r,0]}))$. *For any* $\Lambda > 0$ *there exists a* $C > 0$ *such that, for all* $|\lambda_1|, |\lambda_2| < \Lambda$, *we have*

(i) $E[\|X_T^0(\tilde{\eta}+\lambda_1\xi h) - X_T^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]^{\frac{1}{2}} \leqslant C|\lambda_1-\lambda_2|^2$

(ii) $E[\|DX_T^0(\tilde{\eta}+\lambda_1\xi h)[\lambda_1\xi h]\|_{M_2}^4]^{\frac{1}{2}} \leqslant C|\lambda_1|^2$

(iii) $E[\|DX_T^0(\tilde{\eta}+\lambda_1\xi h)[\lambda_1\xi h] - DX_T^0(\tilde{\eta}+\lambda_2\xi h)[\lambda_2\xi h]\|_{M_2}^2] \leqslant C|\lambda_1-\lambda_2|^2$.

Proof See Appendix. □

Proof (of Theorem 5) By Lemma 4, we know that we can interchange the directional derivative with the expectation. We shall prove that the Skorohod integral in (23) is well defined. For this we apply Proposition 1 and use (22).

Let $\lambda_1, \lambda_2 \in \mathbb{R}$, $|\lambda_1|, |\lambda_2| < \Lambda$. Because of Hypotheses (A), and by Lemma 5, we have that

$$\begin{aligned}
&\|\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)) - \Phi(X_T^0(\tilde{\eta} + \lambda_2 \xi h))\|^2_{L^2(\Omega)} \\
&\quad \leqslant L_\Phi^2 E[\|X_T^0(\tilde{\eta} + \lambda_1 \xi h) - X_T^0(\tilde{\eta} + \lambda_2 \xi h)\|^2_{M_2}] \\
&\quad \leqslant L_\Phi^2 E[\|X_T^0(\tilde{\eta} + \lambda_1 \xi h) - X_T^0(\tilde{\eta} + \lambda_2 \xi h)\|^4_{M_2}]^{\frac{1}{2}} \\
&\quad \leqslant L_\Phi^2 C|\lambda_1 - \lambda_2|^2.
\end{aligned}$$

On the other hand, the chain rule for the Malliavin derivative, the property $\mathscr{D}_s \xi = \xi$, and the fact that for two linear operators A_1 and A_2 it holds $A_1 x_1 - A_2 x_2 = (A_1 - A_2)x_1 + A_2(x_1 - x_2)$ together with the property $|a+b|^2 \leqslant 2|a|^2 + 2|b|^2$ yield

$$\begin{aligned}
&|\mathscr{D}_s\{\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)) - \Phi(X_T^0(\tilde{\eta} + \lambda_2 \xi h))\}|^2 \\
&\leqslant 2|(D\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)) - D\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)))[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h]]|^2 \\
&\quad + 2|D\Phi(X_T^0(\tilde{\eta} + \lambda_2 \xi h)[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h] - DX_T^0(\tilde{\eta} + \lambda_2 \xi h)[\lambda_2 \xi h]]|^2 \\
&\leqslant 2\|D\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)) - D\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h))\|^2 \|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h]]\|^2_{M_2} \\
&\quad + 2\|D\Phi(X_T^0(\tilde{\eta} + \lambda_2 \xi h)\|^2 \|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h] - DX_T^0(\tilde{\eta} + \lambda_2 \xi h)[\lambda_2 \xi h]]\|^2_{M_2} \\
&\leqslant 2L_{D\Phi}^2 \|X_T^0(\tilde{\eta} + \lambda_1 \xi h) - X_T^0(\tilde{\eta} + \lambda_1 \xi h)\|^2_{M_2} \|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h]]\|^2_{M_2} \\
&\quad + 2L_\Phi^2 \|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h] - DX_T^0(\tilde{\eta} + \lambda_2 \xi h)[\lambda_2 \xi h]]\|^2_{M_2},
\end{aligned}$$

where we used Hypothesis (A) in the end. Taking expectations, applying Hölder's inequality and Lemma 5 we finally get

$$\begin{aligned}
&\|\mathscr{D}\{\Phi(X_T^0(\tilde{\eta} + \lambda_1 \xi h)) - \Phi(X_T^0(\tilde{\eta} + \lambda_2 \xi h))\}\|^2_{L^2(\Omega \times [-r,0])} \\
&\leqslant 2L_{D\Phi}^2 E[\|X_T^0(\tilde{\eta} + \lambda_1 \xi h) - X_T^0(\tilde{\eta} + \lambda_1 \xi h)\|^4_{M_2}]^{\frac{1}{2}} E[\|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h]]\|^4_{M_2}]^{\frac{1}{2}} \\
&\quad + 2L_\Phi^2 E[\|[DX_T^0(\tilde{\eta} + \lambda_1 \xi h)[\lambda_1 \xi h] - DX_T^0(\tilde{\eta} + \lambda_2 \xi h)[\lambda_2 \xi h]]\|^2_{M_2}] \\
&\leqslant 2(L_{D\Phi}^2 C^2 |\lambda_1|^2 + L_\Phi^2 C)|\lambda_1 - \lambda_2|^2 \\
&= \mathscr{O}(1)|\lambda_1 - \lambda_2|^2.
\end{aligned}$$

Hence, Proposition 1 guarantees the existence of the evaluation of the Skorohod integral in $\lambda = \frac{1}{\xi}$. □

3.3 Generalization to a Larger Class of Payoff Functions

This section intends to generalize the findings of the previous section for a larger class of pay-off functions. In particular, this has interest in the context of finance where typical pay-off function are not smooth. The results of this section allow to treat, e.g., vanilla options such as European call and put options or even Asian options averaged on the past delay.

Instead of Hypothesis (A), assume now that the following holds:

Hypotheses (A'): The payoff function $\Phi : M_2 \to \mathbb{R}$ is convex, bounded from below and globally Lipschitz continuous with Lipschitz constant L_Φ.

Moreover, consider the *Moreau-Yosida approximations* $\Phi_n : M_2 \to \mathbb{R}$ given by

$$\Phi_n(x) := \inf_{y \in M_2} \left(\Phi(y) + \frac{n}{2} \|x - y\|^2_{M_2} \right). \tag{24}$$

The following lemma summarizes some well-known properties of the Moreau-Yosida approximations in our setup.

Lemma 6 *For Φ and Φ_n as above, the following holds*

(i) $\Phi_n(x) = \Phi(J_n(x)) + \frac{n}{2}\|x - J_n(x)\|^2_{M_2}$, $x \in M_2$, *where J_n is given by*

$$n(x - J_n(x)) \in \partial\Phi(J_n(x)) \text{ or, equivalently } J_n = \left(id + \frac{\partial \Phi}{n} \right)^{-1},$$

where $\partial\Phi(x)$ denotes the subdifferential of Φ in x and $\partial\Phi := \{(x, y) \in M_2 \times M_2 : y \in \partial\Phi(x)\}$.

(ii) *For all $x \in M_2$, $\Phi_n(x) \uparrow \Phi(x)$ and $J_n(x) \to x$, as $n \to \infty$.*

(iii) *Φ_n is Fréchet differentiable and, for all $x \in M_2$, it holds*

$$D\Phi_n(x) = n(x - J_n(x)) \in \partial\Phi(J_n(x))$$

and $D\Phi_n$ is Lipschitz.

(iv) *For each point $x \in dom(\partial\Phi)$,*

$$D\Phi_n(x) \to \partial^0\Phi(x),$$

where $\partial^0\Phi(x)$ denotes the element $y \in \partial\Phi(x)$ with minimal norm.

(v) *For each $x \in M_2$, it holds $\|D\Phi_n(x)\| \leqslant L_\Phi$.*

Proof

(i) See [10, p. 58] or [7, Theorem 3.24, p. 301], .

(ii) See Theorem 2.64 in [7, p. 229].

(iii) See [10, p. 58], and [7, Thm. 3.24].

(iv) See [7, Proposition 3.56 (c), equation (3.136), p. 354].

(v) By (iii), it holds $D\Phi_n(x) \in \partial\Phi(y_0)$ for some $y_0 \in M_2$ (namely $y_0 = J_n(x)$). By the definition of the subdifferential, it holds for every $g \in \partial\Phi(y_0)$ and every $h \in M_2$:

$$\langle g, h\rangle \leqslant \Phi(y_0 + h) - \Phi(y_0) \leqslant L_\Phi \|h\|_{M_2}.$$

In particular, $D\Phi_n(x)[h] \leqslant L_\Phi \|h\|_{M_2}$ and $D\Phi_n(x)[-h] \leqslant L_\Phi \|h\|_{M_2}$ and thus

$$|D\Phi_n(x)[h]| \leqslant L_\Phi \|h\|_{M_2}, \text{ which implies } \|D\Phi_n(x)\| \leqslant L_\Phi. \qquad \square$$

The following proposition shows that we can approximate $p(\eta)$ by a sequence $p_n(\eta)$ using the Moreau-Yosida approximations for the payoff functions.

Proposition 2 *Let the payoff function $\Phi : M_2 \to \mathbb{R}$ be of type* (A'). *Let Φ_n be given by* (24). *Set $p_n(\eta) := E[\Phi_n(X_T^0(\eta))]$ for $\eta \in \mathbb{M}_2$. Then, for all $\eta \in \mathbb{M}_2$, $p_n(\eta) \to p(\eta)$ as $n \to \infty$.*

Proof As Φ is bounded from below, without loss of generality, we can assume Φ being nonnegative. Then it is immediately clear from (24) that also Φ_n is nonnegative for every n. Since $\Phi_n(x) \uparrow \Phi(x)$ from Lemma 6 item (ii), we have that $\Phi_n(X_T^0(\eta, \omega)) \uparrow \Phi(X_T^0(\eta, \omega))$, for a.e. $\omega \in \Omega$, and, therefore, by the monotone convergence theorem

$$\lim_{n\to\infty} p_n(\eta) = \lim_{n\to\infty} E[\Phi_n(X_T^0(\eta))] = E[\Phi(X_T^0(\eta))] = p(\eta). \qquad \square$$

Definition 2 Let $\mathscr{X}$ and $\mathscr{Y}$ be Banach spaces. We call a function $F : \mathscr{X} \to \mathscr{Y}$ *LC directional differentiable* at $x \in \mathscr{X}$ if the directional derivative $\partial_h F(x)$ exists for each direction $h \in \mathscr{X}$ and defines a bounded linear operator from $\mathscr{X}$ to $\mathscr{Y}$.

Lemma 7 *For each point $x \in M_2$ at which Φ is LC directional differentiable, it holds*

$$D\Phi_n(x) \to \partial.\Phi(x).$$

Proof Since Φ is directional differentiable in x in each direction $h \in M_2$, we have that $\partial\Phi(x)$ is a singleton. In fact, by the definitions of the subdifferential and of the directional derivative, we have for all $h \in M_2$, on the one side we have

$$\partial_h \Phi(x) = \lim_{\varepsilon\to 0} \frac{\Phi(y_0 + \varepsilon h) - \Phi(y_0)}{\varepsilon} \geqslant \langle g, h\rangle,$$

for all $g \in \partial\Phi(x)$, and on the other side

$$\partial_h\Phi(x) = -\partial_{-h}\Phi(x) \leqslant -\langle g, -h\rangle = \langle g, h\rangle,$$

for all $g \in \partial\Phi(x)$. Namely, $\partial\Phi(x) = \{\partial.\Phi(x)\}$. By Lemma 6(iv) we have that $D\Phi_n(x) \to \partial^0\Phi(x) = \partial.\Phi(x)$. □

The following lemma, which is directly taken out of [31], shows that the set of points where Φ is not LC directional differentiable, is a Gaussian null set. Recall that a measure μ on a Banach space $\mathscr{B}$ is called *Gaussian* if for any nonzero $b \in \mathscr{B}^*$, the image measure $b_*(\mu) := \mu \circ b^{-1}$ is a Gaussian measure on $\mathbb{R}$. It is called *nondegenerate*, if for any $b \in \mathscr{B}^*$, the variance of $b_*(\mu)$ is nonzero.

Lemma 8 *Let $\mathscr{X}$ be a real separable Banach space, $\mathscr{Y}$ be a real Banach space such that every function $[0, 1] \to \mathscr{Y}$ of bounded variation is a.e. differentiable, $\emptyset \neq G \subset \mathscr{X}$ open. Moreover, let $T : G \to \mathscr{Y}$ be a locally Lipschitz mapping. Then T is LC directional differentiable outside a Gaussian null subset of G, i.e. for every nondegenerate Gaussian measure μ on G,*

$$\mu(\{x \in G : \ T \text{ is not LC directional differentiable in } x\}) = 0.$$

Proof See Theorem 1, Chapter 2 of [4] and Theorem 6 in [31]. □

This motivates the following assumption:

Hypothesis (G): The distribution of $X_T^0(\eta)$ is absolutely continuous with respect to some nondegenerate Gaussian measure, namely it holds $P_{X_T^0(\eta)} := X_T^0(\eta)(P) := P \circ (X_T^0(\eta))^{-1} \ll \mu$ for some nondegenerate Gaussian measure μ.

The following lemma provides a chain rule for $\Phi \circ X_T^0$

Lemma 9 *Let $\eta \in \mathbb{M}_2$ and $h \in M_2$. Under Hypotheses* (EU), (Flow), (H), (A') *and* (G) *it holds that the directional derivative $\partial_h(\Phi \circ X_T^0)(\eta)$ exists a.s. and we have*

$$\partial_h(\Phi \circ X_T^0)(\eta) = \partial_{DX_T^0(\eta)[h]}\Phi(X_T^0(\eta)).$$

Proof By definition of the directional derivative, we have

$$\begin{aligned}\partial_h(\Phi \circ X_T^0)(\eta) &= \lim_{\varepsilon\to 0} \frac{\Phi(X_T^0(\eta+\varepsilon h)) - \Phi(X_T^0(\eta))}{\varepsilon} \\ &= \lim_{\varepsilon\to 0}\Bigg(\frac{\Phi\Big(X_T^0(\eta) + \varepsilon\frac{X_T^0(\eta+\varepsilon h)-X_T^0(\eta)}{\varepsilon}\Big) - \Phi(X_T^0(\eta) + \varepsilon DX_T^0(\eta)[h])}{\varepsilon} \\ &\quad + \frac{\Phi(X_T^0(\eta) + \varepsilon DX_T^0(\eta)[h]) - \Phi(X_T^0(\eta))}{\varepsilon}\Bigg).\end{aligned}$$

Remark that, by Hypothesis (A'), Φ is Lipschitz, and, by Hypotheses (EU), (Flow) and (H), X_T^0 is Fréchet differentiable. Then we have for the first summand in this limit

$$\begin{aligned}
&\left|\frac{\Phi\left(X_T^0(\eta)+\varepsilon\frac{X_T^0(\eta+\varepsilon h)-X_T^0(\eta)}{\varepsilon}\right)-\Phi(X_T^0(\eta)+\varepsilon DX_T^0(\eta)[h])}{\varepsilon}\right| \\
&\quad\leqslant L_\Phi\left|\frac{X_T^0(\eta)+\varepsilon\frac{X_T^0(\eta+\varepsilon h)-X_T^0(\eta)}{\varepsilon}-X_T^0(\eta)-\varepsilon DX_T^0(\eta)[h]}{\varepsilon}\right| \\
&\quad= L_\Phi\left|\frac{X_T^0(\eta+\varepsilon h)-X_T^0(\eta)}{\varepsilon}-DX_T^0(\eta)[h]\right|\to 0, \qquad \text{as } \varepsilon\to 0.
\end{aligned}$$

As for the second summand in the above limit, by Hypothesis (G) and Lemma 8, we immediately have that

$$P(\{\omega\in\Omega:\ \Phi \text{ is not LC directional differentiable in } X_T^0(\eta,\omega)\})=0$$

and thus,

$$\partial_{DX_T^0(\eta)[h]}\Phi(X_T^0(\eta))=\lim_{\varepsilon\to 0}\frac{\Phi(X_T^0(\eta)+\varepsilon DX_T^0(\eta)[h])-\Phi(X_T^0(\eta))}{\varepsilon}$$

exists almost surely. This ends the proof. □

Proposition 3 *Under Hypotheses* (EU), (Flow), (H), (A') *and* (G) *it holds*

$$\partial_h p_n(\eta)\to\partial_h p(\eta). \tag{25}$$

Proof By Lemma 8 and Hypothesis (G), we have that

$$P(\{\omega\in\Omega:\ \Phi \text{ is not LC directional differentiable in } X_T^0(\eta,\omega)\})=0,$$

and thus, by Lemma 7,

$$D\Phi_n(X_T^0(\eta))\to\partial_\cdot\Phi(X_T^0(\eta)),\ \text{a.s.}$$

Therefore, applying the Fréchet differentiability of the mapping $\eta\mapsto X_T^0(\eta)$, the chain rule from Lemma 9 and the fact that the LC directional derivative is a continuous linear mapping (as a function of the direction), we obtain

$$\begin{aligned}
|D(\Phi_n\circ X_T^0)(\eta)[h]-\partial_h(\Phi\circ X_T^0)(\eta)| &= |D\Phi_n(X_T^0(\eta))DX_T^0(\eta)[h]-\partial_{DX_T^0(\eta)[h]}\Phi(X_T^0(\eta))| \\
&= |(D\Phi_n(X_T^0(\eta))-\partial_\cdot\Phi(X_T^0(\eta)))[DX_T^0(\eta)[h]]| \\
&\leqslant \|D\Phi_n(X_T^0(\eta))-\partial_\cdot\Phi(X_T^0(\eta))\|\cdot\|DX_T^0(\eta)\|\cdot\|h\| \\
&\longrightarrow 0,\ \text{a.s.},
\end{aligned}$$

as $n \to \infty$. Moreover, by Lemma 6 item (v) and Lemma 2, it holds

$$|D(\Phi_n \circ X_T^0)(\eta)[h]| \leqslant \|D\Phi_n(X_T^0(\eta))\| \cdot \|DX_T^0(\eta)[h]\| \leqslant L_\Phi \|DX_T^0(\eta)[h]\| \in L^1(\Omega).$$

Furthermore, similarly to the proof of Lemma 4, it can be shown that

$$\partial_h p(\eta) = E[\partial_h(\Phi \circ X_T^0)(\eta)] \text{ and} \tag{26}$$

$$\partial_h p_n(\eta) = E[D(\Phi_n \circ X_T^0)(\eta)[h]], \tag{27}$$

where, as for (26), we used that the LC directional derivative of $\Phi \circ X_T^0$ is defined for a.e. $\omega \in \Omega$ (rather than the Fréchet derivative). It now follows by dominated convergence that

$$\partial_h p_n(\eta) = E[D(\Phi_n \circ X_T^0)(\eta)[h]] \to E[\partial_h(\Phi \circ X_T^0)(\eta)] = \partial_h p(\eta).$$

By this we end the proof. □

Our final theorem summarizes the results of this section and shows that our representation formula (23) can be used in an approximation scheme for the directional derivatives of p in this more general setup:

Theorem 6 *Let Hypotheses* (EU), (Flow), (H), (A') *and* (G) *be fulfilled. Let* Φ_n *denote the nth Moreau-Yosida approximation of* Φ. *Then, for* $\xi = \exp(\mathbb{B}(1_{[-r,0]}))$,

$$\partial_h p(\eta) = -\lim_{n\to\infty} E\left[\left\{\delta\Big(\Phi_n(X_T^0(\tilde{\eta} + \lambda\xi h))a(\cdot)\Big)\right\}\Big|_{\lambda=\frac{1}{\xi}}\right]. \tag{28}$$

Proof As we have shown so far, $\partial_h p(\eta) = \lim_{n\to\infty} E[D(\Phi_n \circ X_T^0)(\eta)[h]]$ from Proposition 3. It follows from Lemma 6 items (iii) and (v) that Φ_n satisfies Hypothesis (A). Therefore, we can apply Theorem 5. □

Remark 2 Making use of the linearity of the derivative operator and the expectation, this result can easily be generalised to Φ being given by the difference of two convex, bounded from below and globally Lipschitz continuous functions $\Phi^{(1)}$ and $\Phi^{(2)}$.

To conclude this section, we provide an example, where the Hypothesis (G) holds.

Example 2 Let $d = m$, $T > r$, f be bounded and $g(s, \varphi) = Id_{d\times d}$, i.e.

$$\begin{cases} {}^\eta x(t) & = \eta(0) + \int_0^t f(s, {}^\eta x(s), {}^\eta x_s)ds + W(t),\ t \in [0, T] \\ {}^\eta x_0 & = \eta. \end{cases}$$

Then, the application of Girsanov's theorem (Novikov's condition is satisfied) yields that

$$ {}^{\eta}\tilde{W}(t) := \int_0^t f(s,\ {}^{\eta}x(s),\ {}^{\eta}x_s)ds + W(t) $$

is an m-dimensional Brownian motion under a measure ${}^{\eta}Q$ equivalent to P. Since $T > r$, we have

$$ X_T^0(\eta) = (\eta(0) + {}^{\eta}\tilde{W}(T), \eta(0) + {}^{\eta}\tilde{W}_T). $$

Now, since $P \ll {}^{\eta}Q$, it holds also

$$ P_{X_T^0(\eta)} \ll {}^{\eta}Q_{X_T^0(\eta)} = {}^{\eta}Q_{(\eta(0)+{}^{\eta}\tilde{W}(T),\eta(0)+{}^{\eta}\tilde{W}_T)}. $$

But ${}^{\eta}Q_{(\eta(0)+{}^{\eta}\tilde{W}(T),\eta(0)+{}^{\eta}\tilde{W}_T)}$ is a Gaussian measure on M_2, as for every $e \in M_2$ and every $A \in \mathscr{B}(\mathbb{R})$

$$ {}^{\eta}Q_{\langle(\eta(0)+{}^{\eta}\tilde{W}(T),\eta(0)+{}^{\eta}\tilde{W}_T),e\rangle}(A) = {}^{\eta}Q(\langle(\eta(0) + {}^{\eta}\tilde{W}(T), \eta(0) + {}^{\eta}\tilde{W}_T), e\rangle \in A) $$

$$ = {}^{\eta}Q\left(\eta(0)\left(e(0) + \int_{-r}^0 e(u)du\right) + {}^{\eta}\tilde{W}(T)e(0) + \int_{-r}^0 {}^{\eta}\tilde{W}(T+u)e(u)du \in A\right) $$

and ${}^{\eta}\tilde{W}$ is a Gaussian process under ${}^{\eta}Q$.

Appendix

Proof of Lemma 5:

(i):

$$ E[\|X_T^0(\tilde{\eta} + \lambda_1\xi h) - X_T^0(\tilde{\eta} + \lambda_2\xi h)\|_{M_2}^4] $$

$$ = E\left[\left(|\ {}^{\tilde{\eta}+\lambda_1\xi h}x(T) - {}^{\tilde{\eta}+\lambda_2\xi h}x(T)|_{\mathbb{R}^d}^2 + \int_{T-r}^T |\ {}^{\tilde{\eta}+\lambda_1\xi h}x(t) - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^2 dt\right)^2\right]. $$

Now splitting up the integral into an integral on $[T-r, T-r\vee 0]$ and an integral on $[T-r\vee 0, T]$ as we have done already in the proof of Lemma 2, we get

$$ \int_{T-r}^T |\ {}^{\tilde{\eta}+\lambda_1\xi h}x(t) - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^2 dt \leqslant r|\lambda_1-\lambda_2|^2|\xi|^2\|h\|_{M_2}^2 + \int_0^T |\ {}^{\tilde{\eta}+\lambda_1\xi h}x(t) $$

$$ - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^2 dt, $$

and therefore,

$$
\begin{aligned}
&E[\|X_T^0(\tilde{\eta}+\lambda_1\xi h) - X_T^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4] \\
&\leqslant \mathscr{O}(1)\Bigg(E\left[|\,{}^{\tilde{\eta}+\lambda_1\xi h}x(T) - {}^{\tilde{\eta}+\lambda_2\xi h}x(T)|_{\mathbb{R}^d}^4\right] + |\lambda_1-\lambda_2|^4 \\
&\quad + E\left[\int_0^T |\,{}^{\tilde{\eta}+\lambda_1\xi h}x(t) - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^4 dt\right]\Bigg).
\end{aligned}
$$

Now consider the term $E\left[|\,{}^{\tilde{\eta}+\lambda_1\xi h}x(t) - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^4\right]$. Similarly to the steps in the proof of Lemma 2 (applying Jensen's inequality, Burkholder-Davis-Gundy's inequality and the Lipschitzianity of f and g), we show that

$$
\begin{aligned}
&E\left[|\,{}^{\tilde{\eta}+\lambda_1\xi h}x(t) - {}^{\tilde{\eta}+\lambda_2\xi h}x(t)|_{\mathbb{R}^d}^4\right] \\
&\leqslant \mathscr{O}(1)\Bigg(|\lambda_1-\lambda_2|^4 + (L_f^4+L_g^4)\int_0^T E[\|X_u^0(\tilde{\eta}+\lambda_1\xi h) - X_u^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]du\Bigg).
\end{aligned}
$$

Finally, we can plug this into the inequality from before and get

$$
\begin{aligned}
&E[\|X_T^0(\tilde{\eta}+\lambda_1\xi h) - X_T^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4] \\
&\leqslant \mathscr{O}(1)\Bigg(|\lambda_1-\lambda_2|^4 + (L_f^4+L_g^4)\int_0^T E[\|X_u^0(\tilde{\eta}+\lambda_1\xi h) - X_u^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]du \\
&\quad + \int_0^T |\lambda_1-\lambda_2|^4 + (L_f^4+L_g^4)\int_0^T E[\|X_u^0(\tilde{\eta}+\lambda_1\xi h) - X_u^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]dudt\Bigg) \\
&\leqslant \mathscr{O}(1)\Bigg(|\lambda_1-\lambda_2|^4 + \int_0^T E[\|X_u^0(\tilde{\eta}+\lambda_1\xi h) - X_u^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]du\Bigg).
\end{aligned}
$$

Since we already know from Lemma 2 that $t \mapsto E[\|X_t^0(\tilde{\eta}+\lambda_1\xi h) - X_t^0(\tilde{\eta}+\lambda_2\xi h)\|_{M_2}^4]$ is integrable on $[0,T]$, the result follows directly by application of Gronwall's inequality and taking the square root.

(ii) and (iii): The proofs follow from the same considerations that we made in (i) and in the proof of Lemma 2, by applying Gronwall's inequality and making use of the fact that we have integrability of the functions $t \mapsto E[\|DX_T^0(\tilde{\eta}+\lambda_1\xi h)[\lambda_1\xi h]\|_{M_2}^4]^{\frac{1}{2}}$ and $t \mapsto E[\|DX_T^0(\tilde{\eta}+\lambda_1\xi h)[\lambda_1\xi h] - DX_T^0(\tilde{\eta}+\lambda_2\xi h)[\lambda_2\xi h]\|_{M_2}^2]$ by Lemma 2. □

Acknowledgements This research is conducted within the projects FINEWSTOCH (239019) and STOCHINF (250768) of the Research Council of Norway (NFR). The support of NFR is thankfully acknowledged.

References

1. Anh, V., Inoue, A.: Financial markets with memory I. Dynamic models. Stoch. Anal. Appl. **23**(2), 275–300 (2005)
2. Anh, V., Inoue, A., Kasahara, Y.: Financial markets with memory II. Dynamic models. Stoch. Anal. Appl. **23**(2), 301–328 (2005)
3. Appleby, J.A.D., Rodkina, A., Swords, C.: Fat tails and bubbles in a discrete time model of an inefficient financial market. Proc. Dyn. Sys. Appl. **5**, 35–45 (2008)
4. Aronszajn, N.: Differentiability of Lipschitz mappings between Banach spaces. Stud. Math. **T. LVIII**, 147–190 (1976)
5. Arriojas, M., Hu, Y., Mohammed, S.-E.A., Pap, G.: A delayed Black and Scholes formula. J. Stoch. Anal. Appl. **25**(2), 471–492 (2007)
6. Arriojas, M., Hu, Y., Mohammed, S.-E.A., Pap, G.: A delayed Black and Scholes formula II (2008). arXiv:math/0604641
7. Attouch, H.: Variational Convergence for Functions and Operators, 1st edn. Pitman, Boston/London/Melbourne (1984)
8. Bahar, A., Mao, X.: Stochastic delay Lotka-Volterra model. J. Math. Anal. Appl. **292**, 364–380 (2004)
9. Baños, D., Cordoni, F., Di Nunno, G., Di Persio, L., Røse, E.: Stochastic systems with memory and jumps. arXiv:1603.00272
10. Borwein, J.M., Noll, D.: Second order differentiability of convex functions in Banach spaces. Trans. Am. Math. Soc. **342**(1), 43–81 (1994)
11. Brett, T., Galla, T.: Stochastic processes with distributed delays: Chemical Langevin Equation and linear-noise approximation. Phys. Rev. Lett. **110**, 250601 (2013)
12. Chang, M.-H., Youree, R.K.: The European option with hereditary price structures: basic theory. Appl. Math. Comput. **102**, 279–296 (1999)
13. Chang, M.-H., Youree, R.K.: Infinite-dimensional Black-Scholes equation with hereditary structure. Appl. Math. Optim. **56**(3), 395–424 (2007)
14. Cont, R., Fournié, D.-A.: Change of variable formulas for non-anticipative functionals on path space. J. Funct. Anal. **259**(4), 1043–1072 (2010)
15. Cont, R., Fournié, D.-A.: Functional Itô calculus and stochastic integral representation of martingales. Ann. Probab. **41**(1), 109–133 (2013)
16. Cosso, A., Russo, F.: Functional and Banach space stochastic calculi: path-dependent Kolmogorov equations associated with the frame of a Brownian motion. In: Benth, F., Di Nunno, G. (eds.) Stochastics of Environmental and Financial Economics. Springer, Cham (2016)
17. Dupire, B.: Functional Itô calculus, Bloomberg Portfolio Research paper (2009)
18. Flandoli, F., Zanco, G.: An infinite-dimensional approach to path-dependent Kolmogorov equations. Ann. Probab. **44**(4), 2643–2693 (2016)
19. Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L., Touzi, N.: Applications of Malliavin calculus to Monte Carlo methods in finance. Finance Stochast. **3**, 391–412 (1999)
20. Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L.: Applications of Malliavin calculus to Monte Carlo methods in finance II. Finance Stochast. **5**, 201–236 (2001)
21. Hale, J.K., Verduyn Lunel, S.M.: Introduction to Functional Differential Equations. Springer, New York (1993)
22. Hobson, D., Rogers, L.C.G.: Complete models with stochastic volatility. Math. Financ. **8**, 27–48 (1998)
23. Kazmerchuk, Y., Swishchuk, A., Wu, J.: A continuous-time GARCH model for stochastic volatility with delay. Can. Appl. Math. Q. **13**, 123–149 (2005)
24. Kazmerchuk, Y., Swishchuk, A., Wu, J.: The pricing of options for securities markets with delayed response. Math. Comput. Simul. **75**, 69–79 (2007)
25. Küchler, U., Platen, E.: Time delay and noise explaining cyclical fluctuations in prices of commodities. Quantitative Finance Research Center, Research Paper 195 (2007)

26. Mao, X., Yuan, C., Zou, J.: Stochastic differential delay equations of population dynamics. J. Math. Anal. Appl. **304**, 296–320 (2005)
27. Miekisz, J., Poleszczuk, J., Bodnar, M., Foryś, U.: Stochastic models of gene expression with delayed degradation. Bull. Math. Biol. **73**(9), 2231–2247 (2011)
28. Mohammed, S.-E.A.: Stochastic Functional Differential Equations. Research Notes in Mathematics, vol. 99. Pitman Advanced Publishing Program, Boston (1984)
29. Mohammed, S.-E.A., Scheutzow, M.K.R.: The stable manifold theorem for non-linear stochastic systems with memory I. Existence of the semiflow. J. Funct. Anal. **205**, 271–305 (2003)
30. Nualart, D.: The Malliavin Calculus and Related Topics, 2nd edn. Springer, Berlin/Heidelberg (2006)
31. Phelps, R.R.: Gaussian null sets and differentiability of Lipschitz map on Banach spaces. Pac. J. Math. **77**(2), 523–531 (1978)
32. Platen, E., Heath, D.: A Benchmark Approach to Quantitative Finance. Springer, Berlin/Heidelberg (2006)
33. Pratt, J.W.: On interchanging limits and integrals. Ann. Math. Stat. **31**, 74–77 (1960)
34. Pronk, M., Veraar, M.: Tools for Malliavin calculus in UMD Banach spaces. Potential Anal. **40**(4), 307–344 (2014)
35. Schilling, R.L.: Measures, Integrals and Martingales. Cambridge University Press, Cambridge (2007)
36. Stoica, G.: A stochastic financial model. Proc. Am. Math. Soc. **133**(6), 1837–1841 (2005)
37. Swishchuk, A.V.: Modelling and Pricing of Swaps for Financial and Energy Markets with Stochastic Volatilities. World Scientific Publishing Co. Pte. Ltd., Singapore (2013)
38. Yan, F., Mohammed, S.: A stochastic calculus for systems with memory. Stoch. Anal. Appl. **23**(3), 613–657 (2005)

Grassmannian Flows and Applications to Nonlinear Partial Differential Equations

Margaret Beck, Anastasia Doikou, Simon J. A. Malham, and Ioannis Stylianidis

Abstract We show how solutions to a large class of partial differential equations with nonlocal Riccati-type nonlinearities can be generated from the corresponding linearized equations, from arbitrary initial data. It is well known that evolutionary matrix Riccati equations can be generated by projecting linear evolutionary flows on a Stiefel manifold onto a coordinate chart of the underlying Grassmann manifold. Our method relies on extending this idea to the infinite dimensional case. The key is an integral equation analogous to the Marchenko equation in integrable systems, that represents the coodinate chart map. We show explicitly how to generate such solutions to scalar partial differential equations of arbitrary order with nonlocal quadratic nonlinearities using our approach. We provide numerical simulations that demonstrate the generation of solutions to Fisher–Kolmogorov–Petrovskii–Piskunov equations with nonlocal nonlinearities. We also indicate how the method might extend to more general classes of nonlinear partial differential systems.

1 Introduction

It is well known that solutions to many integrable nonlinear partial differential equations can be generated from solutions to a linear integrable equation namely the Gel'fand–Levitan–Marchenko equation. It is an example of a generic dressing

M. Beck
Department of Mathematics and Statistics, Boston University, Boston, MA, USA
e-mail: mabeck@bu.edu

A. Doikou · S. J. A. Malham (✉) · I. Stylianidis
Maxwell Institute for Mathematical Sciences, and School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
e-mail: A.Doikou@hw.ac.uk; S.J.A.Malham@hw.ac.uk; is11@hw.ac.uk

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_3

transformation which we shall express in the form

$$g(x, y) = p(x, y) - \int_x^\infty g(x, z) q'(z, y; x)\, \mathrm{d}z, \tag{1}$$

for $y \geqslant x$. See Zakharov and Shabat [49] or Dodd, Eilbeck, Gibbon and Morris [17] for more details. Here all the functions shown may depend explicitly on time t, and we suppose that q' and p represent given data and g is the solution. Typically p represents the scattering data and takes the form $p = p(x + y)$ while q' depends on p, for example $q' = -p$ in the case of the Korteweg de Vries equation. See Ablowitz, Ramani and Segur [2] for more details. Typically given a nonlinear integrable partial differential equation, the function p is the solution to an associated linear system and the solution to the nonlinear integrable equation is given by $u = -2(\mathrm{d}/\mathrm{d}x)g(x, x)$. See for example Drazin and Johnson [19, p. 86] for the case of the Korteweg de Vries equation. The notion that the solution to a corresponding linear partial differential equation can be used to generate solutions to nonlinear integrable partial differential equations is addressed in the review by Miura [33]. An explicit formula was provided by Dyson [20] who showed that for the Korteweg de Vries equation the solution to the Gel'fand–Levitan–Marchenko equation along the diagonal $g = g(x, x)$ can be expressed in terms of the derivative of the logarithm of a tau-function or Fredholm determinant. In a series of papers Pöppe [36–38], Pöppe and Sattinger [39], Bauhardt and Pöppe [5], and Tracy and Widom [47] expressed the solutions to further nonlinear integrable partial differential equations in terms of Fredholm determinants. Importantly Pöppe [36] explicitly states the idea that:

> For every soliton equation, there exists a *linear* PDE (called a base equation) such that a map can be defined mapping a solution p of the base equation to a solution u of the soliton equation. The properties of the soliton equation may be deduced from the corresponding properties of the base equation which in turn are quite simple due to linearity. The map $p \to u$ essentially consists of constructing a set of linear integral operators using p and computing their Fredholm determinants.

From our perspective, the solution g to the dressing transformation represents an element of a Fredholm Grassmann manifold, expressed in a given coordinate patch. Let us briefly explain this perspective here. This will also help motivate the structures we introduce herein. Our original interest in Grassmann manifolds arose in spectral problems associated with nth order linear operators on the real line which can be expressed in the form

$$\partial_t \mathfrak{q} = A\mathfrak{q} + B\mathfrak{p} \tag{2a}$$

$$\partial_t \mathfrak{p} = C\mathfrak{q} + D\mathfrak{p}, \tag{2b}$$

where $\mathfrak{q} = \mathfrak{q}(t) \in \mathbb{C}^k$ and $\mathfrak{p} = \mathfrak{p}(t) \in \mathbb{C}^{n-k}$, with natural numbers $1 \leqslant k < n$. In these equations $A = A(t) \in \mathbb{C}^{k\times k}$, $B = B(t) \in \mathbb{C}^{k\times(n-k)}$, $C = C(t) \in \mathbb{C}^{(n-k)\times k}$ and $D = D(t) \in \mathbb{C}^{(n-k)\times(n-k)}$ are linear matrix operators. We assume as given that the matrix consisting of the blocks A, B, C and D has rank n for

all $t \in \mathbb{R}$. For example, in the case of elliptical eigenvalue problems, we have $(n-k) = k$ and $A = O$, $B = I_k$ and C contains the potential function. Then, with $t \in \mathbb{R}$ representing a spatial coordinate, the equations above are the corresponding first order representation of such an eigenvalue problem and the first equation corresponds to simply setting the variable $\mathfrak{p}$ to be the spatial derivative of $\mathfrak{q}$. The goal is to solve such eigenvalue problems by shooting. In such an approach the far-field boundary conditions, let's focus on the left far-field for the moment, naturally determine a subspace of solutions which decay to zero exponentially fast, though in general at different exponential rates. The choice of k above hitherto was arbitrary, now we retrospectively choose it to be the dimension of this subspace of solutions decaying exponentially in the left far-field. We emphasize it is the data, in this case the far-field data, that determines the dimension k of the subspace we consider. In principle we can integrate k solutions from the left far-field forward in t thus generating a continuous set of k-frames evolving with $t \in \mathbb{R}$. If $(\mathfrak{q}_1, \mathfrak{p}_1)^{\mathrm{T}}, \ldots, (\mathfrak{q}_k, \mathfrak{p}_k)^{\mathrm{T}}$ represent the solutions to the linear system (2) above that make up the components of the k-frame, we can represent them by

$$\begin{pmatrix} Q \\ P \end{pmatrix} \coloneqq \begin{pmatrix} \mathfrak{q}_1 & \cdots & \mathfrak{q}_k \\ \mathfrak{p}_1 & \cdots & \mathfrak{p}_k \end{pmatrix},$$

where $Q \in \mathbb{C}^{k\times k}$ and $P \in \mathbb{C}^{(n-k)\times k}$. From the linear system (2) for $\mathfrak{q}$ and $\mathfrak{p}$ above, the matrices $Q = Q(t)$ and $P = P(t)$ naturally satisfy the linear matrix system

$$\partial_t Q = AQ + BP \tag{3a}$$

$$\partial_t P = CQ + DP. \tag{3b}$$

To determine eigenvalues, by matching with the right far-field boundary conditions, the minimal information required is that for the subspace only and not the complete frame information. The Grassmann manifold $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$ is the set of k-dimensional subpsaces in $\mathbb{C}^n$. It is thus the natural context for the subspace evolution and then matching. See Alexander, Gardner and Jones [3] for a comprehensive account; they used the Plücker coordinate representation for the Grassmannian $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$. In Ledoux, Malham and Thümmler [31], Ledoux, Malham, Niesen and Thümmler [30] and Beck and Malham [7] we chose instead to directly project onto a coordinate patch of the Grassmannian $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$. Assuming that the matrix $Q \in \mathbb{C}^{k\times k}$ has rank k, we can achieve this as follows. We consider the transformation of coordinates here given by Q^{-1} that renders the first $k < n$ coordinates as orthonormal thus generating the matrix

$$\begin{pmatrix} I_k \\ G \end{pmatrix},$$

where $G(t) = P(t)\, Q^{-1}(t)$ for all $t \in \mathbb{R}$. Note this includes the data, i.e. $G(-\infty) = P(-\infty)Q^{-1}(-\infty)$. The key point which underlies the ideas in this paper is that $G = G(t) \in \mathbb{C}^{(n-k)\times k}$ evolves according to the evolutionary Riccati equation

$$\partial_t G = C + DG - G\,(A + BG), \tag{4}$$

where A, B, C and D are the block matrices from the linear evolutionary system (2) above. This equation for G is straightforwardly derived by direct computation. The Grassmannian $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$ is a homogeneous manifold. As such there are many numerical advantages to integrating along it, here this corresponds to computing $G = G(t)$. For instance, the aforementioned differing far-field exponential growth rates are projected out and G provides a useful succinct paramaterization of the subspace. It provides the natural extension of shooting methods to higher order linear spectral problems on the real line; also see Deng and Jones [16]. Let us call the procedure of deriving the Riccati equation (4) from the linear equations (3) the *forward* problem.

However in this finite dimensional context let us now turn this question around. Suppose our goal now is to solve the quadratically nonlinear evolution equation (4) above for some given data $G(-\infty) = G_0$. We assume of course the block matrices A, B, C and D have the properties described above. Let us call this the *inverse* problem. With the forward problem described above in mind, given such a nonlinear evolution equation to solve, we might naturally assume the nonlinear evolution equation (4) resulted from the projection of a linear Stiefel manifold flow onto a coordinate patch of the underlying Grassmann manifold. From this perspective, since all we are given is G or indeed just the data G_0, we can naturally assume G_0 was the result of such a Stiefel to Grassmann manifold projection. In particular we are free to assume that the transformation Q underlying the projection had rank k, and indeed $Q(-\infty)$, was simply I_k itself. Thus if we suppose we were given data $Q(-\infty) = I_k$ and $P(-\infty) = G_0$ and that $Q \in \mathbb{C}^{k\times k}$ and $P \in \mathbb{C}^{(n-k)\times k}$ satisfied the linear evolutionary equations (3) above then indeed $G = PQ^{-1}$ would solve the nonlinear evolution equation (4) above. Note that in this process there is nothing special about the data being prescribed at $t = -\infty$, it could be prescribed at any finite value of t, for example $t = 0$. In summary, suppose we want to solve the nonlinear Riccati equation (4) for some given data $G(0) = G_0 \in \mathbb{C}^{(n-k)\times k}$. Then if we suppose the matrices $Q \in \mathbb{C}^{k\times k}$ and $P \in \mathbb{C}^{(n-k)\times k}$ satisfy the linear system of equations (3) with $Q(0) = I_k$ and $P(0) = G_0$, then the solution $G \in \mathbb{C}^{(n-k)\times k}$ to the linear relation $P = G\,Q$ solves the Riccati equation (4) on some possibly small but non-zero interval of existence.

Our goal herein is to extend the idea just outlined to the infinite dimensional setting. Hereafter we always think of $t \in [0, \infty)$ as an evolutionary time variable. The natural extension of the finite rank (matrix) operator setting above to the infinite dimensional case is to pass over to the corresponding setting with compact operators. Thus formally, now suppose $Q = Q(t)$ and $P = P(t)$ are linear operators satisfying the linear system of evolution equations (3) for $t \geqslant 0$. We assume that A

and C are bounded operators, while the operators B and D may be bounded or unbounded. We suppose the solution operators $Q = Q(t)$ and $P = P(t)$ are such that for some $T > 0$, we have $Q(t) - \mathrm{id}$ and $P(t)$ are Hilbert–Schmidt operators for $t \in [0, T]$. Thus over this time interval $Q(t)$ is a Fredholm operator. If the operators B and D are bounded then we require that P lies in the subset of the class of Hilbert–Schmidt operators characterized by their domains. In addition we now suppose that $P = P(t)$ and $Q = Q(t)$ are related through a Hilbert–Schmidt operator $G = G(t)$ as follows

$$P = G\,Q. \tag{5}$$

We suppose herein this is a Fredholm equation for G and not of Volterra type like the dressing transformation above. We will return to this issue in our concluding section. As in the matrix case above, if we differentiate this Fredholm relation with respect to time using the product rule, insert the evolution equations (3) for $Q = Q(t)$ and $P = P(t)$, and then post-compose by Q^{-1}, then we obtain the Riccati evolution equation (4) for the Hilbert–Schmidt operator $G = G(t)$. We emphasize that, as for the matrix case above, for some time interval of existence $[0, T]$ with $T > 0$, we can generate the solution to the Riccati equation (4) with given initial data $G(0) = G_0$ by solving the two linear evolution equations (3) with the initial data $Q(0) = \mathrm{id}$ and $P(0) = G_0$ and then solving the third linear integral equation (5). This is the inverse problem in the infinite dimensional setting.

We now address how these operator equations are related to evolutionary partial differential equations. Can we use the approach above to find solutions to evolutionary partial differential equations with nonlocal quadratic nonlinearities in terms of solutions to the corresponding linearized evolutionary partial differential equations? Suppose that $\mathbb{V}$ is a closed linear subspace of $\mathbb{H} := L^2(\mathbb{R};\mathbb{R}) \times L^2_{\mathrm{d}}(\mathbb{R};\mathbb{R})$. Here $L^2_{\mathrm{d}}(\mathbb{R};\mathbb{R}) \subseteq L^2(\mathbb{R};\mathbb{R})$ represents the subspace of $L^2(\mathbb{R};\mathbb{R})$ corresponding to the intersection of the domains of the operators B and D. Suppose further that we have the direct sum decomposition $\mathbb{H} = \mathbb{V} \oplus \mathbb{V}^{\perp}$, where $\mathbb{V}^{\perp}$ represents the closed subspace of $\mathbb{H}$ orthogonal to $\mathbb{V}$. As already intimated, suppose for some $T > 0$ that for $t \in [0, T]$ we know: (i) $Q = Q(t)$ is a Fredholm operator from $\mathbb{V}$ to $\mathbb{V}$ of the form $Q = \mathrm{id} + Q'$ where $Q' = Q'(t)$ is a Hilbert–Schmidt operator on $\mathbb{V}$; and (ii) $P = P(t)$ is a Hilbert–Schmidt operator from $\mathbb{V}$ to $\mathbb{V}^{\perp}$. As such for $t \in [0, T]$ there exist integral kernels $q' = q'(x, y; t)$ and $p = p(x, y; t)$ with $x, y \in \mathbb{R}$ representing the action of the operators $Q'(t)$ and $P(t)$, respectively. Let us define the following nonlocal product for any two functions $g, g' \in L^2(\mathbb{R}^2;\mathbb{R})$ by

$$\big(g \star g'\big)(x, y) := \int_{\mathbb{R}} g(x, z)\, g'(z, y)\, \mathrm{d}z.$$

Suppose now that the unbounded operators B and D are now explicitly constant coefficient polynomial functions of ∂_x; let us denote them by $b = b(\partial_x)$ and $d = d(\partial_x)$. Further suppose A and C are bounded Hilbert–Schmidt operators which can be represented via their integral kernels, say $a = a(x, y; t)$ and $c = c(x, y; t)$,

respectively. If $g = g(x, y; t)$ represents the integral kernel corresponding to the Hilbert–Schmidt operator $G = G(t)$ then we observe that the two linear evolutionary equations (3) and linear integral equation (5) can be expressed as follows. We have

1. *Base equation:* $\partial_t p = c \star (\delta + q') + d\, p$;
2. *Aux. equation:* $\partial_t q = a \star (\delta + q') + b\, p$;
3. *Riccati relation:* $p = g \star (\delta + q')$.

Here δ is the Dirac delta function representing the identity at the level of integral kernels. The evolutionary equation for $g = g(x, y; t)$ corresponding to the Riccati evolution equation (4) takes the form

$$\partial_t g = c + d\, g - g \star (a + b\, g).$$

This is an evolutionary partial differential equation for $g = g(x, y; t)$ with a nonlocal quadratic nonlinearity '$g \star (b\, g)$'. Explicitly it has the form

$$\partial_t g(x, y; t) = c(x, y; t) + d(\partial_x)\, g(x, y; t) - \int_{\mathbb{R}} g(x, z; t)\big(a(z, y; t) + b(\partial_z)\, g(z, y; t)\big)\, \mathrm{d}z.$$

The reason underlying the nomination of the base and auxiliary equations above is that in most of our examples we have $a \equiv c \equiv 0$—let us assume this for the sake of our present argument. We have outlined the forward problem identified earlier at the partial differential equation level. However our goal is to solve the inverse problem at this level: given a nonlinear evolutionary partial differential equation of the form above with some arbitrary initial data $g(x, y; 0) = g_0(x, y)$, can we re-engineer solutions to it from solutions to the corresponding base and auxiliary equations? The answer is yes. Given the nonlinear evolutionary partial differential equation $\partial_t g = d\, g - g \star (b\, g)$ with $b = b(\partial_x)$ and $d = d(\partial_x)$ as described above, suppose we solve the corresponding linear base equation for $p = p(x, y; t)$, which is the linearized version of the given equation and consequently solve the auxiliary equation for $q' = q'(x, y; t)$. Then solutions $g = g(x, y; t)$ to the nonlinear evolutionary partial differential equation are re-engineered/generated by solving the Riccati relation for p and q' which is a linear Fredholm integral equation.

We explicitly demonstrate this procedure through two examples. We consider two Fisher–Kolmogorov–Petrovskii–Piskunov type equations with nonlocal nonlinearities. The first has a nonlocal nonlinear term of the form '$g \star g$' where the product '$\star$' represents the special case of convolution. The second has a nonlocal nonlinear term of the form '$g \star (b\, g)$' where $b = b(x)$ is a multiplicative function corresponding to a correlation in the nonlinearity. In this latter case the product '$\star$' has the general form as originally defined above. In both these cases we show how solutions can be generated using the approach we propose from arbitrary initial data. We provide numerical simulations to confirm this. From these examples we also see how our procedure extends straightforwardly to any higher order diffusion. We additionally show how Burgers' equation and its solution using the corresponding

base equation via the Cole–Hopf transformation fits into the context we have described here.

We emphasize that, as is well known for the Gel'fand–Levitan–Marchenko equation above which is of Volterra type, the procedure we have outlined works for most integrable systems, as demonstrated in Ablowitz, Ramani and Segur [2] who assume $p = p(x + y)$ is a Hankel kernel. For example, we can generate solutions to the Korteweg de Vries equation from the Gel'fand–Levitan–Marchenko equation by setting $q' = -p$. As another example, we can generate solutions to the nonlinear Schrödinger equation by assuming $q'(z, y; x) = \pm \int_x^\infty \overline{p}(z, \zeta)\, p(\zeta, y)\, \mathrm{d}\zeta$ where $\overline{p}$ represents the complex conjugate of p. In this case it is also well known that such solutions can be generated from a 2×2 matrix-valued dressing transformation. See Zakharov and Shabat [49], Dodd, Eilbeck, Gibbon and Morris [17] or Drazin and Johnson [19] for more details. Further, the connection between integrable systems and infinite dimensional Grassmann manifolds was first made by Sato [42, 43]. Lastly Riccati systems are a central feature of optimal control systems. The solution to a matrix Riccati equation provides the optimal continuous feedback in optimal linear-quadratic control theory. See for example Bittanti, Laub and Willems [10], Brockett and Byrnes [14], Hermann [25, 26], Hermann and Martin [27], Martin and Hermann [32] and Zelikin [50] for more details. A comprehensive list of the related control literature can also be found in Ledoux, Malham and Thümmler [31].

Our paper is structured as follows. In Sect. 2 we define and outline the Grassmann manifolds in finite and infinite dimensions that we require to give the appropriate context to our procedure. Then in Sect. 3 we show how linear subspace flows induce Riccati flows in coordinate patches of the corresponding Fredholm Grassmannian. We derive the equation for the evolution of the integral kernel associated with the Riccati flow. We then consider two pertinent examples in Sect. 4. Their solutions can be derived by solving the linear base and auxiliary partial differential equations (the subspace flow) and then solving the linear Fredholm equation representing the projection of the subspace flow onto a coordinate patch of the Fredholm Grassmannian. Then finally in Sect. 5 we discuss possible extensions of our approach to other nonlinear partial differential equations.

2 Grassmann Manifolds

Grassmann manifolds underlie the structure, development and solution of the differential equations we consider herein. Hence we introduce them here first in the finite dimensional, and then second in the infinite dimensional, setting. There are many perspectives and prescriptions, we choose the prescriptive path that takes us most efficiently to the infinite dimensional setting we require herein.

Suppose we have a finite dimensional vector space say $\mathbb{H} = \mathbb{C}^n$ of dimension $n \in \mathbb{N}$. Given an integer k with $1 \leqslant k < n$, the Grassmann manifold $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$ is defined to be the set of k-dimensional linear subspaces of $\mathbb{C}^n$. Let $\{\mathrm{e}_j\}_{j\in\{1,\ldots,n\}}$ denote the *canonical basis* for $\mathbb{C}^n$, where e_j is the $\mathbb{C}^n$-valued vector with one in the

jth entry and zeros in all the other entries. Suppose we are given a set of k linearly independent vectors in $\mathbb{C}^n$ and we record them in the following $n \times k$ matrix:

$$W = \begin{pmatrix} w_{1,1} & \cdots & w_{1,k} \\ \vdots & & \vdots \\ w_{n,1} & \cdots & w_{n,k} \end{pmatrix}.$$

Each column is one of the linear independent vectors in $\mathbb{C}^n$. This matrix has rank k. Naturally the columns of W span a k-dimensional subspace or k-plane $\mathbb{W}$ in $\mathbb{C}^n$. Let us denote by $\mathbb{V}_0$ the *canonical subspace* given by $\operatorname{span}\{e_1, \ldots, e_k\}$, i.e. the subspace prescribed by the first k canonical basis vectors which has the representation

$$W_0 \coloneqq \begin{pmatrix} I_k \\ O \end{pmatrix}.$$

Here I_k is the $k \times k$ identity matrix. The span of the vectors $\{e_{k+1}, \ldots, e_n\}$ represents the subspace $\mathbb{V}_0^\perp$, the $(n-k)$-dimensional subspace of $\mathbb{C}^n$ orthogonal to $\mathbb{V}_0$. Suppose we are able to project $\mathbb{W}$ onto $\mathbb{V}_0$. Then the projections $\operatorname{pr}\colon \mathbb{W} \to \mathbb{V}_0$ and $\operatorname{pr}\colon \mathbb{W} \to \mathbb{V}_0^\perp$ respectively give

$$W^{\parallel} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,k} \\ \vdots & & \vdots \\ w_{k,1} & \cdots & w_{k,k} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \quad \text{and} \quad W^{\perp} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ w_{k+1,1} & \cdots & w_{k+1,k} \\ \vdots & & \vdots \\ w_{n,1} & \cdots & w_{n,k} \end{pmatrix}.$$

The existence of this projection presupposes that the rank of the matrix $W^{\parallel}$ on the left above is k, i.e. the determinant of the upper $k \times k$ block say W_{up} is non-zero. This is not always true, we account for this momentarily. The subspace given by the span of the columns of $W^{\parallel}$ naturally coincides with $\mathbb{V}_0$. Indeed since $W^{\parallel}$ has rank k, there exists a rank k transformation from $\mathbb{V}_0 \to \mathbb{V}_0$, given by $W_{\text{up}}^{-1} \in \mathrm{GL}(\mathbb{C}^k)$, that transforms $W^{\parallel}$ to W_0. Thus what distinguishes $\mathbb{W}$ from $\mathbb{V}_0$ is the form of $W^{\perp}$. Under the same transformation of coordinates W_{up}^{-1}, the lower $(n-k) \times k$ matrix say W_{low} of $W^{\perp}$ becomes the $(n-k) \times k$ matrix $G \coloneqq W_{\text{low}} W_{\text{up}}^{-1}$. Or in other words if we perform this transformation of coordinates, the matrix W as a whole becomes

$$\begin{pmatrix} I_k \\ G \end{pmatrix}. \tag{6}$$

Thus any k-dimensional subspace $\mathbb{W}$ of $\mathbb{C}^n$ which can be projected onto $\mathbb{V}_0$ can be represented by this matrix. Conversely any $n \times k$ matrix of this form represents a k-dimensional subspace $\mathbb{W}$ of $\mathbb{C}^n$ that can be projected onto $\mathbb{V}_0$. The matrix G thus paramaterizes all the k-dimensional subspaces $\mathbb{W}$ that can be projected onto $\mathbb{V}_0$. As G varies, the orientation of the subspace $\mathbb{W}$ within $\mathbb{C}^n$ varies.

What about the k-dimensional subspaces in $\mathbb{C}^n$ that cannot be projected onto $\mathbb{V}_0$? This occurs when one or more of the column vectors of W are parallel to one or more of the orthogonal basis vectors $\{\mathrm{e}_{k+1}, \ldots, \mathrm{e}_n\}$. Such subspaces cannot be represented in the form (6) above. Any such matrices W are rank k matrices by choice, their columns span a k-dimensional subspace $\mathbb{W}$ in $\mathbb{C}^n$, it's just that they have a special orientation in the sense just described. We simply need to choose a better representation. Given a multi-index $\mathbb{S} = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ of cardinality k, let $\mathbb{V}_0(\mathbb{S})$ denote the subspace given by $\operatorname{span}\{\mathrm{e}_{i_1}, \ldots, \mathrm{e}_{i_k}\}$. The vectors $\{\mathrm{e}_i\}_{i \in \mathbb{S}^o}$ span the subspace $\mathbb{V}_0^\perp(\mathbb{S})$, the $(n-k)$-dimensional subspace of $\mathbb{C}^n$ orthogonal to $\mathbb{V}_0(\mathbb{S})$. Since W has rank k, there exists a multi-index $\mathbb{S}$ such that the projection $\mathrm{pr}\colon \mathbb{W} \to \mathbb{V}_0(\mathbb{S})$ exists. The arguments above apply with $\mathbb{V}_0(\mathbb{S})$ replacing $\mathbb{V}_0 = \mathbb{V}_0(\{1, \ldots, k\})$. The projections $\mathrm{pr}\colon \mathbb{W} \to \mathbb{V}_0(\mathbb{S})$ and $\mathrm{pr}\colon \mathbb{W} \to \mathbb{V}_0^\perp(\mathbb{S})$ respectively give

$$W_{\mathbb{S}}^{\|} = \begin{pmatrix} W_{\mathbb{S}} \\ O_{\mathbb{S}^o} \end{pmatrix} \qquad \text{and} \qquad W_{\mathbb{S}}^{\perp} = \begin{pmatrix} O_{\mathbb{S}} \\ W_{\mathbb{S}^o} \end{pmatrix}.$$

Here $W_{\mathbb{S}}$ represents the $k \times k$ matrix consisting of the $\mathbb{S}$ rows of W and so forth, and, for example, the form for $W_{\mathbb{S}}^{\|}$ shown is meant to represent the $n \times k$ matrix whose $\mathbb{S}$ rows are occupied by $W_{\mathbb{S}}$ while the remaining rows contain zeros. We can perform a rank k transformation of coordinates $\mathbb{V}_0(\mathbb{S}) \to \mathbb{V}_0(\mathbb{S})$ via $W_{\mathbb{S}}^{-1} \in \mathrm{GL}(\mathbb{C}^k)$ under which the matrix W becomes

$$\begin{pmatrix} I_{\mathbb{S}} \\ G_{\mathbb{S}^o} \end{pmatrix}. \tag{7}$$

Thus $G_{\mathbb{S}^o}$ parameterizes all k-dimensional subspaces $\mathbb{W}$ that can be projected onto $\mathbb{V}_0(\mathbb{S})$. Each possible choice of $\mathbb{S}$ generates a coordinate patch of the Grassmann manifold $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$. For more details on establishing $\mathrm{Gr}(\mathbb{C}^n, \mathbb{C}^k)$ as a compact and connected manifold, see Griffiths and Harris [23, p. 193–4].

Let us now consider the infinite dimensional extension to Fredholm Grassmann manifolds. They are also known as Sato Grassmannians, Segal–Wilson Grassmannians, Hilbert–Schmidt Grassmannians and restricted Grassmannians, as well as simply Hilbert Grassmannians. See Sato [42, 43], Miwa, Jimbo and Date [34], Segal and Wilson [45, Section 2] and Pressley and Segal [40, Chapters 6,7] for more details. In the infinite dimensional setting we suppose the underlying vector space is a separable Hilbert space $\mathbb{H} = \mathbb{H}(\mathbb{C})$. Any separable Hilbert space is isomorphic to the sequence space $\ell^2 = \ell^2(\mathbb{C})$ of square summable complex sequences; see Reed and Simon [41, p. 47]. We will parameterize the $\mathbb{C}$-valued components of

the sequences in $\ell^2 = \ell^2(\mathbb{C})$ by $\mathbb{N}$. This is sufficient as any sequence space $\ell^2 = \ell^2(\mathbb{F}; \mathbb{C})$, where $\mathbb{F}$ denotes a countable field isomorphic to $\mathbb{N}$ that parameterizes the sequences therein, is isomorphic to $\ell^2 = \ell^2(\mathbb{N}; \mathbb{C})$. We recall any $\mathfrak{a} \in \ell^2(\mathbb{C})$ has the form $\mathfrak{a} = \{\mathfrak{a}(1), \mathfrak{a}(2), \mathfrak{a}(3), \ldots\}$ where $\mathfrak{a}(n) \in \mathbb{C}$ for each $n \in \mathbb{N}$. Hereafter we represent such sequences by column vectors $\mathfrak{a} = (\mathfrak{a}(1), \mathfrak{a}(2), \mathfrak{a}(3), \ldots)^{\mathrm{T}}$. Since we require square summability, we must have $\mathfrak{a}^\dagger \mathfrak{a} = \sum_{n\in\mathbb{N}} \mathfrak{a}^*(n)\, \mathfrak{a}(n) < \infty$, where $\dagger$ denotes complex conjugate transpose and $*$ denotes complex conjugate only. We define the inner product $\langle \cdot, \cdot \rangle \colon \ell^2(\mathbb{C}) \otimes \ell^2(\mathbb{C}) \to \mathbb{R}$ by $\langle \mathfrak{a}, \mathfrak{b} \rangle := \sqrt{\mathfrak{a}^\dagger \mathfrak{b}}$ for any $\mathfrak{a}, \mathfrak{b} \in \ell^2(\mathbb{C})$. A natural complete orthonormal basis for $\ell^2(\mathbb{C})$ is the *canonical basis* $\{\mathrm{e}_n\}_{n\in\mathbb{N}}$ where e_n is the sequence whose nth component is one and all other components are zero. We have the following corresponding definition for the Grassmannian of all subspaces comparable in size to a given closed subspace $\mathbb{V} \subset \mathbb{H}$; see Segal and Wilson [45] and Pressley and Segal [40].

Definition 1 (Fredholm Grassmannian) Let $\mathbb{H}$ be a separable Hilbert space with a given decomposition $\mathbb{H} = \mathbb{V} \oplus \mathbb{V}^\perp$, where $\mathbb{V}$ and $\mathbb{V}^\perp$ are infinite dimensional closed subspaces. The Grassmannian $\mathrm{Gr}(\mathbb{H}, \mathbb{V})$ is the set of all subspaces W of $\mathbb{H}$ such that:

1. The orthogonal projection $\mathrm{pr}\colon W \to \mathbb{V}$ is a Fredholm operator, indeed it is a Hilbert–Schmidt perturbation of the identity; and
2. The orthogonal projection $\mathrm{pr}\colon W \to \mathbb{V}^\perp$ is a Hilbert–Schmidt operator.

Herein we exclusively assume that our underlying separable Hilbert space $\mathbb{H}$ and closed subspace $\mathbb{V}$ are of the form

$$\mathbb{H} := \ell^2(\mathbb{C}) \times \ell^2_{\mathrm{d}}(\mathbb{C}) \qquad \text{and} \qquad \mathbb{V} := \ell^2(\mathbb{C}),$$

where $\ell^2_{\mathrm{d}}(\mathbb{C})$ is a closed subspace of $\ell^2(\mathbb{C})$. We thus assume a special form for $\mathbb{H}$. This form is the setting for our applications discussed in our Introduction. We use it to motivate the definition of the Fredholm Grassmannian above and its relation to our applications. Suppose we are given a set of independent sequences in $\ell^2(\mathbb{C}) \times \ell^2_{\mathrm{d}}(\mathbb{C})$ which span $\ell^2(\mathbb{C})$ and we record them as columns in the infinite matrix

$$W = \begin{pmatrix} Q \\ P \end{pmatrix}.$$

Here each column of Q lies in $\ell^2(\mathbb{C})$ and each column of P lies in $\ell^2_{\mathrm{d}}(\mathbb{C})$. We denote by $\mathbb{W}$ the subspace of $\ell^2(\mathbb{C}) \times \ell^2_{\mathrm{d}}(\mathbb{C})$ spanned by the columns of W. Let us denote by $\mathbb{V}_0$ the *canonical subspace* which has the corresponding representation

$$W_0 = \begin{pmatrix} \mathrm{id} \\ O \end{pmatrix},$$

where $\mathrm{id} = \mathrm{id}_{\ell^2(\mathbb{C})}$. As above, suppose we are able to project $\mathbb{W}$ on $\mathbb{V}_0$. The projections $\mathrm{pr}\colon \mathbb{W} \to \mathbb{V}_0$ and $\mathrm{pr}\colon \mathbb{W} \to \mathbb{V}_0^{\perp}$ respectively give

$$W^{\parallel} = \begin{pmatrix} Q \\ O \end{pmatrix} \qquad \text{and} \qquad W^{\perp} = \begin{pmatrix} O \\ P \end{pmatrix}.$$

The existence of this projection presupposes that the determinant of the upper block Q is non-zero. We must now choose in which sense we want this to hold. The columns of Q are $\ell^2(\mathbb{C})$-valued. We now retrospectively assume that we constructed Q so that, not only do its columns span $\ell^2(\mathbb{C})$, it is also a Fredholm operator on $\ell^2(\mathbb{C})$ of the form $Q = \mathrm{id} + Q'$ where $Q' \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2(\mathbb{C})\big)$ and $\mathrm{id} = \mathrm{id}_{\ell^2(\mathbb{C})}$. Here $\mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2(\mathbb{C})\big)$ is the class of Hilbert–Schmidt operators from $\ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, equipped with the norm

$$\|Q'\|^2_{\mathfrak{J}_2(\ell^2(\mathbb{C});\ell^2(\mathbb{C}))} \coloneqq \mathrm{tr}\,(Q')^{\dagger}(Q'),$$

where 'tr' represents the trace operator. For such Hilbert–Schmidt operators Q' we can define the regularized Fredholm determinant

$$\det{}_2\big(\mathrm{id} + Q'\big) \coloneqq \exp\left(\sum_{\ell \geqslant 2} \frac{(-1)^{\ell-1}}{\ell} \mathrm{tr}\,(Q')^{\ell}\right).$$

The operator $Q = \mathrm{id} + Q'$ is invertible if and only if $\det_2\big(\mathrm{id} + Q'\big) \neq 0$. For more details see Simon [46]. Hence, assuming that $Q' \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2(\mathbb{C})\big)$, we can assert that the subspace given by the span of the columns of $W^{\parallel}$ coincides with the subspace spanned by W_0, i.e. with $\mathbb{V}_0$. Indeed the transformation given by $Q^{-1} \in \mathrm{GL}\big(\ell^2(\mathbb{C})\big)$ transforms $W^{\parallel}$ to W_0. Let us now focus on $W^{\perp}$. We now also retrospectively assume that we constructed P so that, not only do its columns span $\ell^2_{\mathrm{d}}(\mathbb{C})$, it is a Hilbert–Schmidt operator from $\ell^2(\mathbb{C})$ to $\ell^2_{\mathrm{d}}(\mathbb{C})$, i.e. $P \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2_{\mathrm{d}}(\mathbb{C})\big)$. Hence under the transformation of coordinates $Q^{-1} \in \mathrm{GL}\big(\ell^2(\mathbb{C})\big)$ the matrix for W becomes

$$\begin{pmatrix} \mathrm{id} \\ G \end{pmatrix},$$

where $G \coloneqq PQ^{-1}$. Thus any subspace $\mathbb{W}$ that can be projected onto $\mathbb{V}_0$ can be represented in this way, and conversely. The operator $G \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2_{\mathrm{d}}(\mathbb{C})\big)$ thus parameterizes all subspaces $\mathbb{W}$ that can be projected onto $\mathbb{V}_0$. We call the Fredholm index of the Fredholm operator Q the *virtual dimension* of W; see Segal and Wilson [45] and Pressley and Segal [40] for more details.

Remark 1 (Canonical coordinate patch) In our applications we consider evolutionary flows in which the operators $Q = Q(t)$ and $P = P(t)$ above evolve, as functions of time $t \geqslant 0$, as solutions to linear differential equations. The initial data in all cases is taken to be $Q(0) = \mathrm{id}$ and $P(0) = G_0$ for some given data $G_0 \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2_{\mathrm{d}}(\mathbb{C})\big)$. By assumption in general and by demonstration in practice, the flows are well-posed and smooth in time for $t \in [0, T]$ for some $T > 0$. Hence there exists a time $T > 0$ such that for $t \in [0, T]$ we know, by continuity, that $Q = Q(t)$ is an invertible Hilbert-Schmidt operator of virtual dimension zero and of the form $Q(t) = \mathrm{id} + Q'(t)$ where $Q'(t) \in \mathfrak{J}_2\big(\ell^2(\mathbb{C}); \ell^2(\mathbb{C})\big)$. For this time the flow for $Q = Q(t)$ and $P = P(t)$ prescibes a flow for $G = G(t)$, with $G(t) = P(t)Q^{-1}(t)$. In addition, for this time, whilst the orientation of the subspace prescribed by $Q = Q(t)$ and $P = P(t)$ evolves, the flow remains within the same coordinate patch of the Grassmannian $\mathrm{Gr}(\mathbb{H}, \mathbb{V})$ prescribed by the initial data as just described and explicitly outlined above.

Remark 2 (Frames) More details on "frames" in the infinite dimensional context can be found in Christensen [15] and Balazs [4].

There are three possible obstructions to the construction of the class of subspaces above as follows, the: (i) Virtual dimension of $\mathbb{W}$, i.e. the Fredholm index of Q, may differ by an integer value; (ii) Operator Q' may not be Hilbert–Schmidt valued—it could belong to a 'higher' Schatten–von Neumann class; or (iii) Determinant of Q may be zero. The consequences of these issues for connected components, submanifolds and coordinate patches of $\mathrm{Gr}(\mathbb{H}, \mathbb{V})$ are covered in detail in general in Pressley and Segal [40, Chap. 7]. These have important implications for regularity of the flows mentioned in Remark 1 above, i.e. for our applications. However we leave these questions for further investigation, see Sect. 5. Suffice to say for the moment, from Pressley and Segal [40, Prop. 7.1.6], we know that given any subspace $\mathbb{W}$ of $\mathbb{H}$ there exists a representation analogous to the general coordinate patch form (7) with $\mathbb{S}$ a suitable countable set. In other words there exists a subspace cover. More details on infinite dimensional Grassmannians can be found in Sato [42, 43], Abbondandolo and Majer [1] and Furitani [21].

We have introduced the Fredholm Grassmannian here in the context where the underlying Hilbert space is $\mathbb{H} = \ell^2(\mathbb{C}) \times \ell^2_{\mathrm{d}}(\mathbb{C})$ and the subspace $\mathbb{V} \cong \ell^2(\mathbb{C})$. In our applications the context will be $\mathbb{H} = L^2(\mathbb{I}; \mathbb{C}) \times L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{C})$ and $\mathbb{V} \cong L^2(\mathbb{I}; \mathbb{C})$ where the continuous interval $\mathbb{I} \subseteq \mathbb{R}$. We include here the cases when $\mathbb{I}$ is finite, semi-infinite of the form $[a, \infty)$ for some real constant a or the whole real line. As above, here $L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{C})$ denotes a closed subspace of $L^2(\mathbb{I}; \mathbb{C})$—corresponding to intersection of the domains of the unbounded operators D and B in our applications. All such spaces $L^2(\mathbb{I}; \mathbb{C})$ are separable and isomorphic to $\ell^2(\mathbb{C})$, and correspondingly for the closed subspaces. See for example Christensen [15] or Blanchard and Brüning [11] for more details. It is straightforward to transfer statements we have made thusfar for the Fredholm Grassmannian in the square-summable sequence space context across to the square integrable function space context. When $\mathbb{H} = L^2(\mathbb{I}; \mathbb{C}) \times L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{C})$ and $\mathbb{V} \cong L^2(\mathbb{I}; \mathbb{C})$ the operators Q' and P are Hilbert–Schmidt operators in the sense

that $Q' \in \mathfrak{J}_2\big(L^2(\mathbb{I};\mathbb{C});L^2(\mathbb{I};\mathbb{C})\big)$ and $P \in \mathfrak{J}_2\big(L^2(\mathbb{I};\mathbb{C});L^2_{\mathrm{d}}(\mathbb{I};\mathbb{C})\big)$. By standard theory such Hilbert–Schmidt operators can be parameterized via integral kernel functions, say, $q' \in L^2(\mathbb{I}^2;\mathbb{C})$ and $p \in L^2(\mathbb{I};L^2_{\mathrm{d}}(\mathbb{I};\mathbb{C}))$ and their actions represented by

$$Q'(f)(x) = \int_{\mathbb{I}} q'(x,y)\, f(y)\, \mathrm{d}y,$$
$$P(f)(x) = \int_{\mathbb{I}} p(x,y)\, f(y)\, \mathrm{d}y,$$

for any $f \in L^2(\mathbb{I};\mathbb{C})$ and where $x \in \mathbb{I}$. Furthermore we know we have the isometries $\|Q'\|_{\mathfrak{J}_2(L^2(\mathbb{I};\mathbb{C});L^2(\mathbb{I};\mathbb{C}))} = \|q'\|_{L^2(\mathbb{I}^2;\mathbb{C})}$ and $\|P\|_{\mathfrak{J}_2(L^2(\mathbb{I};\mathbb{C});L^2_{\mathrm{d}}(\mathbb{I};\mathbb{C}))} = \|p\|_{L^2(\mathbb{I};L^2_{\mathrm{d}}(\mathbb{I};\mathbb{C}))}$; see Reed and Simon [41, p. 210]. Hence the subspace $\mathbb{W}$ above and its representation in the canonical coordinate patch are given by

$$W = \begin{pmatrix} q \\ p \end{pmatrix} \quad \rightsquigarrow \quad \begin{pmatrix} \delta \\ g \end{pmatrix}.$$

Here we suppose $q(x,y) = \delta(x-y) + q'(x,y)$ with $\delta(x-y)$ representing the identity operator in $L^2(\mathbb{I};\mathbb{C})$ at the integral kernel level. The function $g = g(x,y)$ is the $L^2(\mathbb{I};L^2_{\mathrm{d}}(\mathbb{I};\mathbb{C}))$-valued kernel associated with the Hilbert–Schmidt operator G. It is explicitly obtained by solving the Fredholm equation given by

$$p(x,y) = \int_{\mathbb{I}} g(x,z)\, q(z,y)\, \mathrm{d}z.$$

Solving this equation for g is equivalent to solving the operator relation $P = G\,Q$ for G by postcomposing by Q^{-1}.

3 Fredholm Grassmannian Flows

We show how linear evolutionary flows on subspaces of an abstract separable Hilbert space $\mathbb{H}$ generate a quadratically nonlinear flow on a coordinate patch of an associated Fredholm Grassmann manifold. The setting is similar to that outlined at the beginning of the last section. Assume for the moment that $\mathbb{H}$ admits a direct sum orthogonal decomposition $\mathbb{H} = \mathbb{V} \oplus \mathbb{V}^{\perp}$, where $\mathbb{V}$ and $\mathbb{V}^{\perp}$ are closed subspaces of $\mathbb{H}$. The subspace $\mathbb{V}$ is fixed. Now suppose there exists a time $T > 0$ such that for each time $t \in [0,T]$ there exists a continuous path of subspaces $\mathbb{W} = \mathbb{W}(t)$ of $\mathbb{H}$ such that the projections $\mathrm{pr}\colon \mathbb{W}(t) \to \mathbb{V}$ and $\mathrm{pr}\colon \mathbb{W}(t) \to \mathbb{V}^{\perp}$ can be respectively parameterised by the operators $Q(t) = \mathrm{id} + Q'(t)$ and $P(t)$. We in fact assume the path of subspaces $\mathbb{W}(t)$ is smooth in time and $Q'(t)$ and

$P(t)$ are Hilbert–Schmidt operators so that indeed $Q' \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V})\big)$ and $P \in C^\infty\big([0,T];\mathrm{Dom}(D)\cap\mathrm{Dom}(B)\big)$. Here D and B are in general unbounded operators for which, as we see presently, for each $t \in [0,T]$ we require $DP(t) \in \mathfrak{J}_2(\mathbb{V};\mathbb{V}^\perp)$ and $BP(t) \in \mathfrak{J}_2(\mathbb{V};\mathbb{V})$. The subspaces $\mathrm{Dom}(D) \subseteq \mathfrak{J}_2(\mathbb{V};\mathbb{V}^\perp)$ and $\mathrm{Dom}(B) \subseteq \mathfrak{J}_2(\mathbb{V};\mathbb{V})$ are their respective domains. Our analysis also involves two bounded operators $A \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V})\big)$ and $C \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V}^\perp)\big)$. The evolution of $Q = Q(t)$ and $P = P(t)$ is prescribed by the following system of differential equations.

Definition 2 (Linear Base and Auxiliary Equations) We assume there exists a $T > 0$ such that, for the linear operators A, B, C and D described above, the linear operators $Q' \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V})\big)$ and $P \in C^\infty\big([0,T];\mathrm{Dom}(D)\cap\mathrm{Dom}(B)\big)$ satisfy the linear system of operator equations

$$\partial_t Q = AQ + BP,$$
$$\partial_t P = CQ + DP,$$

where $Q = \mathrm{id} + Q'$. We assume at time $t = 0$ that $Q'(0) = O$ so that $Q(0) = \mathrm{id}$ and $P(0) = P_0$ for some given $P_0 \in \mathrm{Dom}(D)\cap\mathrm{Dom}(B)$. We call the evolution equation for $P = P(t)$ the *base equation* and that for $Q = Q(t)$ the *auxiliary equation*.

Remark 3 The initial condition $Q(0) = \mathrm{id}$ and $P(0) = P_0$ means that the corresponding subspace $\mathbb{W}(0)$ is represented in the canonical coordinate chart of $\mathrm{Gr}(\mathbb{H},\mathbb{V})$. Hereafter we will assume that for $t \in [0,T]$ the subspace $\mathbb{W}(t)$ is representable in the canonical coordinate chart and in particular that $\det_2 Q(t) \neq 0$.

The base and auxiliary equations represent two essential ingredients in our prescription, which to be complete, requires a third crucial ingredient. This is to propose a relation between P and Q as follows.

Definition 3 (Riccati Relation) We assume there exists a $T > 0$ such that for $P \in C^\infty\big([0,T];\mathrm{Dom}(D)\cap\mathrm{Dom}(B)\big)$ and $Q' \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V})\big)$ there exists a linear operator $G \in C^\infty\big([0,T];\mathrm{Dom}(D)\cap\mathrm{Dom}(B)\big)$ satisfying the linear Fredholm equation

$$P = G\,Q,$$

where $Q = \mathrm{id} + Q'$. We call this the *Riccati Relation*.

Given solution linear operators $P = P(t)$ and $Q = Q(t)$ to the linear base and auxiliary equations we can prove the existence of a suitable solution $G = G(t)$ to the linear Fredholm equation constituting the Riccati relation. This result is proved in Beck et al. [8]. The result is as follows.

Lemma 1 (Existence and Uniqueness: Riccati relation) *Assume there exists a $T > 0$ such that $P \in C^\infty\big([0,T];\mathrm{Dom}(D)\cap\mathrm{Dom}(B)\big)$, $Q' \in C^\infty\big([0,T];\mathfrak{J}_2(\mathbb{V};\mathbb{V})\big)$ and $Q'(0) = O$. Then there exists a $T' > 0$ with $T' \leqslant T$ such that for $t \in [0,T']$ we*

have $\det_2\big(Q(t)\big) \neq 0$ *and* $\|Q'(t)\|_{\mathfrak{J}_2(\mathbb{V};\mathbb{V})} < 1$. *In particular, there exists a unique solution* $G \in C^\infty\big([0, T']; \mathrm{Dom}(D) \cap \mathrm{Dom}(B)\big)$ *to the Riccati relation.*

The proof utilizes the fact that we assume the solutions $P(t) \in \mathrm{Dom}(D) \cap \mathrm{Dom}(B)$ and $Q(t) \in \mathfrak{J}_2(\mathbb{V}; \mathbb{V})$ are smooth in time and at time $t = 0$ we have $\det_2\big(Q(0)\big) = 1$ and $\|Q'(0)\|_{\mathfrak{J}_2(\mathbb{V};\mathbb{V})} = 0$. Hence for a short time we are guaranteed that $\det_2\big(Q(t)\big)$ is non-zero and $\|Q'(t)\|_{\mathfrak{J}_2(\mathbb{V};\mathbb{V})}$ is sufficiently small to provide suitable bounds on $G(t) = P(t)Q^{-1}(t)$. Our main result now is that the solution $G = G(t)$ to the Riccati relation satisfies a quadratically nonlinear evolution equation as follows.

Theorem 1 (Riccati evolution equation) *Suppose we are given the initial data* $G_0 \in \mathrm{Dom}(D) \cap \mathrm{Dom}(B)$ *and that* $Q'(0) = O$ *and* $P(0) = G_0$. *Assume for some* $T > 0$ *that* $P \in C^\infty\big([0, T]; \mathrm{Dom}(D) \cap \mathrm{Dom}(B)\big)$, $Q' \in C^\infty\big([0, T]; \mathfrak{J}_2(\mathbb{V}; \mathbb{V})\big)$ *satisfy the linear base and auxiliary equations and that* $G \in C^\infty\big([0, T]; \mathrm{Dom}(D) \cap \mathrm{Dom}(B)\big)$ *solves the Riccati relation. Then this solution* G *to the Riccati relation necessarily satisfies* $G(0) = G_0$ *and for* $t \in [0, T]$ *solves the Riccati evolution equation*

$$\partial_t G = C + DG - G\,(A + BG).$$

Proof By direct computation, if we differentiate the Riccati relation $P = G\,Q$ with respect to time using the product rule and use that P and Q satisfy the linear base and auxiliary equations we find $(\partial_t G)\,Q = \partial_t P - G\,\partial_t Q = DP - G\,(AQ + BP) = (DG)\,Q - G\,(A + BG)\,Q$. Postcomposing by Q^{-1} establishes the result. □

We now consider the abstract development above at the partial differential equation level with an eye towards our applications. All our assumptions hitherto in this section apply here as well. As hinted in our Introduction and indicated more explicitly at the end of Sect. 2, suppose our underlying separable Hilbert space is $\mathbb{H} = L^2(\mathbb{I}; \mathbb{R}) \times L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})$, where the continuous interval $\mathbb{I} \subseteq \mathbb{R}$. The function space $L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R}) \subseteq L^2(\mathbb{I}; \mathbb{R})$ is a subspace of $L^2(\mathbb{I}; \mathbb{R})$ which we will explicitly define presently. We assume the closed subspace $\mathbb{V}$ of $\mathbb{H}$ to be $\mathbb{V} \cong L^2(\mathbb{I}; \mathbb{R})$. By assumption we know for some $T > 0$ the operators Q' and P are Hilbert–Schmidt operators with $Q' \in C^\infty\big([0, T]; \mathfrak{J}_2\big(L^2(\mathbb{I}; \mathbb{R}); L^2(\mathbb{I}; \mathbb{R})\big)\big)$ and $P \in C^\infty\big([0, T]; \mathfrak{J}_2\big(L^2(\mathbb{I}; \mathbb{R}); L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)\big)$. By standard theory the actions Q' and P can be represented by the integral kernel functions $q' = q'(x, y; t)$ and $p = p(x, y; t)$, respectively where

$$q' \in C^\infty\big([0, T]; L^2(\mathbb{I}^2; \mathbb{R})\big) \qquad \text{and} \qquad p \in C^\infty\big([0, T]; L^2\big(\mathbb{I}; L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)\big).$$

Henceforth we assume that the operator B is multiplicative corresponding to multiplication by the smooth, bounded, square-integrable and real-valued function $b = b(x)$. We also assume the operator D is the unbounded operator $d = d(\partial_x)$ which is a polynomial in ∂_x with constant real-valued coefficients. We can now specify $L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R}) \subseteq L^2(\mathbb{I}; \mathbb{R})$, it corresponds to the domain of the operator $d = d(\partial_x)$. Also by assumption A and C are Hilbert–Schmidt valued

operators and can thus be represented by integral kernel functions $a = a(x, y; t)$ and $c = c(x, y; t)$, respectively, where $a \in C^\infty\big([0, T]; L^2(\mathbb{I}^2; \mathbb{R})\big)$ and $c \in C^\infty\big([0, T]; L^2\big(\mathbb{I}; L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)\big)$ The linear base and auxiliary equations, since $Q(t) = \mathrm{id} + Q'(t)$, thus have the form

$$\partial_t q'(x, y; t) = a(x, y; t) + \int_{\mathbb{I}} a(x, z; t) q'(z, y; t)\, \mathrm{d}z + b(x)\, p(x, y; t), \tag{8a}$$

$$\partial_t p(x, y; t) = c(x, y; t) + \int_{\mathbb{I}} c(x, z; t) q'(z, y; t)\, \mathrm{d}z + d(\partial_x)\, p(x, y; t). \tag{8b}$$

Remark 4 To be consistent with our assumptions on the properties of $P = P(t)$ and its corresponding integral kernel $p = p(x, y; t)$ for $t \in [0, T]$ as outlined above, we must suitably restrict the choice of the class of operator $d = d(\partial_x)$ appearing in the base equation. In addition to the class properties outlined just above, we assume henceforth, and in particular for all our applications in Sect. 4, that $d = d(\partial_x)$ is diffusive or dispersive as a polynomial operator in ∂_x. Hence for example we could assume that d is a polynomial in only even degree terms in ∂_x with the $2N$th degree term having a real coefficient of sign $(-1)^{N+1}$. Alternatively for example d could have a dispersive form such as $d = -\partial_x^3$.

We are now in a position to prove our main result for evolutionary partial differential equations with nonlocal quadratic nonlinearities.

Corollary 1 (Grassmannian evolution equation) *Given the initial data $g_0 \in C^\infty\big(\mathbb{I}^2; \mathbb{R}\big) \cap L^2\big(\mathbb{I}; L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)$ suppose $q' = q'(x, y; t)$ and $p = p(x, y; t)$ are the solutions to the linear evolutionary base and auxiliary equations (8) with $p(x, y; 0) = g_0(x, y)$ and $q'(x, y; 0) = 0$. Suppose the operator $d = d(\partial_x)$ is of the diffusive or dispersive form described in Remark 4. Then there is a $T > 0$ such that the solution $g \in C^\infty\big([0, T]; L^2\big(\mathbb{I}; L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)\big)$ to the linear Fredholm equation*

$$p(x, y; t) = g(x, y; t) + \int_{\mathbb{I}} g(x, z; t)\, q'(z, y; t)\, \mathrm{d}z, \tag{9}$$

solves the evolutionary partial differential equation with quadratic nonlocal nonlinearities of the form

$$\partial_t g(x, y; t) = c(x, y; t) + d(\partial_x)\, g(x, y; t) - \int_{\mathbb{I}} g(x, z; t)\big(a(z, y; t) + b(z)\, g(z, y; t)\big)\, \mathrm{d}z.$$

Proof That for some $T > 0$ there exists a solution $g \in C^\infty\big([0, T]; L^2\big(\mathbb{I}; L^2_{\mathrm{d}}(\mathbb{I}; \mathbb{R})\big)\big)$ to the linear Fredholm equation (9), i.e. the Riccati relation, follows from Lemma 1 and our assumptions on $q' = q'(x, y; t)$ and $p = p(x, y; t)$ outlined above. That this solution g also solves the evolutionary partial differential equation with quadratic nonlocal nonlinearity shown follows from Theorem 1. □

It is instructive to see the proof of the second of the results from Corollary 1 at the integral kernel level, i.e. the proof that the solution g to the linear Fredholm equation (9) also solves the evolutionary partial differential equation with quadratic nonlocal nonlinearity shown. We present this here. First we differentiate the linear Fredholm equation (9) with respect to time. This generates the relation

$$\partial_t g(x,y;t) + \int_{\mathbb{I}} \partial_t g(x,z;t)\, q'(z,y;t)\,\mathrm{d}z = \partial_t p(x,y;t) - \int_{\mathbb{I}} g(x,z;t)\, \partial_t q'(z,y;t)\,\mathrm{d}z.$$

Second we substitute for $\partial_t q'$ and $\partial_t p$ using their evolution equations. Let us consider the first term on the right above. We find that

$$\begin{aligned}
\partial_t p(x,y;t) &= \int_{\mathbb{I}} c(x,z;t)\,\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z + d(\partial_x)p(x,y;t) \\
&= \int_{\mathbb{I}} c(x,z;t)\,\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z \\
&\quad + d(\partial_x)\bigg(g(x,y;t) + \int_{\mathbb{I}} g(x,z;t)\,q'(z,y;t)\,\mathrm{d}z\bigg) \\
&= \int_{\mathbb{I}} c(x,z;t)\,\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z \\
&\quad + d(\partial_x)\int_{\mathbb{I}} g(x,z;t)\,\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z \\
&= \int_{\mathbb{I}} \big(c(x,z;t) + d(\partial_x)g(x,z;t)\big)\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z.
\end{aligned}$$

Now consider the second term on the right above. We observe

$$\begin{aligned}
&\int_{\mathbb{I}} g(x,z;t)\,\partial_t q'(z,y;t)\,\mathrm{d}z \\
&= \int_{\mathbb{I}} g(x,z;t)\bigg(\int_{\mathbb{I}} a(z,\zeta;t)\,\big(\delta(\zeta-y) + q'(\zeta,y;t)\big)\,\mathrm{d}\zeta\bigg)\mathrm{d}z \\
&\quad + \int_{\mathbb{I}} g(x,z;t)\,\big(b(z)p(z,y;t)\big)\,\mathrm{d}z \\
&= \int_{\mathbb{I}} g(x,z;t)\bigg(\int_{\mathbb{I}} a(z,\zeta;t)\,\big(\delta(\zeta-y) + q'(\zeta,y;t)\big)\,\mathrm{d}\zeta\bigg)\mathrm{d}z \\
&\quad + \int_{\mathbb{I}} g(x,z;t)\bigg(b(z)\int_{\mathbb{I}} g(z,\zeta;t)\,\big(\delta(\zeta-y) + q'(\zeta,y;t)\big)\,\mathrm{d}\zeta\bigg)\mathrm{d}z \\
&= \int_{\mathbb{I}}\bigg(\int_{\mathbb{I}} g(x,\zeta;t)\big(a(\zeta,z;t) + b(\zeta)g(\zeta,z;t)\big)\,\mathrm{d}\zeta\bigg)\big(\delta(z-y) + q'(z,y;t)\big)\,\mathrm{d}z.
\end{aligned}$$

Putting these results together and post-composing by the operator Q^{-1} generates the required result. Another way to enact this last step is to postmultiply the final combined result by '$\delta(y-\eta)+\tilde{q}'(y,\eta;t)$' for some $\eta \in \mathbb{I}$. This is the kernel associated with the inverse operator $Q^{-1} = \mathrm{id}+\tilde{Q}'$ to $Q = \mathrm{id}+Q'$. Then integrating over $y \in \mathbb{I}$ gives the result for $g = g(x,\eta;t)$.

Remark 5 (Nonlocal nonlinearities with derivatives) In the linear base and auxiliary equations (8) we could take b to be a constant coefficient polynomial in ∂_x. With minor modifications the results we derive above still apply.

We now need to demonstrate as a practical procedure, how linear evolutionary partial differential equations for p and q' generate solutions to the evolutionary partial differential equation with quadratic nonlocal nonlinearities at hand. We show this explicitly through two examples in the next section.

4 Examples

We now consider some evolutionary partial differential equations with nonlocal quadratic nonlinearities and explicitly show how to generate solutions to them from the linear base and auxiliary equations and linear Riccati relation. In both examples we take $\mathbb{I} \coloneqq \mathbb{R}$. Note that throughout we define the Fourier transform for any given function $f = f(x)$ and its inverse as

$$\hat{f}(k) \coloneqq \int_{\mathbb{R}} f(x)\mathrm{e}^{2\pi \mathrm{i}kx}\,\mathrm{d}x \qquad \text{and} \qquad f(x) \coloneqq \int_{\mathbb{R}} \hat{f}(k)\mathrm{e}^{-2\pi \mathrm{i}kx}\,\mathrm{d}k.$$

Example 1 (Nonlocal convolution nonlinearity) In this case the target evolutionary partial differential equation has a quadratic nonlinearity in the form of a convolution and is given by

$$\partial_t g = d\,g - g \star g,$$

where $d = d(\partial_x)$ and the $\star$ operation here does indeed represent convolution. In other words for this example we suppose

$$\big(g \star g\big)(x;t) = \int_{\mathbb{R}} g(x-z;t)\,g(z;t)\,\mathrm{d}z.$$

We assume smooth and square-integrable initial data $g_0 = g_0(x)$.

To find solutions via our approach, we begin by assuming the kernel g of the operator G has the convolution form $g = g(x-y;t)$. We further assume the linear base and auxiliary equations have the form

$$\partial_t p(x,y;t) = d(\partial_x)\,p(x,y;t),$$
$$\partial_t q'(x,y;t) = b(x)\,p(x,y;t),$$

with in fact $b \equiv 1$. In addition we suppose $d = d(\partial_x)$ is of diffusive or dispersive form as described in Remark 4. In this case the Grassmannian evolution equation in Corollary 1 has the form

$$\partial_t g(x-y;t) = d(\partial_x)\, g(x-y;t) - \int_{\mathbb{R}} g(x-z;t)\, g(z-y;t)\, \mathrm{d}z,$$

which by setting $y = 0$ matches the system under consideration. We verify the sufficient conditions for Corollary 1 to apply presently. In Fourier space our example partial differential equation naturally takes the form

$$\partial_t \hat{g} = d(2\pi \mathrm{i}k)\, \hat{g} - \hat{g}^2. \tag{10}$$

We generate solutions to the given partial differential equation for g from the linear base and auxiliary equations, for the given initial data g_0, as follows. Note the base equation has the following equivalent form and solution in Fourier space:

$$\partial_t \hat{p}(k,y;t) = d(2\pi \mathrm{i}k)\, \hat{p}(k,y;t) \quad \Leftrightarrow \quad \hat{p}(k,y;t) = \mathrm{e}^{d(2\pi \mathrm{i}k)\, t}\, \hat{p}_0(k,y).$$

Here $\hat{p}_0$ is the Fourier transform of the initial data for p. In Fourier space the auxiliary equation has the form and solution:

$$\partial_t \hat{q}'(k,y;t) = \hat{p}(k,y;t) \quad \Leftrightarrow \quad \hat{q}'(k,y;t) - \hat{q}'_0(k,y) = \frac{\mathrm{e}^{d(2\pi \mathrm{i}k)\, t} - 1}{d(2\pi \mathrm{i}k)}\, \hat{p}_0(k,y).$$

Here $\hat{q}'_0(k,y)$ is the Fourier transform of the initial data for $\hat{q}'$. As per the general theory, we suppose $\hat{q}'_0(k,y) = 0$. This means if we set $t = 0$ in the Riccati relation we find

$$p_0(x,y) = g_0(x-y) \qquad \Leftrightarrow \qquad \hat{p}_0(k,y) = \mathrm{e}^{2\pi \mathrm{i}ky}\, \hat{g}_0(k).$$

where g_0 is the initial data for the partial differential equation for g. Hence explicitly we have

$$\hat{p}(k,y;t) = \mathrm{e}^{d(2\pi \mathrm{i}k)\, t}\, \mathrm{e}^{2\pi \mathrm{i}ky}\, \hat{g}_0(k) \quad \text{and} \quad \hat{q}'(k,y;t) = \frac{\mathrm{e}^{d(2\pi \mathrm{i}k)\, t} - 1}{d(2\pi \mathrm{i}k)}\, \mathrm{e}^{2\pi \mathrm{i}ky}\, \hat{g}_0(k).$$

Note by taking the inverse Fourier transform, we deduce that $p = p(x-y;t)$ and $q' = q'(x-y;t)$. From these explicit forms for their Fourier transforms, we deduce there exists a $T > 0$ such that on the time interval $[0, T]$ we know p and q' have the regularity required so that Corollary 1 applies. Further, the Riccati relation in this

case is

$$p(x,y;t) = g(x-y;t) + \int_{\mathbb{R}} g(x-z;t)\, q'(z,y;t)\,\mathrm{d}z$$

$$\Leftrightarrow \qquad \hat{p}(k,y;t) = \hat{g}(k;t)\big(\mathrm{e}^{2\pi \mathrm{i} k y} + \hat{q}'(k,y;t)\big).$$

Thus using the expressions for $\hat{p}$ and $\hat{q}'$ above we find that

$$\hat{g}(k;t) = \frac{\mathrm{e}^{d(2\pi \mathrm{i}k)\,t}\,\hat{g}_0(k)}{1 + \Big(\big(\mathrm{e}^{d(2\pi \mathrm{i}k)\,t} - 1\big)/d(2\pi \mathrm{i}k)\Big)\,\hat{g}_0(k)}.$$

Direct substitution into the Fourier form (10) of our example partial differential equation verifies it is indeed the solution for the initial data g_0.

In Fig. 1 we show the solution to the nonlocal quadratically nonlinear partial differential equation above, for $d = \partial_x^2 + 1$ and a given generic initial profile g_0. The left panel shows the evolution of the solution profile computed using a direct integration approach. By this we mean we approximated ∂_x^2 by the central difference formula and computed the nonlinear convolution by computing the inverse Fourier transform of $\big(\hat{g}(k)\big)^2$. We used the inbuilt Matlab integrator `ode23s` to integrate in time. Similar direct integration could be achieved by integrating the differential equation (10) for $\hat{g}$ using `ode23s` and then computing the inverse Fourier transform. The right panel in Fig. 1 shows the solution evolution computed using our Riccati approach. As expected, the solutions look identical (up to

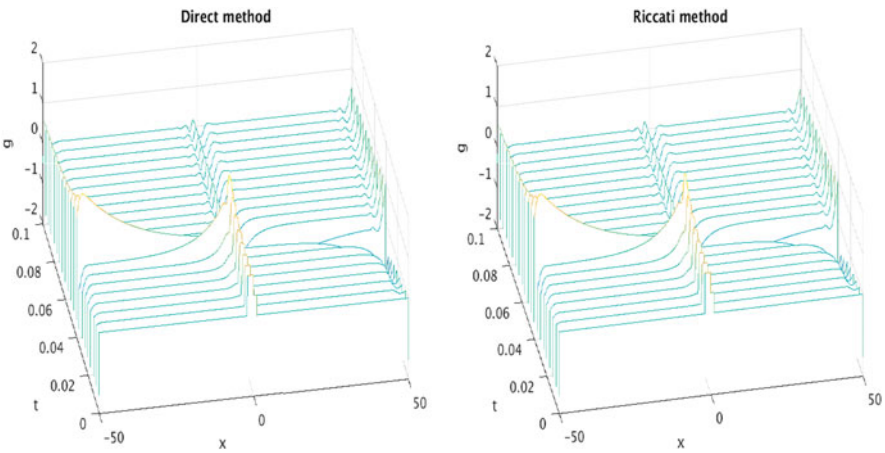


Fig. 1 We plot the solution to the nonlocal quadratically nonlinear partial differential equation from Example 1. We used a generic initial profile g_0 as shown. The left panel shows the solution computed using a direct integration approach while the right panel shows the solution computed using our Riccati approach

numerical precision), even when we continue the solution past the time when the diffusion has reached the boundaries of the finite domain of integration in x, roughly half way along the interval of evolution shown.

Remark 6 (Multi-dimensions) This last example extends to the case where $x, y \in \mathbb{R}^n$ for any $n \geqslant 1$ when d is a scalar operator such as a power of the Laplacian, with p, q' and g all scalar.

Example 2 (Nonlocal quadratic nonlinearity with correlation) In this case the target evolutionary partial differential equation has a nonlocal quadratic nonlinearity involving a correlation function and has the form

$$\partial_t g(x, y; t) = d(\partial_x)\, g(x, y; t) - \int_{\mathbb{R}} g(x, z; t)\, b(z)\, g(z, y; t)\, \mathrm{d}z.$$

This corresponds to the evolutionary partial differential equation with nonlocal quadratic nonlinearity in Corollary 1, with $a = c = 0$ and $b = b(x)$ the scalar smooth, bounded square-integrable function described in the paragraphs preceding it. We also assume that $d = d(\partial_x)$ is of the diffusive or dispersive form described in Remark 4. We assume smooth and square-integrable initial data $g_0 = g_0(x, y)$.

To find solutions to the evolutionary partial differential equation just above using our approach we assume the linear base and auxiliary equations have the form

$$\partial_t p(x, y; t) = d(\partial_x)\, p(x, y; t),$$
$$\partial_t q'(x, y; t) = b(x)\, p(x, y; t).$$

In Fourier space the solution of the base equation has the form

$$\hat{p}(k, y; t) = \mathrm{e}^{d(2\pi \mathrm{i} k)\, t}\, \hat{p}_0(k, y),$$

where $\hat{p}_0$ is the Fourier transform of the initial data for p. The auxiliary equation solution in Fourier space has the form,

$$\hat{q}'(k, y; t) = \int_{\mathbb{R}} \hat{b}(k - \kappa)\, \hat{I}(\kappa, t)\, \hat{p}_0(\kappa, y)\, \mathrm{d}\kappa,$$

where we set

$$\hat{I}(k, t) := \frac{\mathrm{e}^{d(2\pi \mathrm{i} k)\, t} - 1}{d(2\pi \mathrm{i} k)}.$$

As in the last example we took the initial data for q' to be zero and thus the initial data for $\hat{q}'$ is also zero. We also set the initial data for $p = p(x, y; t)$ to be $p_0(x, y) = g_0(x, y)$. In Fourier space this is equivalent to $\hat{p}_0(k, y) = \hat{g}_0(k, y)$. We now derive an explicit form for $q' = q'(x, y; t)$ from $\hat{q}' = \hat{q}'(k, y; t)$ above.

Taking the inverse Fourier transform of $\hat{q}'$, we find that

$$\begin{aligned} q'(x,y;t) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathrm{e}^{-2\pi \mathrm{i} k x}\, \hat{b}(k-\kappa)\, \mathrm{d}k \right) \hat{I}(\kappa,t)\, \hat{g}_0(\kappa,y)\, \mathrm{d}\kappa \\ &= \int_{\mathbb{R}} \left(\mathrm{e}^{-2\pi \mathrm{i} \kappa x}\, b(x) \right) \hat{I}(\kappa,t)\, \hat{g}_0(\kappa,y)\, \mathrm{d}\kappa \\ &= b(x) \int_{\mathbb{R}} \mathrm{e}^{-2\pi \mathrm{i} \kappa x}\, \hat{I}(\kappa,t)\, \hat{g}_0(\kappa,y)\, \mathrm{d}\kappa \\ &= b(x) \int_{\mathbb{R}} I(x-\zeta,t)\, g_0(\zeta,y)\, \mathrm{d}\zeta. \end{aligned}$$

Lastly, the Riccati relation here has the form

$$p(x,y;t) = g(x,y;t) + \int_{\mathbb{R}} g(x,z;t)\, q'(z,y;t)\, \mathrm{d}z.$$

Since we have an explicit expression for $q' = q'(x,y;t)$, and we can obtain one for $p = p(x,y;t)$ by taking the inverse Fourier transform of the explicit expression for $\hat{p} = \hat{p}(k,y;t)$ above, we can solve this linear Fredholm equation for $g = g(x,y;t)$. The solution, by Corollary 1, will be the solution to the evolutionary partial differential equation with the nonlocal quadratic nonlinearity above corresponding to the initial data $g_0(x,y)$.

We solved the Fredholm equation for $g = g(x,y;t)$ numerically. The results are shown in Fig. 2. We set the operator $d = \partial_x^2 + 1$ and took as the generic initial profile $g_0(x,y) := \mathrm{sech}(x+y)\,\mathrm{sech}(y)$. We set $b = b(x)$ to be a mean-zero Gaussian density function with standard deviation 0.01. The top panel in Fig. 2 shows the initial data. The middle panel in the figure shows the solution profile computed at time $t = 2$ using a direct spectral integration approach. By this we mean we solved the equation for $\hat{g}(k,y;t)$ generated by taking the Fourier transform of the equation for $g = g(x,y;t)$. We used the inbuilt Matlab integrator `ode23s` to integrate in time. The bottom panel in Fig. 2 shows the solution computed with the time parameter $t = 2$ using our Riccati approach, i.e. by numerically solving the Fredholm equation for $g = g(x,y;t)$ above by standard methods for such integral equations. As expected, the solutions in the middle and bottom panels look identical (up to numerical precision).

Remark 7 We emphasize that, when we can explicitly solve for $p = p(x,y;t)$ and $q' = q'(x,y;t)$ in our Riccati approach, then time t plays the role of a parameter. One decides the time at which one wants to compute the solution and we then solve

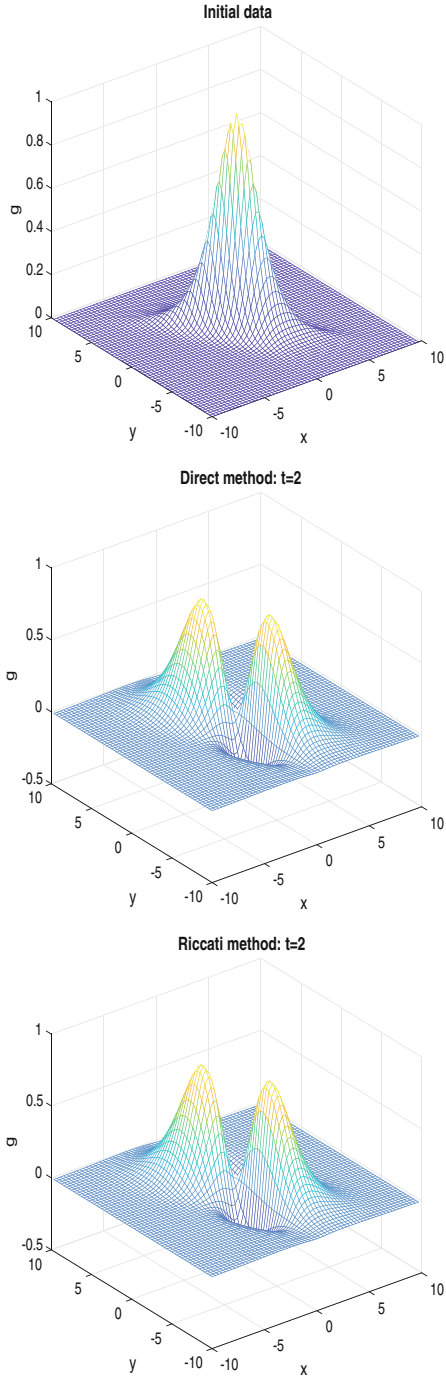


Fig. 2 We plot the solution to the nonlocal quadratically nonlinear partial differential equation with correlation from Example 2. We used a generic initial profile g_0 as shown in the top panel. For time $t = 2$, the middle panel shows the solution computed using a direct integration approach while the bottom panel shows the solution computed using our Riccati approach

the Fredholm equation to generate the solution $g = g(x, y; t)$ for that time t. This is one of the advantages of our method over standard numerical schemes.[1]

Remark 8 (Burgers' equation) Burgers' equation can be considered as a special case of our Riccati approach in the following sense. Suppose the linear base and auxiliary equations are $\partial_t p(x;t) = \partial_x^2 p(x;t)$ and $\partial_t q(x;t) = \partial_x p(x;t)$ for the real valued functions p and q. Further suppose the Riccati relation takes the form $p(x;t) = g(x;t)\, q(x;t)$ where g is also real valued. Note this represents a rank one relation between p and q in the sense that we obtain p from q by a simple multiplication of q by the function g. From the linear base and auxiliary equations, assuming smooth solutions, we deduce that $\partial_t q = \partial_x p = \partial_x^{-1}\partial_x^2 p = \partial_x^{-1}\partial_t p = \partial_t(\partial_x^{-1} p)$, where $\partial_x^{-1} w$ represents the operation $\int_{-\infty}^{x} w(z)\,\mathrm{d}z$ for any smooth integrable function $w = w(x)$ on $\mathbb{R}$. From the above equalities we deduce $p(x;t) = \partial_x q(x;t) + f(x)$ where $f = f(x)$ is an arbitrary function of x only. If we take the special case $f \equiv 0$, then we deduce $p(x;t) = \partial_x q(x;t)$. This also implies $\partial_t q = \partial_x^2 q$. If we insert the relation $p(x;t) = \partial_x q(x;t)$ into the Riccati relation we find

$$g(x;t) = \frac{\partial_x q(x;t)}{q(x;t)}.$$

This is almost the Cole–Hopf transformation, it's just missing the usual '-2' factor on the right-hand side. However carrying through our Riccati approach by direct computation, differentiating the Riccati relation with respect to time, we observe

$$\begin{aligned}
(\partial_t g)\, q &= \partial_t p - g\, \partial_t q \\
&= \partial_x^2 p - g\, \partial_t q \\
&= \partial_x^2 (g\, q) - g\, \partial_t q \\
&= (\partial_x^2 g)\, q + 2\, (\partial_x g)\, \partial_x q + g\, (\partial_x^2 q) - g\, \partial_t q \\
&= (\partial_x^2 g)\, q + 2\, (\partial_x g)\, p \\
&= (\partial_x^2 g)\, q + 2\, (\partial_x g)\, g\, q.
\end{aligned}$$

If we divide through by the function $q = q(x;t)$ we conclude that $g = g(x;t)$ satisfies the nonlinear partial differential equation

$$\partial_t g = \partial_x^2 g + 2\, g\, \partial_x g.$$

However we now observe that '$-2\, g$' indeed satisfies Burgers' equation.

[1] We quote from the referee: "numerical integration in time will usually become inaccurate for large time t, but the nature of the exact solution gives you a precise answer for arbitrary t, and maybe allows access to information about long time behaviour which is inaccessible via standard numerical schemes."

5 Conclusions

There are many extensions of our approach to more general nonlinear partial differential equations. One immediate extension to consider is to multi-dimensions, i.e. where the underlying spatial domain lies in $\mathbb{R}^n$ for some $n \geqslant 1$. This should be straightforward as indicated in Remark 6 above. Another immediate extension is to systems of nonlinear partial differential equations with nonlocal nonlinearites. Indeed we explicitly consider this extension in Beck, Doikou, Malham and Stylianidis [8] where we demonstrate how to generate solutions to certain classes of reaction-diffusion systems with nonlocal quadratic nonlinearities. We also demonstrate therein, how to extend our approach to generate solutions to evolutionary partial differential equations with higher degree nonlocal nonlinearities, including the nonlocal nonlinear Schrödinger equation. Further therein, for arbitrary initial data $g_0 = g_0(x)$, we use our Riccati approach to generate solutions to the nonlocal Fisher–Kolmogorov–Petrovskii–Piskunov equation for scalar $g = g(x;t)$ of the form

$$\partial_t g(x;t) = d(\partial_x)\, g(x;t) - g(x;t) \int_{\mathbb{R}} g(z;t)\, \mathrm{d}z.$$

This has recently received some attention; see Britton [13] and Bian, Chen and Latos [9]. We would also like to consider the extension of our approach to the full range of possible choices of the operators d and b both as unbounded and bounded operators, for example to fractional and nonlocal diffusion cases. We have already considered the extension of our approach to evolutionary stochastic partial differential equations with nonlocal nonlinearities in Doikou, Malham and Wiese [18]. Therein we consider the separate cases when the driving space-time Wiener field appears as a nonhomogeneous additive source term or as a multiplicative but linear source term. Of course, another natural extension is to determine whether we can include the generation of solutions to evolutionary partial differential equations with local nonlinearities within the context of our Riccati approach. One potential approach is to suppose the Riccati relation is of Volterra type. This is an ongoing investigation. Lastly we remark that for the classes of nonlinear partial differential equations we can consider, solution singularities correspond to poor choices of coordinate patches which are related to function space regularity. In principle solutions can be continued by changing coordinate patches; see Schiff and Shnider [44] and Ledoux et al. [31]. This is achieved by pulling back the flow to the relevant general linear group and then projecting down to a more appropriate coordinate patch of the Fredholm Grassmannian. Alternatively, we could continue the flow in the appropriate general linear group via the base and auxiliary equations, and then monitor the relevant projection(s).

Acknowledgements We are very grateful to the referee for their detailed report and suggestions that helped significantly improve the original manuscript. We would like to thank Percy Deift, Kurusch Ebrahimi–Fard and Anke Wiese for their extremely helpful comments and suggestions. The work of M.B. was partially supported by US National Science Foundation grant DMS-1411460.

References

1. Abbondandolo, A., Majer, P.: Infinite dimensional Grassmannians. J. Oper. Theory **61**(1), 19–62 (2009)
2. Ablowitz, M.J., Ramani, A., Segur, H.: A connection between nonlinear evolution equations and ordinary differential equations of P-type II. J. Math. Phys. **21**, 1006–1015 (1980)
3. Alexander, J.C., Gardner, R., Jones, C.K.R.T.: A topological invariant arising in the stability analysis of traveling waves. J. Reine Angew. Math. **410**, 167–212 (1990)
4. Balazs, P.: Hilbert–Schmidt operators and frames—classification, best approximation by multipliers and algorithms. Int. J. Wavelets Multiresolution Inf. Process. **6**(2), 315–330 (2008)
5. Bauhardt, W., Pöppe, C.: The Zakharov–Shabat inverse spectral problem for operators. J. Math. Phys. **34**(7), 3073–3086 (1993)
6. Beals, R., Coifman, R.R.: Linear spectral problems, non-linear equations and the $\overline{\partial}$-method. Inverse Prob. **5**, 87–130 (1989)
7. Beck, M., Malham, S.J.A.: Computing the Maslov index for large systems. PAMS **143**, 2159–2173 (2015)
8. Beck, M., Doikou, A., Malham, S.J.A., Stylianidis, I.: Partial differential systems with nonlocal nonlinearities: generation and solution. Philos. Trans. A **376**(2117) (2018). https://doi.org/10.1098/rsta.2017.0195
9. Bian, S., Chen, L., Latos, E.A.: Global existence and asymptotic behavior of solutions to a nonlocal Fisher–KPP type problem. Nonlinear Anal. **149**, 165–176 (2017)
10. Bittanti, S., Laub, A.J., Willems, J.C. (eds.): The Riccati Equation. Communications and Control Engineering Series. Springer, Berlin/Heidelberg (1991)
11. Blanchard, P., Brüning, E.: Mathematical Methods in Physics: Distributions, Hilbert Space Operators, Variational Methods, and Applications in Quantum Physics, 2nd edn. Progress in Mathematical Physics, vol. 69. Birkhäuser, Berlin (2015)
12. Bornemann, F.: Numerical evaluation of Fredholm determinants and Painlevé transcendents with applications to random matrix theory, talk at the Abdus Salam International Centre for Theoretical Physics (2009)
13. Britton, N.F.: Spatial structures and periodic travelling waves in an integro-differential reaction-diffusion population model. SIAM J. Appl. Math. **50**(6), 1663–1688 (1990)
14. Brockett, R.W., Byrnes, C.I.: Multivariable Nyquist criteria, root loci, and pole placement: a geometric viewpoint. IEEE Trans. Automat. Control **26**(1), 271–284 (1981)
15. Christensen, O.: Frames and Bases. Springer (2008). https://doi.org/10.1007/978-0-8176-4678-3_3
16. Deng, J., Jones, C.: Multi-dimensional Morse index theorems and a symplectic view of elliptic boundary value problems. Trans. Am. Math. Soc. **363**(3), 1487–1508 (2011)
17. Dodd, R.K., Eilbeck, J.C., Gibbon, J.D., Morris H.C.: Solitons and Non-linear Wave Equations. Academic, London (1982)
18. Doikou, A., Malham, S.J.A., Wiese, A.: Stochastic partial differential equations with nonlocal nonlinearities and their simulation. (2018, in preparation)
19. Drazin, P.G., Johnson, R.S.: Solitons: An Introduction. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (1989)
20. Dyson, F.J.: Fredholm determinants and inverse scattering problems. Commun. Math. Phys. **47**, 171–183 (1976)

21. Furutani, K.: Review: Fredholm–Lagrangian–Grassmannian and the Maslov index. J. Geom. Phys. **51**, 269–331 (2004)
22. Grellier, S., Gerard, P.: The cubic Szegö equation and Hankel operators (2015). arXiv:1508.06814
23. Griffiths, P., Harris, J.: Principles of Algebraic Geometry. Wiley Classics Library, New York (1994)
24. Guest, M.A.: From Quantum Cohomology to Integrable Systems. Oxford University Press, Oxford/New York (2008)
25. Hermann, R.: Cartanian Geometry, Nonlinear Waves, and Control Theory: Part A. Interdisciplinary Mathematics, vol. XX. Math Sci Press, Brookline (1979)
26. Hermann, R.: Cartanian Geometry, Nonlinear Waves, and Control Theory: Part B. Interdisciplinary Mathematics, vol. XXI. Math Sci Press, Brookline (1980)
27. Hermann, R., Martin, C.: Lie and Morse theory for periodic orbits of vector fields and matrix Riccati equations, I: general Lie-theoretic methods. Math. Syst. Theory **15**, 277–284 (1982)
28. Karambal, I., Malham, S.J.A.: Evans function and Fredholm determinants. Proc. R. Soc. A **471**(2174) (2015). https://doi.org/10.1098/rspa.2014.0597
29. McKean, H.P.: Fredholm determinants. Cent. Eur. J. Math. **9**(2), 205–243 (2011)
30. Ledoux, V., Malham, S.J.A., Niesen, J., Thümmler, V.: Computing stability of multidimensional travelling waves. SIAM J. Appl. Dyn. Syst. **8**(1), 480–507 (2009)
31. Ledoux, V., Malham, S.J.A., Thümmler, V.: Grassmannian spectral shooting. Math. Comput. **79**, 1585–1619 (2010)
32. Martin, C., Hermann, R.: Applications of algebraic geometry to systems theory: the McMillan degree and Kronecker indicies of transfer functions as topological and holomorphic system invariants. SIAM J. Control Optim. **16**(5), 743–755 (1978)
33. Miura, R.M.: The Korteweg–De Vries equation: a survey of results. SIAM Rev. **18**(3), 412–459 (1976)
34. Miwa, T., Jimbo, M., Date, E.: Solitons: Differential Equations, Symmetries and Infinite Dimensional Algebras. Cambridge University Press, Cambridge (2000)
35. Piccione, P., Tausk, D.V.: A Student's Guide to Symplectic Spaces, Grassmannians and Maslov Index (2008). www.ime.usp.br/~piccione/Downloads/MaslovBook.pdf
36. Pöppe, C.: Construction of solutions of the sine-Gordon equation by means of Fredholm determinants. Physica D **9**, 103–139 (1983)
37. Pöppe, C.: The Fredholm determinant method for the KdV equations. Physica D **13**, 137–160 (1984)
38. Pöppe, C.: General determinants and the τ function for the Kadomtsev–Petviashvili hierarchy. Inverse Prob. **5**, 613–630 (1984)
39. Pöppe, C., Sattinger, D.H.: Fredholm determinants and the τ function for the Kadomtsev–Petviashvili hierarchy. Publ. RIMS Kyoto Univ. **24**, 505–538 (1988)
40. Pressley, A., Segal, G.: Loop Groups, Oxford Mathematical Monographs. Clarendon Press, Oxford (1986)
41. Reed, M., Simon, B.: Methods of Modern Mathematical Physics: I Functional Analysis. Academic, New York/London (1980)
42. Sato, M.: Soliton equations as dynamical systems on a infinite dimensional Grassmann manifolds. RIMS **439**, 30–46 (1981)
43. Sato, M.: The KP hierarchy and infinite dimensional Grassmann manifolds. Proc. Symposia Pure Math. **49**(Part 1), 51–66 (1989)
44. Schiff, J., Shnider, S.: A natural approach to the numerical integration of Riccati differential equations. SIAM J. Numer. Anal. **36**(5), 1392–1413 (1999)
45. Segal, G., Wilson, G.: Loop groups and equations of KdV type. Inst. Hautes Etudes Sci. Publ. Math. N **61**, 5–65 (1985)
46. Simon, B.: Trace Ideals and Their Applications, 2nd edn. Mathematical Surveys and Monographs, vol. 120. AMS, Providence (2005)
47. Tracy, C.A., Widom, H.: Fredholm determinants and the mKdV/Sinh-Gordon hierarchies. Commun. Math. Phys. **179**, 1–10 (1996)

48. Wilson, G.: Infinite-dimensional Lie groups and algebraic geometry in soliton theory. Trans. R. Soc. Lond. A **315**(1533), 393–404 (1985)
49. Zakharov, V.E., Shabat, A.B.: A scheme for integrating the non-linear equation of mathematical physics by the method of the inverse scattering problem I. Funct. Anal. Appl. **8**, 226 (1974)
50. Zelikin, M.I.: Control Theory and Optimization I. Encyclopedia of Mathematical Sciences, vol. 86. Springer, Berlin/Heidelberg (2000)

Gog and Magog Triangles

Philippe Biane

Abstract We survey the problem of finding an explicit bijection between Gog and Magog triangles, a combinatorial problem which has been open since the 1980s. We give some of the ideas behind a recent approach to this question and also prove some properties of the distribution of inversions and coinversions in Gog triangles.

1 Introduction

An alternating sign matrix is a square matrix having coefficients in $\{-1, 0, 1\}$ so that, in each row and in each column, if one forgets the zeros, the 1 and -1 entries alternate and the sum is 1, e.g.

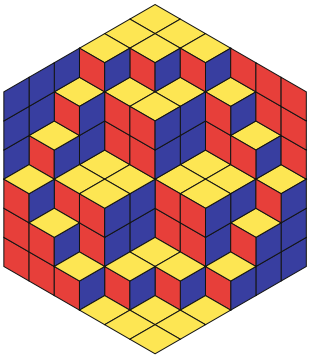
$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

These matrices were investigated by Robbins and Rumsey [24] in 1986 as a generalization of permutation matrices, after they discovered some startling properties of the λ-determinant, a deformation of usual determinants of matrices.

Around the same time interest in the enumeration of plane partitions led to the question of enumerating several symmetry classes of plane partitions. Among these classes the so-called totally symmetric self complementary plane partitions (TSSCPP in short), as the one below,

P. Biane (✉)
Institut Gaspard-Monge, Université Paris-Est Marne-la-Vallée, Marne-la-Vallée cedex 2, France
e-mail: biane@univ-mlv.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_4



stood as the most symmetric ones. Empirical data showed that both the enumeration of ASM and of TSSCPP according to their size started with the numbers $1, 2, 7, 42, 429, 7436, 218348, 10850216\ldots$ which could be put into the closed form $A_n = \prod_{j=0}^{n-1} \frac{(3j+1)!}{(n+j)!}$. There was then two main problems: prove the enumeration formula and find a bijective proof of the coincidence, if it exists. As we explain below, the enumeration problems were finally solved around 1994/5, but the proofs gave no hint as to the existence of a natural bijection between these objects. It turns out that both classes of objects, the ASM and the TSSCPP, can be encoded as triangles of positive integers whose rows interlace, like the following (such triangles are usually called Gelfand-Tsetlin triangles in representation theory where they occur as labels for bases of irreducible representations)

$$\begin{array}{ccccccccc} & & 1 & & 1 & & 3 & & 5 & & 5 \\ & & & 1 & & 2 & & 4 & & 5 & \\ & & & & 2 & & 2 & & 5 & & \\ & & & & & 2 & & 3 & & & \\ & & & & & & 3 & & & & \end{array}$$

Thus one can reformulate the problem as finding a bijection between two species of such Gelfand-Tsetlin triangles, which makes this problem completely elementary.

The purpose of this paper is to present a recent approach to this question which has led to some progress towards the solution of the bijection problem. In particular we will introduce a new class of objects, the GOGAm triangles and present some new conjectures. We also study the distribution of inversions and coinversions in Gog triangles. In particular we determine the pairs (p, q) for which there exists Gog triangles of a given size with p inversions and q coinversions.

The paper is organized as follows. In the first part we present a short survey of plane partitions and their enumeration, in particular the TSSCPP and the Magog triangles. In the second part we present the other combinatorial objects in which we are interested, the alternating sign matrices. As we briefly explain, these objects arise in many places of algebra, representation theory and statistical physics. Much more information about these two parts can be found for example in the book by D.

Bressoud [11]. Then we present the approach to the bijection problem. Finally we give some results on the joint enumeration of inversions and coinversions in Gog triangles.

I thank both referees for their useful remarks and comments which lead to improvements in the presentation of this paper, in particular for correcting the statement of Proposition 2.

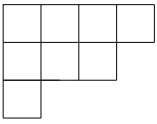
2 Plane Partitions

2.1 Partitions, Tableaux and Triangles

A partition of n is a nonincreasing sequence of nonnegative integers with sum n

$$n = \lambda_1 + \lambda_2 + \ldots, \qquad \lambda_1 \geq \lambda_2 \geq \ldots, \qquad \lambda_i \geq 0.$$

This is a fundamental notion in mathematics which occurs in algebra, representation theory, combinatorics, number theory etc. See e.g. Andrews [1], Fulton [15], Macdonald [21], Ramanujan [23]. The usual way to depict a partition is by drawing superposed rows of squares with λ_i squares in row i from above:



$$8 = 4 + 3 + 1$$

It is easy to derive the following generating series for the set of all partitions, where $|\lambda| = \sum_i \lambda_i$

$$\sum_{\lambda} q^{|\lambda|} = \prod_{n=1}^{\infty} \frac{1}{(1-q^n)}$$

which is closely related to Dedekind's eta function.

A *semi-standard tableau* is obtained by putting positive integers in the boxes of a partition which are

(i) weakly increasing from left to right
(ii) strictly increasing from top to bottom

as below.

1	1	1	3
2	2	4	
4			

The shape of the tableau is the underlying partition λ.

The semi-standard tableaux themselves can be encoded by Gelfand-Tsetlin triangles.

Definition 1 A Gelfand-Tsetlin triangle of size n is a triangular array $X = (X_{i,j})_{n \geqslant i \geqslant j \geqslant 1}$ of nonnegative integers

$$\begin{array}{ccccccccccc} X_{n,1} & & X_{n,2} & & \dots & & X_{n,n-1} & & X_{n,n} \\ & X_{n-1,1} & & X_{n-1,2} & & \dots & & X_{n-1,n-1} & \\ & & \dots & & \dots & & \dots & & \\ & & & X_{2,1} & & X_{2,2} & & & \\ & & & & X_{1,1} & & & & \end{array}$$

such that

$$X_{i+1,j} \leqslant X_{i,j} \leqslant X_{i+1,j+1} \quad \text{for } n-1 \geqslant i \geqslant j \geqslant 1.$$

Given a semi-standard tableau filled with numbers from 1 to n, one can construct a Gelfand-Tsetlin triangle of size n whose row k, as counted from below, consists of the partition, read backwards, formed by the boxes containing numbers from 1 to k in the semi-standard tableau. In the case of the semi-standard tableau above this gives

$$\begin{array}{ccccccc} 0 & & 1 & & 3 & & 4 \\ & 0 & & 2 & & 4 & \\ & & 2 & & 3 & & \\ & & & 3 & & & \end{array}$$

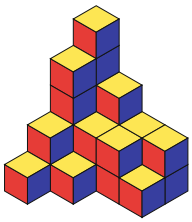
Let $(x_i)_{i \geq 1}$ be a family of indeterminates. For a semi-standard tableau t, let t_i be the number of i occurring in the tableau and $x^t = \prod_i x_i^{t_i}$. The generating function of semi-standard tableaux with shape λ, filled with numbers from 1 to n, is a Schur function

$$s_\lambda(x_1, \dots, x_n) = \sum_{t \text{ tableau of shape } \lambda} x^t.$$

These are symmetric functions, which occur as characters of irreducible representations of the group GL_n (see e.g. Macdonald [21]).

2.2 Plane Partitions

A plane partition is a stack of cubes in a corner. Putting the stack on a square basis and collecting the heights of the piles of cubes, one gets an array of integers:



$$\begin{array}{ccccccccc} & & & & 5 & & & & \\ & & & 4 & & 3 & & & \\ & & 2 & & 2 & & 2 & & \\ & 1 & & 1 & & 2 & & 2 & \\ & & 0 & & 0 & & 1 & & \\ & & & 0 & & 0 & & & \\ & & & & 0 & & & & \end{array}$$

Splitting the array into its left and right parts yields two Gelfand-Tsetlin triangles sharing the same upper row (which is the vertical diagonal of the square array):

$$\begin{array}{ccccc} 0 & 0 & 2 & 5 & \\ & 0 & 1 & 4 & \\ & & 0 & 2 & \\ & & & 1 & \end{array} \qquad \begin{array}{ccccc} 0 & 0 & 2 & 5 & \\ & 0 & 2 & 3 & \\ & & 1 & 2 & \\ & & & 2 & \end{array}$$

From this one can infer that the generating series of plane partitions π according to their size (i.e. the number of cubes in the stack) is equal to

$$\sum_{\pi} q^{|\pi|} = \sum_{\lambda} s_{\lambda}(q, q^2, \ldots, q^j, \ldots)^2.$$

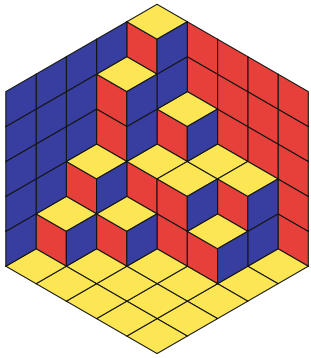
Using Cauchy's formula

$$\sum_{\lambda} s_{\lambda}(x_1, x_2, \ldots, x_j, \ldots) s_{\lambda}(y_1, y_2, \ldots, y_j, \ldots) = \prod_{i,j} \frac{1}{1 - x_i y_j} \tag{2}$$

one obtains Mac Mahon's formula which gives the generating function for plane partitions according to their size

$$\sum_{\pi} q^{|\pi|} = \prod_{n=1}^{\infty} \frac{1}{(1 - q^n)^n}.$$

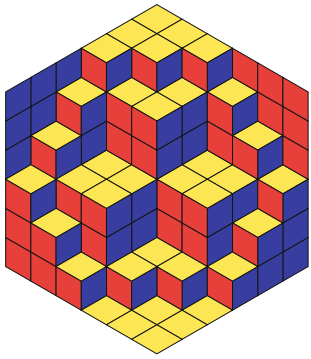
Choosing a large cube that contains a plane partition, one can also encode it as a lozenge tiling of an hexagon.



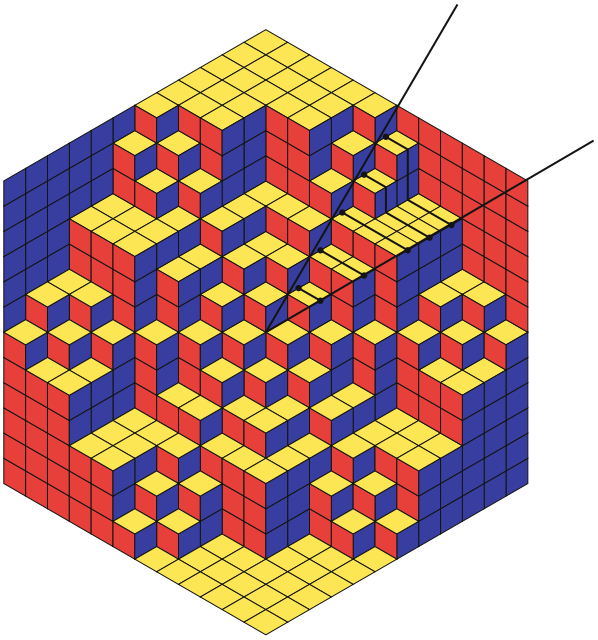
The symmetry group of the hexagon is a dihedral group. For each subgroup of this group, one can consider the class of plane partitions which are invariant under these symmetries of the hexagon. Various enumeration formulas have been derived for such symmetry classes. We will be interested in one of them.

2.3 Totally Symmetric Self-Complementary Plane Partitions

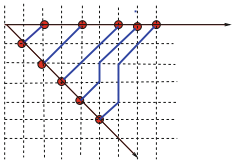
A Totally Symmetric Self-Complementary Plane Partition (TSSCPP in short), of size n, is a plane partition, inside a cube of side $2n$, such that the lozenge tiling has all the dihedral symmetries of the hexagon, as in the picture below, where $n = 3$. Remarkably, a plane partition with all these symmetries can be superposed with its complement in the cube [22].



TSSCPPs can be encoded by non-intersecting paths as below, where the paths go through the yellow and blue tiles of a fundamental domain for the action of the dihedral group:



The paths can be drawn on a lattice using vertical steps to record blue tiles and diagonal steps for yellow tiles:



This correspondence has allowed J. Stembridge [26], building upon the Lindström-Gessel-Viennot technique, to express the number of these paths as a Pfaffian

$$t_n = pf(a_{ij})_{\epsilon_n \le i < j \le n-1}$$

$$a_{ij} = \sum_{r=2i-j+1}^{2j-i} \binom{i+j}{r}, \qquad \epsilon_n = 0 \text{ if } n \text{ is even, 1 if it is odd}$$

It is not an easy task however to evaluate this Pfaffian explicitly, but this was done by G. Andrews [2], who proved that

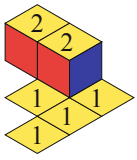
$$t_n = \prod_{j=0}^{n-1} \frac{(3j+1)!}{(n+j)!} \tag{3}$$

$$t_n = 1 \quad 2 \quad 7 \quad 42 \quad 429 \quad 7436 \quad 218348 \quad \cdots$$

2.4 Magog Triangles

Definition 2 A Magog triangle of size n is a Gelfand-Tsetlin triangle of positive integers such that $X_{jj} \leq j$ for all $1 \leq j \leq n$.

Reading the heights of the cubes of a TSSCPP in a fundamental domain of the dihedral group gives a Magog triangle e.g. for the example above, with the heights starting at 1



1 1 2
1 2
1

This gives a bijection between Magog triangles of size n and TSSCPPs of size n. Thus the rather complicated objects which are TSSCPPs can be encoded by these triangles, satisfying a very simple condition.

3 Alternating Sign Matrices

3.1 Jacobi-Desnanot Identity and Dodgson Algorithm

There are many polynomial identities relating the different minors of a matrix. One of them is the Jacobi-Desnanot identity which we now explain. For a square $n \times n$ matrix M let $M^{i_1 \dots i_r}_{j_1 \dots j_r}$ be the matrix obtained by deleting rows $i_1, \dots, i_r$ and columns $j_1, \dots, j_r$. Then one has

$$\det(M) \det(M^{1n}_{1n}) = \det(M^1_1) \det(M^n_n) - \det(M^1_n) \det(M^n_1).$$

For a 2×2 matrix (the empty determinant is 1) this is just

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

Using this identity Charles Dodgson (better known under the name of Lewis Carroll) devised an algorithm for computing the determinant of a matrix which uses only the computation of 2×2 determinants. For example, if you want to compute the determinant of the matrix

$$A = \begin{pmatrix} 1 & 4 & 6 & 0 \\ 2 & 1 & -3 & 1 \\ 3 & 2 & 1 & 5 \\ 3 & 2 & 2 & 0 \end{pmatrix}$$

start with two matrices, A and another matrix B, of size $(n-1) \times (n-1)$, with all its entries equal to one, then inside A insert a (red) matrix formed with the two by two minors of A divided by the corresponding entries of B; inside B insert the (blue) values of A in the inner columns and rows

$$\begin{pmatrix} 1 & 4 & 6 & 0 \\ -7 & -18 & 6 & \\ 2 & 1 & -3 & 1 \\ 1 & 7 & -16 & \\ 3 & 2 & 1 & 5 \\ 0 & 2 & -10 & \\ 3 & 2 & 2 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & -3 & \\ 1 & 1 & 1 \\ 2 & 1 & \\ 1 & 1 & 1 \end{pmatrix}$$

then iterate with the new pair of matrices

$$A' = \begin{pmatrix} -7 & -18 & 6 \\ 1 & 7 & -16 \\ 0 & 2 & -10 \end{pmatrix} \qquad B' = \begin{pmatrix} 1 & -3 \\ 2 & 1 \end{pmatrix}$$

to get

$$\begin{pmatrix} -7 & -18 & 6 \\ -31 & -82 & \\ 1 & 7 & -16 \\ 1 & -38 & \\ 0 & 2 & -10 \end{pmatrix} \quad \begin{pmatrix} 1 & -3 \\ 7 & \\ 2 & 1 \end{pmatrix}$$

finally

$$\det(A) = \frac{\begin{vmatrix} -31 & -82 \\ 1 & -38 \end{vmatrix}}{7} = 180.$$

3.2 The λ-Determinant

In 1983 David Robbins had the idea of replacing, in the above algorithm, every occurrence of $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$ by $\begin{vmatrix} a & b \\ c & d \end{vmatrix}_\lambda = ad + \lambda bc$, for some indeterminate λ. This defines the λ-determinant. The result is surprising, indeed although the algorithm implies taking a lot of quotients of rational fractions the result is always a Laurent polynomial in the coefficients of the matrix. Namely one has, for a $d \times d$ matrix A,

Theorem 1 (D. Robbins, H. Rumsey [24])

$$det_\lambda(A) = \sum_{M \in ASM(d)} (1+\lambda)^{s(M)} \lambda^{i(M)} \prod_{ij} A_{ij}^{M_{ij}} \tag{4}$$

The sum is over the set of alternating sign matrices, defined at the beginning of the introduction while $i(M)$ is the number of inversions of M (to be defined later) and $s(M)$ is the number of -1 coefficients.

This is an example of the "Laurent phenomenon" which is at the heart of the deep theory of cluster algebras, see e.g. [14].

3.3 Alternating Sign Matrices

For the convenience of the reader we remind the definition of alternating sign matrices.

Definition 3 An alternating sign matrix is a square matrix having coefficients in $\{-1, 0, 1\}$ so that, in each row and in each column, if one forgets the zeros, the 1 and -1 entries alternate and the sum is 1.

Here is an example where we show an alternating sign matrix and the alternance of $+1$ and -1 in each row and column, once the zeros are removed:

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 \\ 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} & & & 1 & \\ & & 1 & -1 & 1 \\ 1 & & -1 & 1 & \\ & 1 & & & \\ & & 1 & & \end{pmatrix}$$

In particular, the alternating sign matrices without -1 are exactly the permutation matrices and (4) for $\lambda = -1$ gives the classical formula for the usual determinant.

It turns out that alternating sign matrices occur in a number of different contexts, in statistical physics, representation theory, or combinatorics. We shall give a few examples now.

3.3.1 The Six-Vertex Model

An entry of an alternating sign matrix can take at most three values $\{-1, 0, 1\}$. For each entry with value zero consider the sum of entries lying respectively, on the right and on the left of this entry, then one of these sums is equal to 0 and the other is equal to 1. A similar property holds for the sum of entries lying above and the sum of entries lying below. It follows that one can divide the entries of the matrix into six groups, two corresponding to entries with the value 1 and -1 and four groups corresponding to the configurations of an entry with value 0. There are thus six possible configurations of each entry of an alternating sign matrix, listed below

$$\begin{array}{cc cccc} & & 0 & 1 & 0 & 1 \\ \mathbf{1} & \mathbf{-1} & 1\,\mathbf{0}\,0 & 0\,\mathbf{0}\,1 & 0\,\mathbf{0}\,1 & 1\,\mathbf{0}\,0 \\ & & 1 & 0 & 1 & 0 \end{array} \tag{5}$$

The configurations so obtained form an instance of a famous statistical physic model, known as the six-vertex model, which is one of the most studied models in statistical mechanics (see e.g. [4]). In order to study this model it is convenient to weigh the configurations as follows. Introduce indeterminates q, x_i and y_j, the indices ranging from 1 to n and corresponding to the rows and columns of the matrix. Endow each entry of an alternating sign matrix with a weight $w(i, j)$ (where i, j are the row and column of the entry) given by the following value, according to the configuration of the entry, as in (5)

$$x_i/y_j \qquad y_j/x_i \qquad [qx_i/y_j] \qquad [qx_i/y_j] \qquad [x_i/y_j] \qquad [x_i/y_j] \tag{6}$$

The convention used here is that $[a] = \frac{a-a^{-1}}{q-q^{-1}}$. One can then put on every alternating sign matrix the product of the weights of its entries. It turns out that the particular form of the weights (6) allows one to use the Yang-Baxter equation to compute the partition function of this model, i.e. the sum over all alternating sign matrices of the weights, under the form of a remarkable determinant, the Izergin-Korepin determinant [16]

$$\sum_{ASM} \prod_{ij} w(i,j) = \frac{\prod_i (x_i/y_i) \prod_{i,j} [x_i/y_j][qx_i/y_j]}{\prod_{i,j}[x_i/x_j][y_i/y_j]} \det\left[\frac{1}{[x_i/y_j][qx_i/y_j]}\right]$$

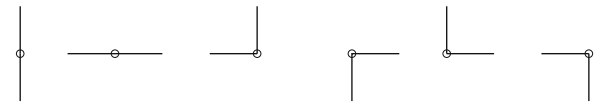
Using appropriate specializations of the variables x_i, y_j and the parameter q, G. Kuperberg [17] was able to deduce from this that the number of alternating sign matrices of size n is again, as in (3)

$$A_n = \prod_{j=0}^{n-1} \frac{(3j+1)!}{(n+j)!} \tag{7}$$

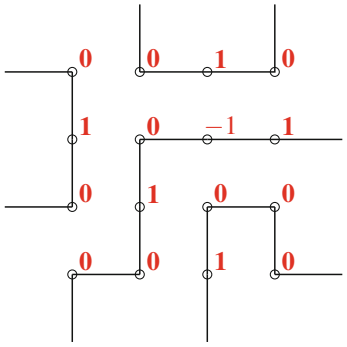
This result had been obtained earlier by D. Zeilberger [28] in an indirect way, by showing that the alternating sign matrices of sign n are equinumerous with TSSCPPs of the same size and using (3). Recently a new proof of this result and related enumerations has been given by I. Fischer [13].

3.3.2 Fully Packed Loops

Another way to encode the six-vertex model is to replace each of the possible six configurations by one of the following



After fixing boundary conditions, there is a unique way to complete the diagram in a fully packed loop as in the picture below, which corresponds to the matrix (1).



Observe that in each configuration the $4n$ vertices on the boundary are related by noncrossing paths. This observation has lead to the famous Razumov-Stroganov conjecture [25], relating alternating sign matrices and the $O(n)$ model, which has been solved recently [12].

3.3.3 Alternating Sign Matrices, the Bruhat Order and Gog Triangles

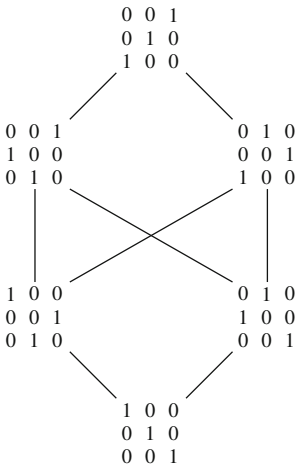
Recall that almost all invertible matrices can be factorized as $X = LU$ into a product of a lower and an upper triangular matrix (the LU-factorization). This can be refined into the Bruhat decomposition, expressing the general linear group as a disjoint union of cells indexed by the symmetric group

$$GL_d = \cup_{w \in S_d} BwB$$

where B is the Borel subgroup of upper triangular matrices. For example, the matrices X having LU factorization are those such that $w_0 X \in Bw_0B$ where w_0 is the permutation of $[1, d]$ such that $w_0(i) = d + 1 - i$. They form the cell of largest dimension. This decomposition induces an order relation (the Bruhat order) on the symmetric group by declaring for $\sigma, \tau \in S_n$:

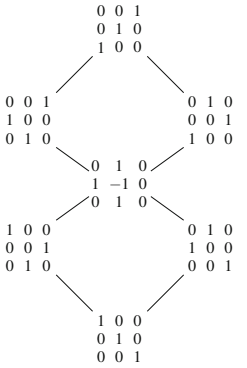
$$\sigma \leq \tau \quad \text{iff} \quad B\sigma B \subset \overline{B\tau B}$$

e.g. for S_3 we get the order relation with Hasse diagram



The Bruhat order on S_3

As we are going to explain, alternating sign matrices can be used to complete this order into a lattice order as in the following diagram



The lattice of 3×3 alternating sign matrices

For this we need to introduce a new species of Gelfand-Tsetlin triangles.

Definition 4 A Gog triangle of size n is a Gelfand-Tsetlin triangle such that

(i) $$X_{i,j} < X_{i,j+1}, \qquad j < i \leqslant n-1$$

in other words, such that its rows are strictly increasing, and such that

(ii) $$X_{n,j} = j, \qquad 1 \leqslant j \leqslant n.$$

There is a simple bijection between the sets of Gog triangles and of Alternating sign matrices of the same size, which goes as follows: If $(M_{ij})_{1\leqslant i,j\leqslant n}$ is an ASM of size n, then the matrix $\tilde{M}_{ij} = \sum_{k=i}^{n} M_{kj}$ has exactly $i-1$ entries 0 and $n-i+1$ entries 1 on row i. Let $(X_{ij})_{j=1,\ldots,i}$ be the columns (in increasing order) with a 1 entry of $\tilde{M}$ on row $n-i+1$. The triangle $X = (X_{ij})_{n\geqslant i\geqslant j\geqslant 1}$ is the Gog triangle corresponding to M.

For example, below are an alternating sign matrix of size 5 and its associated Gog triangle

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \qquad \begin{array}{ccccccccc} 1 & & 2 & & 3 & & 4 & & 5 \\ & 1 & & 3 & & 4 & & 5 & \\ & & 1 & & 4 & & 5 & & \\ & & & 2 & & 4 & & & \\ & & & & 3 & & & & \end{array}$$

There is an order relation on Gog triangles obtained by entrywise comparison: for triangles X, Y of the same size, $X \leq Y$ if and only if each entry of X is smaller than the corresponding entry of Y. Clearly the Gog triangles of fixed size form a lattice for this order. It turns out that the restriction of this order relation to Gog triangles corresponding to permutations is exactly the reversed Bruhat order. The set of alternating sign matrices thus appears as the lattice completion of the set of permutations endowed with the Bruhat order, as first proved by Lascoux and Schützenberger [19].

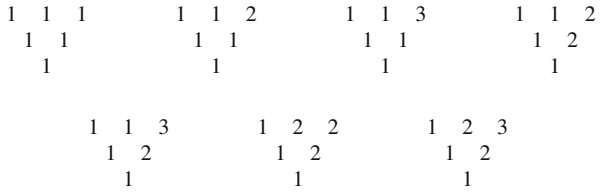
4 The Gog-Magog Problem

4.1 The Question

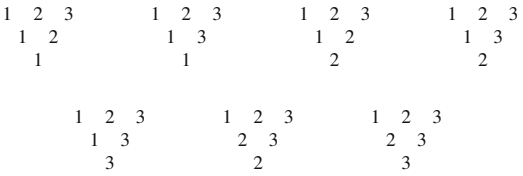
Since the sets of Gog and Magog triangles of size n have the same number of elements it is sensible to ask, in view of their very similar definitions, if there exists a natural bijection between these two sets. This problem is at the time of this writing still largely open. Observe that, although in our discussion we have encountered rather sophisticated mathematical objects, the actual definitions of the Gog and Magog triangles are completely elementary. One needs only to know what are the positive integers and how to compare two positive integers, it is not even necessary to know how to add or multiply them! Also many results have been obtained on the refined enumeration of Gog and Magog triangles according to different statistics and it has been observed that some of these refined enumerations coincide cf. [5]. All these facts point towards the existence of a mathematical structure which would explain all these coincidences by showing that Gog and Magog triangles give two

different ways of parametrizing the same mathematical objects, however for the moment the nature of this mathematical structure remains elusive.

Here are the seven Gog and Magog triangles of size 3. Already finding a "natural" bijection between them does not seem so obvious.



Magog triangles of size 3



Gog triangles of size 3

4.2 Gog and Magog Trapezoids

Definition 5 An (n, k) right (resp. left) Gog trapezoid (for $k \leq n$) is an array of positive integers formed from the k rightmost SW-NE diagonals (resp. leftmost NW-SE diagonals) of some Gog triangle of size n.

Below are two (5, 2) Gog trapezoids.

left trapezoid

right trapezoid

Analogously there is a notion of right and left Magog trapezoids. We will use only the right ones, of which below is a $(5, 2)$ example

$$\begin{array}{ccccccccc} & & & & 2 & & 3 & & \\ & & & 1 & & 3 & & & \\ & & 1 & & 2 & & & & \\ & 1 & & 2 & & & & & \\ & & 1 & & & & & & \end{array}$$

There is no known simple formula for enumerating Gog or Magog trapezoids of a fixed shape, however the following holds

Theorem 2 (Zeilberger [28]) *For all $k \leq n$, the (n, k) right Gog and Magog trapezoids are equinumerous.*

The proof of Zeilberger uses transformations of generating series for these objects and it does not seem possible to transform it into a bijective proof. Some conjectures on the enumeration of Gog and Magog trapezoids refined by some further statistics have been formulated by Krattenthaler [18]. A bijection between permutation matrices and a subset of Magog triangles has been proposed by J. Striker [27]. In the case of $(n, 2)$ right trapezoids a bijective proof incorporating a further statistic has been obtained in [9]. This proof is based on the Schützenberger involution, to be defined below, and uses the inversions of a Gog triangle. Bettinelli [8] found another, simpler bijection which however does not seem to preserve any of the statistics considered by Krattenthaler.

4.3 An Approach to the Bijection Problem

In this section we will describe an approach to the bijection problem which has led to some recent progress. For this approach we need to introduce some statistics on Gog and Magog triangles.

For a Gog triangle X we define

$$\beta_{Gog}(X) = X_{1,1}$$

For a Magog triangle of size n we let

$$\beta_{Magog}(X) = \sum_{i=1}^{n} X_{n,i} - \sum_{i=1}^{n-1} X_{n-1,i}$$

Remark that if we identify a Gog triangle with an alternating sign matrix, then the statistics $\beta_{Gog}(X)$ corresponds to the position of the 1 in the bottom line. Some recent results on the joint enumeration of this and other similar statistics can be found in [3, 5]. In particular, it is known that the number of Gog triangles of size n with $\beta_{Gog}(X) = k$ is equal to the number of Magog triangles of size n with $\beta_{Magog}(X) = k$.

Consider now the bottom triangle

$$\begin{array}{ccc} a & & b \\ & c & \end{array}$$

made of the two lowest rows of some Gog triangle of size $n \geq 2$. Thus a, b, c are integers satisfying the inequalities

$$1 \leq a \leq c \leq b \leq n; \quad a < b.$$

Consider now a triangle

$$\begin{array}{ccc} a' & & b' \\ & c' & \end{array}$$

extracted from the two rightmost NW-SE diagonals of Magog triangle of size n, such as this one:

$$\begin{array}{ccccccc} 1 & & 1 & & 1 & a' & b' \\ & 1 & & 1 & & 1 & c' \\ & & 1 & & 1 & & 1 \\ & & & 1 & & 1 & \\ & & & & 1 & & \end{array}$$

These triangles are characterized by the inequalities

$$a' \leq c' \leq b' \leq n; \quad c' \leq n-1.$$

It is now easy to find a bijection between these two sets of triangles, mapping the statistics β_{Gog} to β_{Magog} i.e. c to $a' + b' - c'$, as follows:

start from a triangle extracted from a Gog triangle $\begin{array}{ccc} a & & b \\ & c & \end{array}$ with $1 \leq a \leq c \leq b \leq n; \quad a < b$ and then

- if $a < c$ map $\begin{array}{ccc} a & & b \\ & c & \end{array}$ to $\begin{array}{ccc} a & & b \\ & a+b-c & \end{array}$
- if $a = c$ map $\begin{array}{ccc} a & & b \\ & a & \end{array}$ to $\begin{array}{ccc} a & & b-1 \\ & b-1 & \end{array}$.

We leave to the reader the task of verifying that this is a bijection. Observe that it can be obtained in two steps: first we make the transformation

$$\begin{array}{ccc} a & b & \\ & c & \end{array} \rightarrow \begin{array}{ccc} a & b & \\ & c & \end{array} \qquad \text{if } a < c$$

$$\begin{array}{ccc} a & b & \\ & a & \end{array} \rightarrow \begin{array}{ccc} a & b-1 & \\ & a & \end{array} \qquad \text{if } a = c$$

then a symmetry

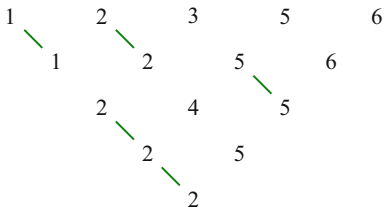
$$\begin{array}{ccc} a & b & \\ & c & \end{array} \rightarrow \begin{array}{ccc} a & & b \\ & a+b-c & \end{array}$$

This idea was generalized in [9]. The first step leads to considering inversions of Gog triangles while the second leads to the Schützenberger involution. We shall explain these two ideas now.

4.3.1 Inversions

An inversion in a Gog triangle X is a pair (i, j) such that $X_{i,j} = X_{i+1,j}$.

For example the triangle below has 5 inversions.



4.3.2 Schützenberger Involution

The Schützenberger involution is a fundamental tool in the theory of Young tableaux, which has a nice geometric interpretation [20]. Its simplest description uses the RSK transformation, which is a bijection between the set of two-dimensional arrays of nonnegative integers, $(M_{ij})_{i,j\geq 1}$ and pairs (S, T) of semi-standard Young tableaux of the same shape λ. This bijection allows to give a bijective proof of Cauchy's identity (2). Using the bijection between semi-standard tableaux and Gelfand-Tsetlin triangles the Schützenberger involution can be transported to Gelfand-Tsetlin triangles. The following description of this involution has been given by Berenstein and Kirillov [7].

First define involutions $\mathfrak{s}_k$, for $k \leqslant n-1$, acting on the set of Gelfand-Tsetlin triangles of size n. If $X = (x_{i,j})_{n \geqslant i \geqslant j \geqslant 1}$ is such a triangle the action of $\mathfrak{s}_k$ on X is given by $s_k X = (\tilde{X}_{i,j})_{n \geqslant i \geqslant j \geqslant 1}$ with

$$\begin{aligned} \tilde{X}_{i,j} &= X_{i,j}, \qquad \text{if } i \neq k \\ \tilde{X}_{k,j} &= \max(X_{k+1,j}, X_{k-1,j-1}) + \min(X_{k+1,j+1}, X_{k-1,j}) - X_{i,j} \end{aligned}$$

It is understood that $\max(a,b) = \max(b,a) = a$ and $\min(a,b) = \min(b,a) = a$ if the entry b of the triangle is not defined. The geometric meaning of the transformation of an entry is the following: on row k, any entry $X_{k,j}$ is surrounded by four (or less if it is on the boundary) numbers, increasing from left to right.

$$\begin{array}{ccc} X_{k+1,j} & & X_{k+1,j+1} \\ & X_{k,j} & \\ X_{k-1,j-1} & & X_{k-1,j} \end{array}$$

These four numbers determine a smallest interval containing $X_{k,j}$, namely

$$\left[\max(X_{k+1,j}, X_{k-1,j-1}), \min(X_{k+1,j+1}, X_{k-1,j})\right]$$

and the transformation maps $X_{k,j}$ to its mirror image with respect to the center of this interval.

Define $\omega_j = \mathfrak{s}_j \mathfrak{s}_{j-1} \ldots \mathfrak{s}_2 \mathfrak{s}_1$.

Theorem 3 (Berenstein and Kirillov [7]) *The Schützenberger involution, acting on Gelfand-Tsetlin triangles of size n, is given by the formula*

$$S = \omega_1 \omega_2 \ldots \omega_{n-1}$$

Using inversions and the Schützenberger involution a bijection between $(n, 2)$ Gog and Magog trapezoids was given in [9].

4.3.3 GOGAm Triangles and Trapezoids

Definition 6 A GOGAm triangle of size n is a Gelfand-Tsetlin triangle which is the image by the Schützenberger involution of a Magog triangle (of size n).

Remark 1 The name GOGAm is obtained from Magog by reading backwards and changing the case as a reminder of the description of the Schützenberger involution on words (cf. [15]).

It is shown in [9] that the GOGAm triangles of size n are the Gelfand-Tsetlin triangles $X = (X_{i,j})_{n \geqslant i \geqslant j \geqslant 1}$ such that $X_{nn} \leq n$ and, for all $1 \leq k \leq n-1$,

and all $n = j_0 > j_1 > j_2 \ldots > j_{n-k} \geq 1$, one has

$$\left(\sum_{i=0}^{n-k-1} X_{j_i+i, j_i} - X_{j_{i+1}+i, j_{i+1}} \right) + X_{j_{n-k}+n-k, j_{n-k}} \leq k \tag{8}$$

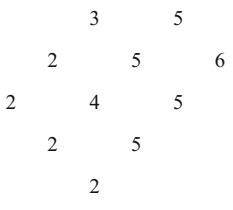
The problem of finding an explicit bijection between Gog and Magog triangles can therefore be reduced to that of finding an explicit bijection between Gog and GOGAm triangles. Again one can define right or left GOGAm trapezoids.

Conjecture 1 For all $k \leq n$, the (n, k) left Gog and GOGAm trapezoids are equinumerous.

In [10] it was shown that the ideas of [9] could be used to provide a simple bijection between $(n, 2)$ Gog and GOGAm left trapezoids. These bijections suggested some further conjectures which we describe in the next section.

4.3.4 Pentagons

Definition 7 For (n, k, l, m), with $n \geq k, l, m$, an (n, k, l, m) Gog (resp. GOGAm) pentagon is an array of positive integers $X = (x_{i,j})_{n \geqslant i \geqslant j \geqslant 1; k \geqslant j; j+l \geqslant i+1}$ formed from the intersection of the k leftmost NW-SE diagonals, the l rightmost SW-NE diagonals and the m bottom lines of a Gog (resp. GOGAm) triangle of size n.



A $(6, 4, 3, 5)$ Gog pentagon

Remark that if $m \geq k + l - 1$ then the pentagon is a rectangle, whereas if $m \leq k, l$ then it is a Gelfand-Tsetlin triangle of size m.

Conjecture 2 For any n, k, l, m the (n, k, l, m) Gog and GOGAm pentagons are equinumerous.

This conjecture can even be refined into

Conjecture 3 For each n, k, l the (n, k) left Gog and GOGAm trapezoids with bottom entry $X_{1,1} = l$ are equinumerous.

Some numerical evidence for these conjectures has been given in [10].

5 On the Distribution of Inversions and Coinversions

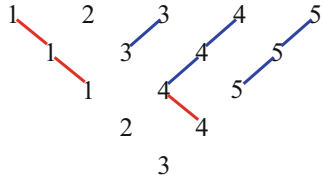
5.1 Inversions and Coinversions

We recall the definition of inversions and introduce the dual notion of coinversion.

Definition 8 An inversion in a Gog triangle X is a pair (i, j) such that $X_{i,j} = X_{i+1,j}$.

A coinversion is a pair (i, j) such that $X_{i,j} = X_{i+1,j+1}$.

For example, the Gog triangle in (9) contains three inversions, (2, 2), (3, 1), (4, 1) and five coinversions, (3, 2), (3, 3), (4, 2), (4, 3), (4, 4).



(9)

We denote by $\mu(X)$ (resp. $\nu(X)$) the number of inversions (resp. coinversions) of a Gog triangle X. Since a pair (i, j) cannot be an inversion and a coinversion at the same time in a Gog triangle and the top row does not contain any inversion or coinversion, one has

$$\nu(X) + \mu(X) \leq \frac{n(n-1)}{2}$$

Actually one can easily see that $\frac{n(n-1)}{2} - \nu(X) - \mu(X)$ is the number of -1's in the alternating sign matrix associated to the Gog triangle X. Also inversions and coinversions correspond to different types of vertices in the six vertex model, see e.g. [6].

Let us denote by $Z(n, x, y)$ the generating function of Gog triangles of size n according to ν and μ.

$$Z(n, x, y) = \sum_{X \in Gog_n} x^{\nu(X)} y^{\mu(X)}. \tag{10}$$

where the sum is over the set Gog_n of Gog triangles of size n.

The following formula has been proved in [6], using properties of the six vertex model.

Proposition 1

$$Z(n, x, y) = \det_{0 \leq i,j \leq n-1} \left(-y^i \delta_{i,j+1} + \sum_{k=0}^{\min(i,j+1)} \binom{i-1}{i-k} \binom{j+1}{k} x^k \right). \tag{11}$$

For example, for Gog triangles of size 3, we have

$$Z(3,x,y)=\det\begin{pmatrix}1 & 1 & 1\\ -y+x & 2x & 3x\\ x & -y^2+2x+x^2 & 3x+3x^2\end{pmatrix}=x^3+xy+y^3+2x^2y+2xy^2. \tag{12}$$

which matches part (a) of Table 1.

It is however not so easy to use this formula in order to prove results on the distribution of inversion and coinversions.

5.2 Distribution of Inversions and Coinversions

Table 1 below shows the joint distribution of μ and ν, for $n=3$ and $n=4$.

We remark that the numbers on the antidiagonal are the Mahonian numbers counting permutations according to the number of their inversions.

Let us denote by $A_{n,k}$ the set of pairs of nonnegative integers (i,j) such that

$$i\geq\frac{k(k+1)}{2},\qquad j\geq\frac{(n-k-1)(n-k)}{2},\qquad i+j\leq\frac{n(n-1)}{2}$$

and let

$$A_n=\cup_{k=0}^{n-1}A_{n,k}.$$

We will give a simple combinatorial proof of the following.

Theorem 4 *There exists a Gog triangle of size n, with i inversions and j coinversions, if and only if (i,j) belongs to the set A_n. If $i=\frac{k(k+1)}{2}$ and $j=\frac{(n-k-1)(n-k)}{2}$ for some $k\in[0,n-1]$ then this triangle is unique, furthermore its bottom value is $n-k$.*

Remark 2 Note, for future reference, that if (l,m) belongs to the set A_n and if $l<\frac{p(p+1)}{2}$ then $m\geq\frac{(n-p)(n-p+1)}{2}$.

Table 1 The number of Gog triangles of size 3 (a) and 4 (b) with k inversions (horizontal values) and l coinversions (vertical values)

	0	1	2	3
0				1
1		1	2	
2		2		
3	1			

	0	1	2	3	4	5	6
0							1
1			1	2	3		
2			6	5			
3		1	6	6			
4		2	5				
5		3					
6	1						

5.3 *Proof of Theorem 4*

5.3.1 Existence

First we show that there exists a triangle of size n with $\frac{k(k+1)}{2}$ inversions and $\frac{(n-k-1)(n-k)}{2}$ coinversions. Indeed the triangle is defined by

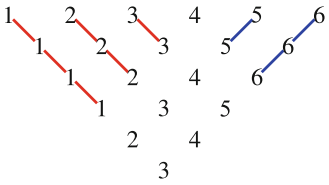
$$X_{ij} = j \quad \text{for} \quad j \leq i - n + k \tag{13}$$

$$X_{ij} = n + j - i \quad \text{for} \quad j \geq k + 1 \tag{14}$$

$$X_{ij} = n - k + 2j - i - 1 \quad \text{for} \quad i - n + k + 1 \leq j \leq k \tag{15}$$

The bottom entry of this triangle is $n - k$, as expected.

We give an example below: for $n = 6$ and $k = 3$, the triangle has 6 inversions and 3 coinversions:

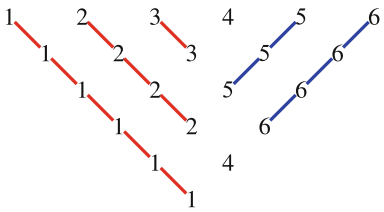


Observe that the entries which are neither inversions nor coinversions form a rectangle of size $k \times (n - k - 1)$ at the bottom of the triangle.

The ASM corresponding to such a triangle has a diamond shape:

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Starting from this triangle, it is not difficult, for a pair of integers (l, m) such that $l \geq \frac{k(k+1)}{2}$, $m \geq \frac{(n-k-1)(n-k)}{2}$, and $l + m \leq \frac{n(n-1)}{2}$, to construct (at least) one triangle with l inversions and m coinversions, for example one can add inversions by decreasing some entries, starting from the westmost corner of the rectangle, and add coinversions by increasing entries, starting from the eastmost corner. Here is an example with $n = 6, l = 9, m = 5$, details of the general case are left to the reader.



5.3.2 Standardization of Gog Triangles

In order to prove the only if part of the Theorem, as well as the uniqueness statement, we now introduce two *standardization* operations. These operations build a Gog triangle of size $n-1$ from a Gog triangle of size n.

5.3.3 Left Standardization

Let X be a Gog triangle of size n then its $(n-1)$th row (counted from bottom to top) has the form $1, 2, \ldots, k, k+2, \ldots, n$ for some $k \in [1, n]$. For $j \leq k$, let m_j be the smallest integer such that $X_{n,j} = X_{m_j,j} = j$.

The left standardization of X is the triangle LX of size $n-1$ obtained as follows:

$$LX_{i,j} = X_{i,j} = j \quad \text{for} \quad j \leq k \quad \text{and} \quad n-1 \geq i \geq m_j. \tag{16}$$

$$LX_{i,j} = X_{i,j} - 1 \quad \text{for other values of } i, j. \tag{17}$$

5.3.4 Right Standardization

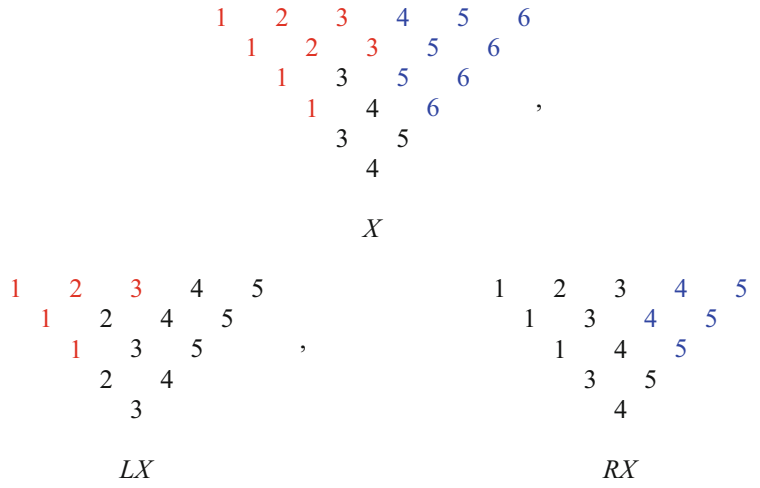
Let X be a Gog triangle of size n with $(n-1)$th row of the form $1, 2, \ldots, k, k+2, \ldots, n$, and for $j \geq k+1$ let $p_j \geq 1$ be the largest integer such that $X_{n,j} = X_{n-p_j,j+1-p_j} = j+1$.

The right standardization of X is the triangle RX of size $n-1$ obtained as follows:

$$RX_{n-l,j+1-l} = X_{n-l,j+1-l} - 1 = j \quad \text{for} \quad j \geq k+1 \quad \text{and} \quad 1 \leq l \leq p_j. \tag{18}$$

$$RX_{i,j} = X_{i,j} \quad \text{for other values of i,j.} \tag{19}$$

Below are a Gog triangle of size 6, together with its left and right standardizations:



X

LX *RX*

5.3.5 An Intermediate Result

Proposition 2 [1] *Let X be a Gog triangle of size n, with l inversions and m coinversions with $(n-1)$th row $1, 2, 3, \ldots, k, k+2, \ldots n$.*

Then LX is a Gog triangle of size $n-1$ with at most $m-n+k+1$ coinversions and at most l inversions.

Similarly, RX is a Gog triangle of size $n-1$ with at most $l-k$ inversions and at most m coinversions.

Proof We prove the proposition only for the left standardization. The case of right standardization can be proven in an analogous way, or by noticing that the mapping $X \mapsto (n+1-X_{i,i+1-j})_{i,j}$ is an involution, which exchanges inversions and coinversions.

Observe that there are exactly k inversions and $n-k-1$ coinversions on row $n-1$ of X. It follows that X has $l-k$ inversions and $m-n+k+1$ coinversions on rows $1, \ldots, n-2$.

We first prove that LX is a Gog triangle. For any i, j we have to prove that

$$LX_{i,j} \geq LX_{i-1,j-1}, \quad LX_{i,j} \geq LX_{i+1,j}, \quad LX_{i,j} > LX_{i,j-1}.$$

[1] In the first version of this paper the statement of this proposition was incorrect. I would like to thank the referees for pointing out the mistake.

Since X is a Gog triangle one has $X_{i,j} \geq X_{i-1,j-1}$, therefore the first inequality may fail only if $LX_{i,j} = X_{i,j} - 1$ and $LX_{i-1,j-1} = X_{i-1,j-1}$. If this is the case then $X_{i,j} > j$ and $X_{i-1,j-1} = j - 1$, therefore $LX_{i,j} > LX_{i-1,j-1}$. This shows also that LX cannot have more coinversions than X on its first $n-2$ rows, therefore the number of coinversions of LX is at most $m-n+k+1$. A similar reasoning yields the other two inequalities, moreover the number of inversions of LX can increase at most by k with respect to that of X in its first $n-2$ rows, more precisely by at most one in each of the k leftmost NW-SE diagonals. It follows that LX has at most l inversions.

5.3.6 End of the Proof

We can now finish the proof of Theorem 4 by induction on n. For $n = 3$ or 4, the claim follows by inspection of Table 1. Let X be a Gog triangle of size n, with l inversions and m coinversions. We have to prove that (l, m) belongs to some $A_{n,r}$. We have seen that LX is a Gog triangle of size $n-1$ with at most $m - n + k + 1$ coinversions and at most l inversions, whereas RX is a Gog triangle of size $n-1$ with at most $l - k$ inversions at most m coinversions. By the induction hypothesis there exists some p such that

$$l \geq \frac{p(p+1)}{2}, \qquad m - n + k + 1 \geq \frac{(n-p-2)(n-p-1)}{2} \tag{20}$$

and there exists q such that

$$l - k \geq \frac{q(q+1)}{2}, \qquad m \geq \frac{(n-q-2)(n-q-1)}{2}. \tag{21}$$

If $p > q$, then (20) implies $l \geq \frac{(q+1)(q+2)}{2}$ and since $m \geq \frac{(n-q-2)(n-q-1)}{2}$ by (21) one has $(l, m) \in A_{n,q+1}$.

Similarly if $q > p$ then (21) implies $l \geq \frac{(p+1)(p+2)}{2}$ and (20) implies $m \geq \frac{(n-p-2)(n-p-1)}{2}$ so that $(l, m) \in A_{n,p+1}$.

If now $p = q$ then either $k > p$ and then $l - k \geq \frac{q(q+1)}{2}$ implies $l \geq \frac{(p+1)(p+2)}{2}$ and $(l, m) \in A_{n,p+1}$, or $k \leq p$ then $m - n + k + 1 \geq \frac{(n-p-2)(n-p-1)}{2}$ implies $m \geq \frac{(n-p-1)(n-p)}{2}$ and $(l, m) \in A_{n,p}$.

Suppose now that $l = \frac{p(p+1)}{2}$, $m = \frac{(n-p-1)(n-p)}{2}$. We wish to prove that there exists a unique Gog triangle with these numbers of inversions and coinversions. Let X be such a triangle and consider RX, which has at most $l - k$ inversions. If $k > p$ then $l - k < \frac{(p-1)p}{2}$, therefore, by Remark 2, RX has at least $\frac{(n-p)(n-p+1)}{2}$ coinversions, which contradicts the fact that RX has at most $m = \frac{(n-p-1)(n-p)}{2}$ coinversions; it follows that $k \leq p$. A similar reasoning with LX shows that in fact $k = p$, and RX has at most $\frac{(p-1)p}{2}$ inversions, and at most $m = \frac{(n-p-1)(n-p)}{2}$ coinversions. By the induction hypothesis RX is the unique Gog triangle with

$l = \frac{(p-1)p}{2}$ inversions, and $\frac{(n-p-1)(n-p)}{2}$ coinversions. The triangle X is completely determined by RX and k and we have $k = p$ therefore X is unique. Comparing with the formula (13), (14), and (15) for this RX, we check that X is the unique triangle of size n with $l = \frac{p(p+1)}{2}$, $m = \frac{(n-p-1)(n-p)}{2}$. □

References

1. Andrews G.E.: The Theory of Partitions. Encyclopedia of Mathematics and its Applications, vol. 2. Addison-Wesley Publishing Co., Reading/London/Amsterdam (1976)
2. Andrews G.E.: Plane partitions. V. The TSSCPP conjecture. J. Combin. Theor. Ser. A **66**(1), 28–39 (1994)
3. Ayyer, A., Romik, D.: New enumeration formulas for alternating sign matrices and square ice partition functions. Adv. Math. **235**, 161–186 (2013)
4. Baxter, R.: Exactly Solved Models in Statistical Mechanics. Academic, London (1982)
5. Behrend R.E.: Multiply-refined enumeration of alternating sign matrices. Adv. Math. **245**, 439–499 (2013)
6. Behrend, R.E., Di Francesco, P., Zinn-Justin, P.: On the weighted enumeration of alternating sign matrices and descending plane partitions. J. Combin. Theor. Ser. A **119**(2), 331–363 (2012)
7. Berenstein, A.D., Kirillov, A.N.: Groups generated by involutions, Gelfand-Tsetlin patterns and combinatorics of Young tableaux. St. Petersburg Math. J. **7**(1), 77–127 (1996)
8. Bettinelli, J.: A simple explicit bijection between (n,2)-Gog and Magog trapezoids. Lotharingien Séminaire Lotharingien de Combinatoire **75**, 1–9 (2016). Article B75e
9. Biane, P., Cheballah, H.: Gog and Magog triangles and the Schützenberger involution. Séminaire Lotharingien de Combinatoire, B66d (2012)
10. Biane, P., Cheballah, H.: Gog and GOGAm pentagons. J. Combin. Theor. Ser. A **138**, 133–154 (2016) JCTA
11. Bressoud D.M.: Proofs and Confirmations, the Story of the Alternating Sign Matrix Conjecture. Cambridge University Press, Cambridge (1999)
12. Cantini, L., Sportiello, A.: Proof of the Razumov-Stroganoff conjecture. J. Combin. Theor. Ser. A **118**(5), 1549–1574 (2011)
13. Fischer I.: A new proof of the refined alternating sign matrix theorem. J. Combin. Theor. Ser. A **114**(2), 253–264 (2007)
14. Fomin, S., Zelevinsky, A.: The Laurent phenomenon. Adv. Appl. Math. **28**, 119–144 (2002)
15. Fulton, W.: Young Tableaux. London Mathematical Society, Student Texts, vol. 35. Cambridge University Press, Cambridge (1997)
16. Izergin, A.G.: Partition function of a six vertex model in a finite volume. Soviet Phys. Dokl. **32**, 878–879 (1987)
17. Kuperberg G.: Another proof of the alternating sign matrix conjecture. Int. Math. Res. Not. **1996**, 139–150 (1996)
18. Krattenthaler, C.: A Gog-Magog conjecture. http://www.mat.univie.ac.at/~kratt/artikel/magog.html
19. Lascoux, A., Schützenberger, M.P.: Treillis et bases des groupes de Coxeter. Electron. J. Combin. **3**(2), 27, 35pp (1996)
20. van Leeuwen, M.A.: Flag varieties and interpretations of Young tableaux algorithms. J. Algebra **224**, 397–426 (2000)
21. Macdonald, I.G.: Symmetric Functions and Hall Polynomials, 2nd edn. With contributions by Zelevinsky, A. Oxford Mathematical Monographs. Oxford Science Publications/The Clarendon Press/Oxford University Press, New York (1995)

22. Mills, W.H., Robbins, D.P., Rumsey, H.: Self complementary totally symmetric plane partitions. J. Combin. Theor. Ser. A **42**, 277–292 (1986)
23. Andrews, G.E., Berndt, B.C.: Ramanujan's Lost Notebook. Part III. Springer, New York (2012)
24. Robbins, D.P., Rumsey, H.: Determinants and alternating sign matrices. Adv. Math. **62**, 169–184 (1986)
25. Razumov, A.V., Stroganov, Y.G.: Combinatorial nature of ground state vector of O(1) loop model. Theor. Math. Phys. **138**, 333–337 (2004)
26. Stembridge, J.: Nonintersecting paths, Pfaffians, and plane partitions. Adv. Math. **83**, 96–131 (1990)
27. Striker, J.: A direct bijection between permutations and a subclass of totally symmetric self-complementary plane partitions. In: 25th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2013), Paris, pp. 803–812. Discrete Mathematics and Theoretical Computer Science Proceedings, AS
28. Zeilberger, D.: Proof of the alternating sign matrix conjecture. Electron. J. Combin. **3**, R13 (1996)

The Clebsch Representation in Optimal Control and Low Rank Integrable Systems

Anthony M. Bloch, François Gay-Balmaz, and Tudor S. Ratiu

Abstract Certain kinematic optimal control problems (the Clebsch problems) and their connection to classical integrable systems are considered. In particular, the rigid body problem and its rank $2k$ counterparts, the geodesic flows on Stiefel manifolds and their connection with the work of Moser, flows on symmetric matrices, and the Toda flows are studied.

1 Introduction

We study a class of kinematic optimal control problems and their relationship with certain integrable systems. In particular, we consider the so-called Clebsch optimal control problem, as analyzed in [8, 21]. We discuss geometrical aspects of the optimal dynamics and their relationship to some classical integrable systems. In particular, we examine the formulation of integrable systems discussed in [32] which includes the free rigid body, their low rank counterparts, flows on Stiefel manifolds, the geodesic spray on the ellipsoid, and the Neumann problem. We also consider the flows on symmetric matrices [4, 12, 13] and the full Toda flows [5, 14, 15] which generalize the classical Toda lattice [18, 19]. We show in this paper

A. M. Bloch (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
e-mail: abloch@umich.edu

F. Gay-Balmaz
CNRS – LMD – IPSL, Ecole Normale Supérieure, Paris, France
e-mail: francois.gay-balmaz@lmd.ens.fr

T. S. Ratiu
School of Mathematics, Shanghai Jiao Tong University, Shanghai, China

Section de Mathématiques, Université de Genéve, Genéve, Switzerland

École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: ratiu@sjtu.edu.cn; tudor.ratiu@epfl.ch

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_5

how the Clebsch approach leads naturally to generalizations of the flows considered by Moser both to higher rank systems and other integrable Hamiltonian systems of interest. We also interpret the Moser formulation geometrically in terms of the momentum map.

More details on the integrability of low rank systems using this approach will appear in [11], extending the work of, e.g., [1] and, in the Stiefel case, that of [17].

In several examples, including the rigid body equations, the Toda lattice, the Bloch-Iserles system, and optimal control on Stiefel manifolds, the Clebsch formulation allows one to naturally formulate the evolution equations on a Cartesian product, rather than on a tangent or cotangent space. This is referred to as the symmetric representation and offers a direct link with the discrete evolution equation obtained by variational discretization, as both the continuous equations in the Clebsch formulation and the discrete equations evolve on the same Cartesian product.

2 The Clebsch Optimal Control Problem

2.1 Review of the Clebsch Optimal Control Problem

We recall from [21] some facts concerning the Clebsch optimal control problem. Let $\Phi : Q \times G \to Q$ be a right action of a Lie group G on a smooth manifold Q. We denote by $qg := \Phi(q, g)$ the action of $g \in G$ on $q \in Q$. Given $u \in \mathfrak{g}$, where $\mathfrak{g}$ is the Lie algebra of G, we denote by $u_Q \in \mathfrak{X}(Q)$ the infinitesimal generator of the action. Recall that u_Q is the vector field on Q defined at $q \in Q$ by $u_Q(q) := \frac{d}{dt}\Big|_{t=0} q \exp(tu)$, where $\exp : \mathfrak{g} \to G$ is the exponential map.

Given a cost function $\ell : \mathfrak{g} \times Q \to \mathbb{R}$, also called here a Lagrangian, the *Clebsch optimal control problem* is

$$\min_{u(t)} \int_0^T \ell(u(t), q(t))dt \tag{1}$$

subject to the following conditions:

(A) $\dot{q}(t) = u(t)_Q(q(t))$;
(B) $q(0) = q_0$ and $q(T) = q_T$.

In order to formulate the main properties of this optimal control problem, we need to recall some definitions. The *partial functional derivative* of ℓ relative to $u \in \mathfrak{g}$ is the function $\frac{\delta \ell}{\delta u}(u, q) \in \mathfrak{g}^*$ defined by

$$\left\langle \frac{\delta \ell}{\delta u}(u, q), \delta u \right\rangle := \frac{d}{dt}\Big|_{s=0} \ell(u + s\delta u, q)$$

for any $\delta u \in \mathfrak{g}$, where $\langle \cdot, \cdot \rangle : \mathfrak{g}^* \times \mathfrak{g} \to \mathbb{R}$ denotes the (a weakly, for infinite dimensional $\mathfrak{g}$) non-degenerate duality pairing. If $\mathfrak{g}$ is infinite dimensional, we assume that $\frac{\delta \ell}{\delta u}$ exists. Usually, we just write $\frac{\delta \ell}{\delta u}$, the dependence on $(u, q) \in \mathfrak{g} \times Q$ being understood.

The *partial functional derivative* of ℓ relative to $q \in Q$ is the function $\frac{\delta \ell}{\delta q}(u, q) \in T_q^* Q$ defined by

$$\left\langle \frac{\delta \ell}{\delta q}(u, q), \delta q \right\rangle := \left. \frac{d}{dt} \right|_{s=0} \ell(u, q(s))$$

for any $\delta q \in T_q Q$, where $q(s) \in Q$ is a curve with $q(0) = q$, $\dot{q}(0) = \delta q$ and $\langle \cdot, \cdot \rangle : T_q^* Q \times T_q Q \to \mathbb{R}$, for all $q \in Q$ denotes the (a weakly, for infinite dimensional Q) non-degenerate duality pairing. If Q is infinite dimensional, we assume that $\frac{\delta \ell}{\delta q}$ exists. As before, the dependence of $\frac{\delta \ell}{\delta q}$ on $(u, q) \in \mathfrak{g} \times Q$ is understood, without being explicitly written.

The *Legendre transformation* of ℓ at $q \in Q$ is defined by $\mathfrak{g} \ni u \mapsto \mu(u, q) := \frac{\delta \ell}{\delta u}(u, q) \in \mathfrak{g}^*$. We say that ℓ is *hyperregular* if this map is a diffeomorphism, for every $q \in Q$. Under this hypothesis, we denote by $\mathfrak{g}^* \ni \mu \mapsto u(\mu, q) \in \mathfrak{g}$ its inverse, and let $h : \mathfrak{g}^* \times Q \to \mathbb{R}$ be the *associated Hamiltonian* given by $h(\mu, q) = \langle \mu, u(\mu, q) \rangle - \ell(u(\mu, q), q)$.

The *momentum map* for the cotangent lifted G-action on T^*Q is the map $\mathbf{J} : T^*Q \to \mathfrak{g}^*$, defined by $\langle \mathbf{J}(\alpha_q), u \rangle := \langle \alpha_q, u_Q(q) \rangle$, for any $\alpha_q \in T^*Q$, $u \in \mathfrak{g}$. This map is equivariant relative to the cotangent lifted G-action on T^*Q and the coadjoint G-action on $\mathfrak{g}^*$.

Finally, given $\alpha, \beta \in T_q^* Q$, the *vertical lift of β relative to α* is defined by

$$\mathrm{Ver}_\alpha \, \beta := \left. \frac{d}{ds} \right|_{s=0} (\alpha + s\beta) \in T_\alpha (T^* Q).$$

The main geometric properties of the Clebsch optimal control problem and its link with Lagrangian and Hamiltonian dynamics are summarized in the following theorem.

Theorem 1 *Let the Lie group G act on the right on Q and let $\ell : \mathfrak{g} \times Q \to \mathbb{R}$ be a cost function. Then:*

- *If $t \mapsto (u(t), q(t)) \in \mathfrak{g} \times Q$ is an extremal solution of the Clebsch optimal control problem (1), then there is a curve $t \mapsto \alpha(t) \in T^*Q$ satisfying $\pi(\alpha(t)) = q(t)$, where $\pi : T^*Q \to Q$ is the cotangent bundle projection, such that the following equations hold:*

$$\frac{\delta \ell}{\delta u} = \mathbf{J}(\alpha), \quad \dot{\alpha} = u_{T^*Q}(\alpha) + \mathrm{Ver}_\alpha \frac{\delta \ell}{\delta q}. \tag{2}$$

- *Equations* (2) *imply (a generalization of) the Euler-Poincaré equations for the control u, given by*

$$\frac{d}{dt}\frac{\delta \ell}{\delta u} = \mathrm{ad}^*_u \frac{\delta \ell}{\delta u} + \mathbf{J}\left(\frac{\delta \ell}{\delta q}\right). \tag{3}$$

- *If ℓ is hyperregular, then the second equation in* (2)*, in which the first equation is used, is Hamiltonian on T^*Q for the Hamiltonian*

$$H(\alpha_q) = h(\mathbf{J}(\alpha_q), q), \tag{4}$$

where $h : \mathfrak{g}^* \times Q \to \mathbb{R}$ is the Hamiltonian obtained from ℓ by Legendre transformation.
- *If ℓ is hyperregular, then equations* (3) *together with condition* (A) *can be equivalently written in terms of the Hamiltonian h as follows:*

$$\begin{cases} \dfrac{d}{dt}\mu = \mathrm{ad}^*_{\frac{\delta h}{\delta \mu}}\, \mu - \mathbf{J}\left(\dfrac{\delta h}{\delta q}\right) \\ \dfrac{d}{dt}q = \left(\dfrac{\delta h}{\delta \mu}\right)_Q (q). \end{cases} \tag{5}$$

Equations (5) *are Hamiltonian with respect to the Poisson bracket*

$$\{f, h\} = -\left\langle \mu, \left[\frac{\delta f}{\delta \mu}, \frac{\delta h}{\delta \mu}\right]\right\rangle + \left\langle \mathbf{J}\left(\frac{\delta f}{\delta q}\right), \frac{\delta h}{\delta \mu}\right\rangle - \left\langle \mathbf{J}\left(\frac{\delta h}{\delta q}\right), \frac{\delta f}{\delta \mu}\right\rangle$$

on $\mathfrak{g}^* \times Q$.

We refer to [21] for a proof of this theorem.

Remark 1 Equations (3), with initial condition $q_0 \in Q$ for the curve $q(t)$, can be obtained by Lagrangian reduction of the Euler-Lagrange equations associated to a Lagrangian $L : TG \to \mathbb{R}$ invariant under the action of the isotropy group $G_{q_0} := \{g \in G \mid q_0 g = q_0\} \subset G$ of q_0. From this point of view, the cost function ℓ emerges as the reduced Lagrangian associated to L via the relation $L(g, \dot{g}) = \ell(g^{-1}\dot{g}, q_0 g)$; see [22]. This explains why we alternatively called the cost function a Lagrangian. A similar comment applies, on the Hamiltonian side, to equations (5). Equations (3), resp., (5), are generalization of the Euler-Poincaré, resp., Lie-Poisson, equations for semidirect products (see [26]) and of the affine Euler-Poincaré, resp., affine Lie-Poisson, equations (see [20]).

Equations (3), together with $\dot{q} = u_Q(q)$ can be interpreted as the Euler-Lagrange equations (in the Lie algebroid sense) on the transformation Lie algebroid $E = \mathfrak{g} \times Q \to Q$ with anchor map $\rho : E \to TQ$, $\rho(u, q) = u_Q(q)$, see [38] and [29]. In this formalism, the condition $\dot{q} = u_Q(q)$ expresses the condition of admissibility for a curve $(u(t), q(t))$ in the algebroid E.

2.2 Restriction to G-Orbits

Note that, due to condition (A), the solution $q(t)$ of the Clebsch optimal control problem (1) necessarily preserves the G-orbit $\mathscr{O}$ of the initial condition q_0. Therefore, we always assume that q_0 and and q_T belong to the same G-orbit, in order to have a well posed problem. As a consequence, the Clebsch optimal control problem (1) on $\mathfrak{g} \times Q$ with $q_0, q_T \in \mathscr{O}$ has the same solutions as the *restricted Clebsch optimal control problem* on $\mathfrak{g} \times \mathscr{O}$ given by

$$\min_{u(t)} \int_0^T \ell^{\mathscr{O}}(u(t), q(t))dt \tag{6}$$

subject to the following conditions:

(A) $\dot{q}(t) = u(t)_{\mathscr{O}}(q(t))$;
(B) $q(0) = q_0$ and $q(T) = q_T$.

In (6), the cost function $\ell^{\mathscr{O}} : \mathfrak{g} \times \mathscr{O} \to \mathbb{R}$ is defined by $\ell^{\mathscr{O}}(u, q) = \ell(u, i(q))$, where $i : \mathscr{O} \hookrightarrow Q$ is the inclusion, and $u_{\mathscr{O}}$ denotes the infinitesimal generator of the G-action on $\mathscr{O}$. We have the relation $Ti(u_{\mathscr{O}}(q)) = u_Q(i(q))$, for all $q \in \mathscr{O}$.

Let us comment on the link between the stationarity conditions of both problems. We denote by $\mathbf{J}^{\mathscr{O}} : T^*\mathscr{O} \to \mathfrak{g}^*$, the momentum map associated to the cotangent lifted G-action on $T^*\mathscr{O}$. We have $\mathbf{J}^{\mathscr{O}}(T^*i(\alpha_q)) = \mathbf{J}(\alpha_q)$, for all $\alpha_q \in T^*Q|_{\mathscr{O}}$, where $T^*i : T^*Q|_{\mathscr{O}} \to T^*\mathscr{O}$ is cotangent map defined by i. Using these relations, one observes that if $\alpha(t) \in T^*Q$ is a solution of (2) with $\pi(\alpha(0)) = q_0$, then $\alpha(t) \in T^*Q|_{\mathscr{O}}$ and $\beta(t) := T^*i(\alpha(t)) \in T^*\mathscr{O}$ is a solution of

$$\frac{\delta \ell^{\mathscr{O}}}{\delta u} = \mathbf{J}^{\mathscr{O}}(\beta), \quad \dot{\beta} = u_{T^*\mathscr{O}}(\beta) + \mathrm{Ver}_\beta \frac{\delta \ell^{\mathscr{O}}}{\delta q}, \tag{7}$$

which is the stationarity condition of problem (6).

Note that if ℓ is hyperregular with associated Hamiltonian h, then $\ell^{\mathscr{O}}$ is hyperregular, with Hamiltonian $h^{\mathscr{O}} : \mathfrak{g}^* \times \mathscr{O} \to \mathbb{R}$ given by $h^{\mathscr{O}}(\mu, q) = h(\mu, i(q))$. Let $H^{\mathscr{O}} : T^*\mathscr{O} \to \mathbb{R}$ be the collective Hamiltonian associated to $h^{\mathscr{O}}$, i.e., $H^{\mathscr{O}}(\beta_q) := h^{\mathscr{O}}\left(\mathbf{J}^{\mathscr{O}}(\beta_q), q\right)$, for all $\beta_q \in T_q^*\mathscr{O}$. Then we have the relation $H^{\mathscr{O}} \circ T^*i = H$ on $T^*Q|_{\mathscr{O}}$, where $H : T^*Q \to \mathbb{R}$ is the collective Hamiltonian of h. This relation completely characterizes $H^{\mathscr{O}}$. If $\alpha(t)$ is a solution of Hamilton's equations for H on T^*Q, with $\pi(\alpha(0)) = q_0$, then necessarily $\alpha(t) \in T^*Q|_{\mathscr{O}}$ and the curve $\beta(t) := T^*i(\alpha(t))$ is a solution of the Hamilton equations for $H^{\mathscr{O}}$ on $T^*\mathscr{O}$.

2.3 Quadratic Cost Functions and the Normal Metric

In this paragraph, we study the Clebsch optimal control problem in the special case where its cost function is given by the kinetic energy of a given inner product on the Lie algebra. We then show that the extremals are geodesics relative to an induced Riemannian metric on orbits. Let γ be the inner product on $\mathfrak{g}$ and consider

$$\ell(u, q) = \frac{1}{2}\gamma(u, u). \tag{8}$$

Defining the flat operator $\mathfrak{g} \ni u \mapsto u^\flat \in \mathfrak{g}^*$ by $u^\flat := \gamma(u, \_)$, we have the functional derivatives

$$\frac{\delta \ell}{\delta u} = u^\flat \quad \text{and} \quad \frac{\delta \ell}{\delta q} = 0.$$

The stationarity conditions (2) and the Euler-Poincaré equations (3) read

$$\dot{\alpha}_q = u_{T^*Q}(\alpha), \qquad u^\flat = \mathbf{J}(\alpha_q), \qquad \text{and} \qquad \frac{d}{dt}u^\flat = \mathrm{ad}^*_u\, u^\flat.$$

The Hamiltonian Since the Lagrangian ℓ is hyperregular we can consider its associated Hamiltonian

$$h(\mu, q) = \frac{1}{2}\gamma(\mu^\sharp, \mu^\sharp),$$

where the sharp operator $\mathfrak{g}^* \ni \mu \mapsto \mu^\sharp \in \mathfrak{g}$ is defined as the inverse of the flat operator. The Hamiltonian $H : T^*Q \to \mathbb{R}$ defined in (4) is thus

$$H(\alpha_q) = \frac{1}{2}\gamma\left(\mathbf{J}(\alpha_q)^\sharp, \mathbf{J}(\alpha_q)^\sharp\right) =: \frac{1}{2}\kappa(q)(\alpha_q, \alpha_q),$$

where we defined the symmetric positive 2-contravariant tensor κ on Q by

$$\kappa(q)\left(\alpha_q, \beta_q\right) := \gamma\left(\mathbf{J}(\alpha_q)^\sharp, \mathbf{J}(\beta_q)^\sharp\right), \quad \text{for all } \alpha_q, \beta_q \in T^*Q.$$

Note that κ is not a co-metric, in general, since it has the kernel $[\mathfrak{g}_Q(q)]^\circ = [T_q\mathscr{O}]^\circ$, where $\mathscr{O}$ is the G-orbit containing q and $\mathfrak{g}_Q(q) = \{u_Q(q) \mid u \in \mathfrak{g}\}$. It is a co-metric if and only if the G-action is infinitesimally transitive, i.e., $\mathfrak{g}_Q(q) = T_qQ$ for all $q \in Q$.

We shall show below that the tensor κ, and hence the Hamiltonian H, are closely related to a particular Riemannian metric on the G-orbits, called the normal metric.

The normal metric on orbits We now recall from [21] the definition of the normal metric on G-orbits. Given $q \in Q$, let $\mathfrak{g}_q := \{\xi \in \mathfrak{g} \mid \xi_Q(q) = 0\}$ denote the isotropy Lie algebra of q. Using the inner product γ on $\mathfrak{g}$, orthogonally decompose $\mathfrak{g} = \mathfrak{g}_q \oplus \mathfrak{g}_q^{\perp}$, and denote by $u = u_q + u^q$ the associated splitting of $u \in \mathfrak{g}$ in this direct sum. With these notations, the *normal metric* on a G-orbit $\mathscr{O}$ is defined by

$$\gamma_{\mathscr{O}}(q)(u_Q(q), v_Q(q)) := \gamma(u^q, v^q), \text{ for all } q \in Q \text{ and } u, v \in \mathfrak{g}. \tag{9}$$

For infinite dimensional $\mathfrak{g}$ and a weak inner product γ, the existence of the relevant objects is either postulated or verified in concrete cases.

Theorem 2 *Let G be a Lie group acting on the right on the smooth manifold Q and let γ be an inner product on $\mathfrak{g}$. Define the following symmetric positive 2-contravariant tensor on Q:*

$$\kappa(q)(\alpha_q, \beta_q) := \gamma\left(\mathbf{J}(\alpha_q)^\sharp, \mathbf{J}(\beta_q)^\sharp\right), \quad \alpha_q, \beta_q \in T^*Q. \tag{10}$$

Then:

- *κ is non-degenerate if and only if the G-action on Q is infinitesimally transitive.*
- *κ induces a well-defined co-metric $\kappa_{\mathscr{O}}$ on each G-orbit $\mathscr{O}$ of Q, through the following relation*

$$\kappa_{\mathscr{O}}(q)\left(T^*i\left(\alpha_{i(q)}\right), T^*i\left(\beta_{i(q)}\right)\right) = \kappa(i(q))\left(\alpha_{i(q)}, \beta_{i(q)}\right), \tag{11}$$

  for $q \in \mathscr{O}$ and $\alpha_{i(q)}, \beta_{i(q)} \in T^*_{i(q)}Q$. The co-metric $\kappa_{\mathscr{O}}$ is explicitly given by

$$\kappa_{\mathscr{O}}(q)(\alpha_q, \beta_q) = \gamma\left(\mathbf{J}^{\mathscr{O}}(\alpha_q)^\sharp, \mathbf{J}^{\mathscr{O}}(\beta_q)^\sharp\right), \quad \text{for} \quad \alpha_q, \beta_q \in T^*\mathscr{O}. \tag{12}$$

- *$\kappa_{\mathscr{O}}$ is the co-metric associated to the normal metric on $\mathscr{O}$, i.e.,*

$$\kappa(q)(\alpha_q, \beta_q) = \gamma_{\mathscr{O}}(q)(\alpha_q^\sharp, \beta_q^\sharp), \quad \text{for all} \quad q \in \mathscr{O} \quad \text{and all} \quad \alpha_q, \beta_q \in T^*_q\mathscr{O}, \tag{13}$$

  where $T^*_q\mathscr{O} \ni \alpha_q \mapsto \alpha_q^\sharp \in T_q\mathscr{O}$ is the sharp operator associated to $\gamma_{\mathscr{O}}$.

Proof Since the kernel of κ is $[\mathfrak{g}_Q(q)]^\circ$, κ is non-degenerate if and only if $\mathfrak{g}_Q(q) = T_qQ$, i.e., the action is infinitesimally transitive.

Let us show that $\kappa_{\mathscr{O}}$ in (11) is well-defined. If $\alpha_{i(q)}, \alpha'_{i(q)} \in T^*_{i(q)}Q$ are such that $T^*i(\alpha_{i(q)}) = T^*i(\alpha'_{i(q)})$, then $\alpha_{i(q)} - \alpha'_{i(q)} \in [T_q\mathscr{O}]^\circ$. Hence $\kappa(i(q))(\alpha_{i(q)}, \beta_{i(q)}) = \kappa(i(q))(\alpha'_{i(q)}, \beta_{i(q)})$, and similarly for $\beta_{i(q)}$. Since the kernel of $\kappa(q)$, for $q \in \mathscr{O}$, is $[\mathfrak{g}_Q(q)]^\circ = \ker(T^*_qi)$, it follows that $\kappa_{\mathscr{O}}$ is non-degenerate and hence a co-metric. Formula (12) follows from the relations (11) and $\mathbf{J}^{\mathscr{O}} \circ T^*i = \mathbf{J}$ on $T^*Q|_{\mathscr{O}}$.

To prove (13), we first note that for $\alpha_q \in T_q^*\mathscr{O}$ and $u \in \mathfrak{g}$,

$$\begin{aligned}
\alpha_q^\sharp = u_Q(q) &\Leftrightarrow \langle \alpha_q, v_Q(q) \rangle = \gamma_Q(q)(u_Q(q), v_Q(q)), \quad \forall\, v \in \mathfrak{g} \\
&\Leftrightarrow \left\langle \mathbf{J}^{\mathscr{O}}(\alpha_q), v \right\rangle = \gamma(u^q, v^q), \quad \forall\, v \in \mathfrak{g} \\
&\Leftrightarrow \left\langle \mathbf{J}^{\mathscr{O}}(\alpha_q), v \right\rangle = \gamma(u^q, v), \quad \forall\, v \in \mathfrak{g} \\
&\Leftrightarrow \gamma(\mathbf{J}^{\mathscr{O}}(\alpha_q)^\sharp, v) = \gamma(u^q, v), \quad \forall\, v \in \mathfrak{g},
\end{aligned}$$

so, we get $\mathbf{J}^{\mathscr{O}}(\alpha_q)^\sharp = u^q$, where $\sharp$ is associated to γ. Similarly, $\beta_q^\sharp = v_Q(q) \Leftrightarrow \mathbf{J}^{\mathscr{O}}(\beta_q)^\sharp = v^q$. We can thus write

$$\begin{aligned}
\kappa(q)(\alpha_q, \beta_q) &= \gamma\left(\mathbf{J}^{\mathscr{O}}(\alpha_q)^\sharp, \mathbf{J}^{\mathscr{O}}(\beta_q)^\sharp\right) = \gamma(u^q, v^q) \\
&= \gamma_{\mathscr{O}}(q)(u_Q(q), v_Q(q)) = \gamma_{\mathscr{O}}(q)(\alpha_q^\sharp, \beta_q^\sharp)
\end{aligned}$$

as requested. ■

As we have seen in Sect. 2.2, we can restrict the Clebsch optimal control problem to the G-orbit $\mathscr{O}$ containing q_0. In this case, by using Theorem 2, the collective Hamiltonian turns out to be the kinetic energy of the normal metric, i.e.,

$$H^{\mathscr{O}}(\alpha_q) = \frac{1}{2}\kappa_{\mathscr{O}}(q)(\alpha_q, \alpha_q) = \frac{1}{2}\gamma_{\mathscr{O}}(\alpha_q^\sharp, \alpha_q^\sharp). \tag{14}$$

We thus obtain the following instance of Theorem 1 which allows to interpret the solution $q(t)$ of the Clebsch optimal control problem for (8) as geodesics on G-obits.

Corollary 1 (Clebsch optimal control and geodesics of the normal metric) *Let the Lie group G act on the right on Q, let γ be an inner product, suppose $q_0, q_T \in \mathscr{O}$, and consider the cost function $\ell(u, q) = \frac{1}{2}\gamma(u, u)$. Then:*

- *If $t \mapsto (u(t), q(t)) \in \mathfrak{g} \times \mathscr{O}$ is an extremal solution of the Clebsch optimal control problem (1), then there is a curve $t \mapsto \alpha(t) \in T^*\mathscr{O}$ covering $q(t)$, such that the following equations holds:*

$$u^\flat = \mathbf{J}(\alpha), \quad \dot{\alpha} = u_{T^*Q}(\alpha). \tag{15}$$

- *Equations (15) imply the Euler-Poincaré equations for the control u*

$$\frac{d}{dt}u^\flat = \mathrm{ad}_u^*\, u^\flat. \tag{16}$$

- *The second equation in* (15)*, in which the first equation is used, is Hamiltonian on* $T^*\mathscr{O}$ *for the Hamiltonian* (14)*. Therefore,* $q(t)$ *is a geodesic on* $\mathscr{O}$ *with respect to the normal metric* $\gamma_{\mathscr{O}}$*.*

  The previous discussion can be easily adapted to the case with a potential, i.e.,

$$\ell(u,q) = \frac{1}{2}\gamma(u,u) - \mathscr{V}(q).$$

Equations (15) and (16) then become

$$u^\flat = \mathbf{J}(\alpha), \quad \dot{\alpha} = u_{T^*Q}(\alpha) - \mathrm{Ver}_\alpha \frac{\delta \mathscr{V}}{\delta q} \quad \text{and} \quad \frac{d}{dt}u^\flat = \mathrm{ad}^*_u u^\flat - \mathbf{J}\left(\frac{\delta \mathscr{V}}{\delta q}\right). \tag{17}$$

The Hamiltonian $H^{\mathscr{O}} : T^*\mathscr{O} \to \mathbb{R}$ takes the standard kinetic plus potential form

$$H^{\mathscr{O}}(\alpha_q) = \frac{1}{2}\gamma_{\mathscr{O}}(\alpha_q^\sharp, \alpha_q^\sharp) + \mathscr{V}(q).$$

2.4 Optimal Control Associated to Geodesics

Suppose that (Q, g) is a Riemannian manifold and consider the minimization of the Riemannian distance

$$\min \int_0^T \frac{1}{2}\|\dot{q}(t)\|^2 dt \tag{18}$$

subject to the condition $q(0) = q_0$ and $q(T) = q_T$. Suppose that there is a *transitive* action of the Lie group G on Q. Then this minimization problem can be reformulated as a Clebsch optimal control problem, namely,

$$\min_{u(t)} \int_0^T \frac{1}{2}\|u_Q(q)\|^2 dt \tag{19}$$

subject to the following conditions:

(A) $\dot{q}(t) = u(t)_Q(q(t))$;
(B) $q(0) = q_0$ and $q(T) = q_T$.

We can thus write the cost function as

$$\ell(u,q) = \frac{1}{2}\|u_Q(q)\|^2 = \frac{1}{2}\left\langle \mathbb{I}(q)u, u\right\rangle,$$

where for each $q \in Q$, $\mathbb{I}(q)$ is the *locked inertia tensor* $\mathbb{I}(q) : \mathfrak{g} \to \mathfrak{g}^*$ defined by

$$\langle \mathbb{I}(q)u, v \rangle := g(q)(u_Q(q), v_Q(q)),$$

for any $u, v \in \mathfrak{g}$. The functional derivatives are

$$\frac{\delta \ell}{\delta u} = \mathbb{I}(q)u \in \mathfrak{g}_q^\circ \quad \text{and} \quad \frac{\delta \ell}{\delta q} = g(u_Q(q), \nabla u_Q(q)) = \frac{1}{2}\langle \mathbf{d}\mathbb{I}(q)(\cdot)u, u \rangle \in T_q^*Q,$$

where ∇ is the covariant derivative corresponding to the Riemannian metric. We note that $\ker(\mathbb{I}(q)) = \mathfrak{g}_q$ and $\mathrm{im}(\mathbb{I}(q)) = \mathfrak{g}_q^\circ$, therefore, ℓ is hyperregular if and only if the action is infinitesimally free, i.e., $\mathfrak{g}_q = \{0\}$.

In the hyperregular case, we obtain the Hamiltonian $h : \mathfrak{g}^* \times Q \to \mathbb{R}$, given by

$$h(\mu, q) = \frac{1}{2}\left\langle \mu, \mathbb{I}(q)^{-1}\mu \right\rangle, \tag{20}$$

and the Hamiltonian $H : T^*Q \to \mathbb{R}$ defined in (4) reads

$$H(\alpha_q) = h(\mathbf{J}(\alpha_q), q) = \frac{1}{2}\left\langle \mathbf{J}(\alpha_q), \mathbb{I}(q)^{-1}\mathbf{J}(\alpha_q) \right\rangle.$$

We extend now the definition of these Hamiltonians to the non-regular case. Let us fix an inner product γ on $\mathfrak{g}$. We write

$$\mathbb{I}(q) : \mathfrak{g} = \mathfrak{g}_q \oplus \mathfrak{g}_q^\perp \to \mathfrak{g}_q^\circ \oplus (\mathfrak{g}_q^\circ)^\perp = \mathfrak{g}_q^\circ \oplus (\mathfrak{g}_q^\perp)^\circ$$

relative to the orthogonal decomposition with respect to γ and $\gamma^\sharp$. Decomposing $\mu \in \mathfrak{g}^*$ as $\mu = \mu_1 + \mu_2 \in \mathfrak{g}_q^\circ \oplus (\mathfrak{g}_q^\perp)^\circ$, we define $\widehat{\mathbb{I}(q)}^{-1} : \mathfrak{g}^* \to \mathfrak{g}$ as

$$\widehat{\mathbb{I}(q)}^{-1}(\mu_1 + \mu_2) := \left(\mathbb{I}(q)|_{\mathfrak{g}_q^\perp}\right)^{-1}(\mu_1) \subset \mathfrak{g}_q^\perp,$$

where we note that $\mathbb{I}(q)|_{\mathfrak{g}_q^\perp} : \mathfrak{g}_q^\perp \to \mathfrak{g}_q^\circ$ is an isomorphism. We can thus define the Hamiltonians

$$h : \mathfrak{g}^* \times Q \to \mathbb{R}, \quad h(\mu, q) := \frac{1}{2}\left\langle \mu, \widehat{\mathbb{I}(q)}^{-1}(\mu) \right\rangle$$

$$H : T^*Q \to \mathbb{R}, \quad H(\alpha_q) := \frac{1}{2}g(q)(\alpha_q^\sharp, \alpha_q^\sharp).$$

Clearly, h extends (20) to the non-regular case. Concerning H, we have the following result.

Theorem 3 *Assume that the G action is transitive (but not necessarily free). Then, we have the relation*

$$H(\alpha_q) = h(\mathbf{J}(\alpha_q), q). \tag{21}$$

Moreover, similarly with the hyperregular case, the second equation in (2) (for the problem (19)), in which the first equation is used, is Hamiltonian on T^*Q for the Hamiltonian (21).

Proof If the action is transitive, we have $\left(\widehat{\mathbb{I}(q)}^{-1}\mathbf{J}(\alpha_q)\right)_Q(q) = \alpha_q^\sharp$. Indeed, for all $u \in \mathfrak{g}$, since $\mathbf{J}(\alpha_q) \in \mathfrak{g}_q^\circ$, we have

$$\begin{aligned} g\Big(\big(\widehat{\mathbb{I}(q)}^{-1}\mathbf{J}(\alpha_q)\big)_Q(q), u_Q(q)\Big) &= \Big\langle \mathbb{I}(q)\big(\widehat{\mathbb{I}(q)}^{-1}\mathbf{J}(\alpha_q)\big), u\Big\rangle \\ &= \Big\langle \mathbf{J}(\alpha_q), u\Big\rangle = \langle \alpha_q, u_Q(q)\rangle = g(\alpha_q^\sharp, u_Q(q)). \end{aligned}$$

Using this identity, we can check (21) as follows

$$h(\mathbf{J}(\alpha_q), q) := \frac{1}{2}\Big\langle \mathbf{J}(\alpha_q), \widehat{\mathbb{I}(q)}^{-1}(\mathbf{J}(\alpha_q))\Big\rangle = \frac{1}{2}\langle \alpha_q, \alpha_q^\sharp\rangle = H(\alpha_q).$$

To show the second result we note that in our case the first equation in (2) is $\mathbf{J}(\alpha_q) = \mathbb{I}(q)u \in \mathfrak{g}_q^\circ$. This relation is not invertible, but it tells us that u is equal to $\widehat{\mathbb{I}(q)}^{-1}(\mathbb{J}(\alpha_q))$ modulo an element in $\mathfrak{g}_q$. In the second equation in (2), we thus have $u_{T^*Q}(\alpha) = \big(\widehat{\mathbb{I}(q)}^{-1}(\mathbb{J}(\alpha_q))\big)_{T^*Q}$.

To check that this coincides with the Hamiltonian equation for H, we note that the Hamiltonian vector field of H in (21) is $X_H(\alpha) = \big(\frac{\delta h}{\delta \mu}(\mathbf{J}(\alpha_q), q)\big)_{T^*Q}(\alpha) - \mathrm{Ver}_\alpha \frac{\delta h}{\delta q}(\mathbf{J}(\alpha_q), q)$. Using the expression of h, we have $\frac{\delta h}{\delta \mu}(\mathbf{J}(\alpha_q), q) = \widehat{\mathbb{I}(q)}^{-1}(\mathbf{J}(\alpha_q))$. This proves the result. ■

3 Optimal Control on Stiefel Manifolds

An optimal control problem on Stiefel manifolds is introduced and studied in [9], as a generalization of the geodesic flow on the sphere (case $n = 1$) and the motion of the free N-dimensional rigid body (case $n = N$). In [21] this problem was generalized to arbitrary Lagrangians and formulated as a Clebsch optimal control problem of the form (1).

In this section, we show that the Clebsch optimal control problem on Stiefel manifolds offers a unified point of view for the formulation of several integrable systems. These systems turn out to be associated to two classes of cost functions, corresponding to the two situations studied in Sects. 2.3 and 2.4. From this setting,

we also deduce a geodesic interpretation of the solution of some of these integrable systems.

3.1 Stiefel Manifolds

For $n \leq N$, define the *Stiefel manifold* $V_n(\mathbb{R}^N)$ to be the set of orthonormal n-frames in $\mathbb{R}^N$ (i.e., an ordered set of n orthonormal vectors) (see, e.g., [25, page 301], [24, Chapter 5, §4,5]). So, $V_n(\mathbb{R}^N)$ is the set of linear isometric embeddings of $\mathbb{R}^n$ into $\mathbb{R}^N$. Let S^{N-1} denoted the unit sphere in $\mathbb{R}^N$. Since $V_n(\mathbb{R}^N) \subset \left(S^{N-1}\right)^n$ is closed, it follows that $V_n(\mathbb{R}^N)$ is compact. Collect the n vectors of an orthonormal frame in $\mathbb{R}^N$ as columns of a $N \times n$ matrix $Q \in V_n(\mathbb{R}^N)$. If $\mathrm{Mat}(N \times n)$ denotes the vector space of matrices having N rows and n columns, then the Stiefel manifold can be described as

$$V_n(\mathbb{R}^N) = \{Q \in \mathrm{Mat}(N \times n) \mid Q^{\mathsf{T}} Q = I_n\}, \tag{22}$$

where I_n is the $n \times n$ identity matrix. The dimension of $V_n(\mathbb{R}^N)$ is $Nn - (n+1)n/2$.

The characterization (22) of $V_n(\mathbb{R}^N)$ immediately shows that if $n = 1$, then $V_1(\mathbb{R}^N) = S^{N-1}$ and if $n = N$, then $V_N(\mathbb{R}^N) = O(N)$, the group of orthogonal isomorphisms of $\mathbb{R}^N$. If $n = 2$, then $V_2(\mathbb{R}^N)$ is the unit tangent bundle of S^{N-1}. Indeed, if $\{\mathbf{u}_1, \mathbf{u}_2\} \subset \mathbb{R}^N$ is an orthonormal frame, think of $\mathbf{u}_1$ as a point in S^{N-1} and of $\mathbf{u}_2$ as a unit vector in the tangent space $T_{\mathbf{u}_1} S^{N-1}$, and vice versa. If $n = N-1$ and $Q \in V_{N-1}(\mathbb{R}^N)$ has orthonormal columns $\{\mathbf{u}_1, \ldots, \mathbf{u}_{N-1}\}$ there is a unique unit vector $\mathbf{u}_0 \in \mathbb{R}^N$, orthogonal to the vector subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_{N-1}\}$, such that $\{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{N-1}\}$ is positively oriented, i.e., the determinant of the matrix $\widetilde{Q}$ whose columns are these basis vectors is > 0. Therefore, $\widetilde{Q} \in SO(N)$, the orthogonal orientation preserving isomorphisms of $\mathbb{R}^N$, i.e., the special orthogonal group. Conversely, given an element of $SO(N)$, the $N \times (N-1)$ matrix Q formed by the last $N-1$ columns is an element of $V_{N-1}(\mathbb{R}^N)$. This shows that $V_{N-1}(\mathbb{R}^N) = SO(N)$.

This last construction generalizes to give another characterization of $V_n(\mathbb{R}^N)$. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ be the standard orthonormal basis of $\mathbb{R}^N$ and $R \in O(N)$. Then the ith column of R is $R\mathbf{e}_i$. Define $\pi : O(N) \to V_n(\mathbb{R}^N)$, by $\pi(R) := [R\mathbf{e}_{N-n+1} \ \ldots \ R\mathbf{e}_N]$, where $[\mathbf{u}_i \ \ldots \ \mathbf{u}_j]$, $i < j$, denotes the matrix whose columns (in $\mathbb{R}^N$) are $\mathbf{u}_i, \ldots, \mathbf{u}_j$. Thus π maps $R \in O(N)$ (a rotation matrix in $\mathbb{R}^N$) to the orthonormal frame formed by its last n columns. This map is clearly surjective, since any orthonormal frame formed by $n \leq N$ vectors can be completed to an orthonormal basis and the matrix whose columns are the elements of such a basis is in $O(N)$. The rotation group $O(N)$ acts on $V_n(\mathbb{R}^N)$ by multiplication on the left. This action is transitive. Indeed, given two orthonormal frames of n vectors, i.e., $Q_1, Q_2 \in V_n(\mathbb{R}^N)$, complete each to an orthonormal basis of $\mathbb{R}^N$, i.e., obtain $\widetilde{Q}_1 = [Q_1' | Q_1]$, $\widetilde{Q}_2 = [Q_2' | Q_2] \in O(N)$, where Q_1', Q_2' have $N-n$ columns which

are orthonormal vectors in $\mathbb{R}^N$. Then $\widetilde{Q}_2\widetilde{Q}_1^\mathsf{T} \in O(N)$ and, since $(Q_1')^\mathsf{T}Q_1 = 0$, $Q_1^\mathsf{T}Q_1 = I_n$, we have

$$\widetilde{Q}_2\widetilde{Q}_1^\mathsf{T}Q_1 = [Q_2'\,Q_2][Q_1'\,Q_1]^\mathsf{T}Q_1 = [Q_2'\,Q_2]\begin{bmatrix}(Q_1')^\mathsf{T}Q_1\\ Q_1^\mathsf{T}Q_1\end{bmatrix} = [Q_2'\,Q_2]\begin{bmatrix}0\\ I_n\end{bmatrix} = Q_2.$$

The isotropy of the element $[\mathbf{e}_{N-n+1}\ \dots\ \mathbf{e}_N] = [0|I_n]^\mathsf{T} \in V_n(\mathbb{R}^N)$ consists of matrices $R \in O(N)$ satisfying

$$[0\ \dots\ 0\,R\mathbf{e}_{N-n+1}\ \dots\ R\mathbf{e}_N] = R[0\ \dots\ 0\,\mathbf{e}_{N-n+1}\ \dots\ \mathbf{e}_N] = [0\ \dots\ 0\,\mathbf{e}_{N-n+1}\ \dots\ \mathbf{e}_N],$$

i.e., the matrix R is of the form

$$R = \begin{bmatrix}R_1 & 0\\ R_2 & I_n\end{bmatrix}.$$

Since $R \in O(N)$, we must have

$$\begin{bmatrix}I_{N-n} & 0\\ 0 & I_n\end{bmatrix} = I_N = RR^\mathsf{T} = \begin{bmatrix}R_1 & 0\\ R_2 & I_n\end{bmatrix}\begin{bmatrix}R_1^\mathsf{T} & R_2^\mathsf{T}\\ 0 & I_n\end{bmatrix} = \begin{bmatrix}R_1R_1^\mathsf{T} & R_1R_2^\mathsf{T}\\ R_2R_1^\mathsf{T} & R_2R_2^\mathsf{T} + I_n\end{bmatrix}$$

and hence $R_1 \in O(N-n)$, $R_2R_2^\mathsf{T} = 0$. Taking the trace of the second relation gives the sum of squares of all entries of R_2, which implies that $R_2 = 0$. This shows that the $O(N)$-isotropy of the element $[\mathbf{e}_{N-n+1}\ \dots\ \mathbf{e}_N]$ is $O(N-n)$ embedded in $O(N)$ as the upper left $(N-n)\times(N-n)$ diagonal block and the lower right $n \times n$ block equal to I_n.

Conclusion: *$V_n(\mathbb{R}^N)$ is diffeomorphic to $O(N)/O(N-n)$ and the quotient map is $\pi : O(N) \ni [R\mathbf{e}_1\ \dots\ R\mathbf{e}_N] \mapsto [R\mathbf{e}_{N-n+1}\ \dots\ R\mathbf{e}_N] \in V_n(\mathbb{R}^N)$* (as an easy verification shows).

Note that $O(N-n)$ acts on $O(N)$ by multiplication on the right, where $O(N-n)$ is regarded as a subgroup of $O(N)$ when viewed as a 2×2 block matrix with $O(N-n)$ embedded as the upper left block, I_n the lower right block and the off-diagonal blocks equal to zero.

If $n < N$, all considerations above work with $O(N)$ and $O(N-n)$ replaced by $SO(N)$ and $SO(N-n)$, respectively; in constructing orthonormal bases in $\mathbb{R}^N$, always choose positively oriented ones, by changing the sign, if necessary, of one of the vectors used to complete the basis. Therefore, the conclusion above holds with $O(N)$ and $O(N-n)$ replaced by $SO(N)$ and $SO(N-n)$, respectively. Note that for $n = N$, these statements are false.

The tangent space at $Q \in V_n(\mathbb{R}^N)$ to the Stiefel manifold $V_n(\mathbb{R}^N)$ is given by

$$T_QV_n(\mathbb{R}^N) = \{V \in \mathrm{Mat}(N\times n) \mid V^\mathsf{T}Q + Q^\mathsf{T}V = 0\}. \tag{23}$$

We identify $T^*V_n(\mathbb{R}^N)$ with $TV_n(\mathbb{R}^N)$ using the pairing $T^*_Q V_n(\mathbb{R}^N) \times T_Q V_n(\mathbb{R}^N) \ni (P_Q, V_Q) \mapsto \text{Trace}\left(P_Q^\mathsf{T} V_Q\right) \in \mathbb{R}$ for every $Q \in V_n(\mathbb{R}^N)$.

Remark 2 It is also known that

$$V_n(\mathbb{R}^N) = SO(N)/SO(N-n) \to SO(N)/(SO(n) \times SO(N-n)) =: \widetilde{Gr}_n(\mathbb{R}^N)$$

is a principal $SO(n)$-bundle, where $\widetilde{Gr}_n(\mathbb{R}^N)$ is the Grassmannian of oriented n-planes in $\mathbb{R}^N$. The notation $Gr_n(\mathbb{R}^N)$ is reserved for the Grassmannian of n-planes in $\mathbb{R}^N$ (regardless of orientation). We will not use these bundles in our considerations.

Note that for $N > 1$, $V_1(\mathbb{R}^N) = S^{N-1} = \widetilde{Gr}_1(\mathbb{R}^N)$, while $Gr_1(\mathbb{R}^N) = \mathbb{RP}^{N-1}$. ◊

3.2 Clebsch Optimal Control on Stiefel Manifolds

From now on, we assume that $n < N$. We consider the *right* $SO(N)$-action on $V_n(\mathbb{R}^N)$ given by $Q \mapsto R^{-1}Q$ for $R \in SO(N)$. The infinitesimal generator of this action is $U_{V_n(\mathbb{R}^N)}(Q) = -UQ \in T_Q V_n(\mathbb{R}^N)$, $U \in \mathfrak{so}(N)$.

Given $Q_0, Q_T \in V_n(\mathbb{R}^N)$, the Clebsch optimal control problem (1) reads

$$\min_{U(t)} \int_0^T \ell(U(t), Q(t))dt \tag{24}$$

subject to the following conditions:

(A) $\dot{Q}(t) = -U(t)Q(t)$;
(B) $Q(0) = Q_0$ and $Q(T) = Q_T$.

We identify the dual $\mathfrak{so}(N)^*$ with itself using the non-degenerate pairing $\mathfrak{so}(N) \times \mathfrak{so}(N) \ni (U_1, U_2) \mapsto \text{Trace}(U_1^\mathsf{T} U_2) \in \mathbb{R}$. The cotangent bundle momentum map $\mathbf{J} : T^*V_n(\mathbb{R}^N) \to \mathfrak{so}(N)^*$ is easily verified to be

$$\mathbf{J}(Q, P) = \frac{1}{2}\left(QP^\mathsf{T} - PQ^\mathsf{T}\right).$$

The optimal control is thus given by $\delta\ell/\delta U = \left(QP^\mathsf{T} - PQ^\mathsf{T}\right)/2$ (see (2)). The cotangent lifted action on $T^*V_n(\mathbb{R}^N)$ reads $(Q, P) \mapsto (R^\mathsf{T}Q, R^\mathsf{T}P)$ and hence Hamilton's equations (2) become

$$\dot{Q} = -UQ, \quad \dot{P} = -UP + \frac{\delta\ell}{\delta Q}, \tag{25}$$

in this particular case. Recall that here $\delta\ell/\delta Q \in T^*_Q V_n(\mathbb{R}^N)$ denotes the functional derivative of ℓ relative to the above defined pairing. The optimal control U, given algebraically by $\delta\ell/\delta U = \left(QP^\mathsf{T} - PQ^\mathsf{T}\right)/2$, is necessarily the solution of the Euler-Poincaré equation (3) given in this particular case by

$$\frac{d}{dt}\frac{\delta\ell}{\delta U} = \left[\frac{\delta\ell}{\delta U}, U\right] + \frac{1}{2}\left(Q\left(\frac{\delta\ell}{\delta Q}\right)^\mathsf{T} - \frac{\delta\ell}{\delta Q}Q^\mathsf{T}\right). \tag{26}$$

3.2.1 Example 1: N-Dimensional Free Rigid Body

We consider as a cost function the free rigid body Lagrangian $\ell(U) = \frac{1}{2}\langle U, J(U)\rangle$, where $J(U) = \Lambda U + U\Lambda$, $\Lambda = \mathrm{diag}(\Lambda_1, \ldots \Lambda_N)$, $\Lambda_i + \Lambda_j > 0$ for $i \neq j$. The corresponding Clebsch optimal control falls into the setting studied in Sect. 2.3.

Since

$$M := \frac{\delta\ell}{\delta U} = J(U), \quad \frac{\delta\ell}{\delta Q} = 0,$$

equations (25) and (26) become

$$\dot{Q} = -UQ, \qquad \dot{P} = -UP \tag{27}$$

and

$$\dot{M} = [M, U], \quad \text{where} \quad M = J(U) = \frac{1}{2}\left(QP^\mathsf{T} - PQ^\mathsf{T}\right).$$

From Corollary 1, the solution $Q(t)$ is a geodesic on the $SO(N)$-orbit of $Q(0)$ in $V_n(\mathbb{R}^N)$, relative to the normal metric induced by the inner product $\gamma(U, V) := \langle U, J(V)\rangle$ on this orbit.

It is a remarkable fact that the free rigid body equations $\dot{M} = [M, U]$, $M = J(U) = \Lambda U + U\Lambda$ on $\mathfrak{so}(N)$, and indeed their generalization on any semisimple Lie algebra, are integrable [30]. A key observation in this regard, pointed out for the first time in [28], was that one can write the generalized rigid body equations as a Lax equation with parameter:

$$\frac{d}{dt}(M + \lambda\Lambda^2) = [M + \lambda\Lambda^2, U + \lambda\Lambda]. \tag{28}$$

The nontrivial coefficients of λ in the traces of the powers of $M + \lambda\Lambda^2$ then yield the right number of independent integrals in involution to prove integrability of the flow on the generic coadjoint orbits of $SO(n)$ (see also [35]).

Equation (28) is of the form $\dot{L} = [L, B]$ with L expressed in terms of the canonical variables as $L(Q, P) = \frac{1}{2}\left(QP^{\mathsf{T}} - PQ^{\mathsf{T}}\right) + \Lambda^2\lambda$.

Example 1A: symmetric representation of the N-dimensional free rigid body Consider the special case $n = N - 1$, i.e. $V_{N-1}(\mathbb{R}^N) = SO(N)$. Note that if the initial condition $P(0) \in SO(N)$, then the solution $(Q(t), P(t))$ of (27) preserves $SO(N) \times SO(N)$. Since in this case the formulation (27) of the free rigid body equation is symmetric in Q and P, it is called the *symmetric representation of the rigid body on* $SO(N) \times SO(N)$. As before, if (Q, P) is a solution of (27), then (Q, M), where $M = J(U)$ and $U = -\dot{Q}Q^{-1}$, satisfies the rigid body equations $\dot{Q} = -UQ$, $\dot{M} = [M, U]$ (see [6, 7] for a study of this system and its discretization).

Example 1B: $n = 1$, the rank 2 free rigid body We compute equations (25) and (26) for the case $n = 1$, i.e., $V_1(\mathbb{R}^N) = S^{N-1}$. For $(\mathbf{q}, \mathbf{p}) \in T^*S^{N-1}$ we get

$$\dot{\mathbf{q}} = -U\mathbf{q}, \qquad \dot{\mathbf{p}} = -U\mathbf{p}$$

and

$$\dot{M} = [M, U], \quad \text{where} \quad M = J(U) = \frac{1}{2}\left(\mathbf{q} \otimes \mathbf{p} - \mathbf{p} \otimes \mathbf{q}\right). \tag{29}$$

Note that, generically, M has rank 2. Associated to the Manakov equation (28), Moser [32, page 155] introduces the Lax pair matrices L and B given by

$$\begin{aligned} L(\mathbf{q}, \mathbf{p}) &= \Lambda^2 + a\mathbf{q} \otimes \mathbf{q} + b\mathbf{q} \otimes \mathbf{p} + c\mathbf{p} \otimes \mathbf{q} + d\mathbf{p} \otimes \mathbf{p}, \\ B(\mathbf{q}, \mathbf{p}) &= J^{-1}(\mathbf{q} \otimes \mathbf{p} - \mathbf{p} \otimes \mathbf{q}) + \lambda\Lambda, \end{aligned} \tag{30}$$

with $a = d = 0$, $b = -c = \frac{1}{2\lambda}$. For these values of the parameters, the expression of the matrix L in (30) is reminiscent of the expression of the momentum map (29) arising from the Clebsch optimal control formulation. We have

$$J^{-1}(\mathbf{q} \otimes \mathbf{p} - \mathbf{p} \otimes \mathbf{q}) = \frac{q_i p_j - q_j p_i}{2(\Lambda_i + \Lambda_j)}.$$

The geometric structures underlying definitions (30) will be given in [11].

Recall that the equations for the rank 2 free rigid body arise from an optimal control problem on S^{N-1}, rather than on the orthogonal group: minimize $\frac{1}{2}\int_0^T \langle U, J(U)\rangle dt$, where U is a skew symmetric control, subject to $\dot{\mathbf{q}} = -U\mathbf{q}$ as in (1).

From the result of Corollary 1, the curve $\mathbf{q}(t) \in S^{N-1}$ (there is only one orbit for $n = 1$) is a geodesic on S^{N-1} relative to the normal metric induced from the inner product $\gamma(U, V) := \langle U, J(V)\rangle$.

3.2.2 Example 2

We consider as a cost function the expression

$$\ell(U, Q) = \frac{1}{2}\langle \Lambda U Q, U Q\rangle - \mathscr{V}(Q), \tag{31}$$

where Λ is a given symmetric positive definite $N \times N$ matrix and $\mathscr{V} \in C^\infty(V_n(\mathbb{R}^N))$. The case $\mathscr{V} = 0$ is the geodesic problem studied in [9]. The first term in (31) is the kinetic energy associated to the Riemannian metric g on $V_n(\mathbb{R}^N)$ defined by

$$g_Q(V, W) = \langle \Lambda V, W\rangle = \mathrm{Tr}(V^\mathsf{T} \Lambda W) \qquad V, W \in T_Q V_n(\mathbb{R}^N).$$

The corresponding Clebsch optimal control falls into the setting studied in Sect. 2.4. In each of the examples mentioned below, the Clebsch optimal control formulation allows us to efficiently derive the explicit form of geodesic equations; see (37) and (41). This approach also yields a natural setting for generalizing certain integrable systems from the sphere to the Stiefel manifold, such as the C. Neumann problem.

Before analyzing this formulation for various examples, we first compute, in general, the stationarity conditions associated to the Clebsch optimal control problem.

The functional derivative of ℓ with respect to U is $\delta\ell/\delta U = \frac{1}{2}(Q Q^\mathsf{T} U \Lambda + \Lambda U Q Q^\mathsf{T})$. The relationship $\delta\ell/\delta U = \mathbf{J}(Q, P)$, which is equivalent to

$$Q Q^\mathsf{T} U \Lambda + \Lambda U Q Q^\mathsf{T} = Q P^\mathsf{T} - P Q^\mathsf{T}, \tag{32}$$

cannot be inverted in order to get U as a function of (Q, P) because the associated locked inertia tensor is not invertible since the $SO(N)$-action is not free.

Next, we calculate $\delta\ell/\delta Q$. Denoting $S := U\Lambda U$ (a symmetric matrix), we have

$$\left\langle \frac{\delta \ell}{\delta Q}, \delta Q \right\rangle = -\left\langle SQ + \frac{\delta \mathscr{V}}{\delta Q}, \delta Q \right\rangle = -\left\langle SQ - QQ^\mathsf{T} SQ + \frac{\delta \mathscr{V}}{\delta Q}, \delta Q \right\rangle$$

because $\left\langle QQ^\mathsf{T} SQ, \delta Q \right\rangle = 0$. Since $SQ - QQ^\mathsf{T} SQ \in T^*_Q V_n(\mathbb{R}^N)$ (see (23)), we get

$$\frac{\delta \ell}{\delta Q} = -SQ + QQ^\mathsf{T} SQ - \frac{\delta \mathscr{V}}{\delta Q}. \tag{33}$$

Thus, Hamilton's equations (2) become in this case

$$\dot{Q} = -UQ, \qquad \dot{P} = -UP + [QQ^\mathsf{T}, U\Lambda U]Q - \frac{\delta \mathscr{V}}{\delta Q}. \tag{34}$$

The corresponding Euler-Poincaré equations (26) are

$$\dot{M} = [M, U] + \frac{1}{2}[U\Lambda U, QQ^{\mathsf{T}}] - \frac{1}{2}\left(Q\left(\frac{\delta \mathscr{V}}{\delta Q}\right)^{\mathsf{T}} - \frac{\delta \mathscr{V}}{\delta Q}Q^{\mathsf{T}}\right), \tag{35}$$

where $M = \frac{\delta \ell}{\delta U} = \frac{1}{2}\left(QQ^{\mathsf{T}}U\Lambda + \Lambda UQQ^{\mathsf{T}}\right) \overset{(32)}{=} \frac{1}{2}\left(QP^{\mathsf{T}} - PQ^{\mathsf{T}}\right)$.

Example 2A: $n = 1$, $\mathscr{V} = 0$, geodesics on the ellipsoid Let us consider the case $n = 1$, i.e., $V_1(\mathbb{R}^N) = S^{N-1}$. The geodesic flow on S^{N-1} for the metric $g(q)(u, v) := \langle u, \Lambda v\rangle$, for $q \in S^{N-1}$, $u, v \in T_qS^{N-1}$ is equivalent to the geodesic flow on the ellipsoid $\bar{q}^{\mathsf{T}}\Lambda^{-1}\bar{q} = 1$, with $q = \Lambda^{-1/2}\bar{q}$. Equations (34) and (35) yield

$$\dot{q} = -Uq, \quad \dot{p} = -Up + [qq^{\mathsf{T}}, U\Lambda U]q, \quad \dot{M} = [M, U] + \frac{1}{2}[U\Lambda U, qq^{\mathsf{T}}] \tag{36}$$

where $M = \delta\ell/\delta U = \frac{1}{2}\left(qq^{\mathsf{T}}U\Lambda + \Lambda Uqq^{\mathsf{T}}\right) \overset{(32)}{=} \frac{1}{2}\left(qp^{\mathsf{T}} - pq^{\mathsf{T}}\right)$.

We now deduce from (36) the geodesic equations for the ellipsoid (see Theorem 1, (2)). Using the equality $M = \frac{1}{2}\left(qq^{\mathsf{T}}U\Lambda + \Lambda Uqq^{\mathsf{T}}\right)$, we get

$$\dot{M}\Lambda^{-1}q = \frac{1}{2}\left(qq^{\mathsf{T}}U^2q + \Lambda\dot{U}q(q^{\mathsf{T}}\Lambda^{-1}q) - \Lambda U^2q(q^{\mathsf{T}}\Lambda^{-1}q) + \Lambda Uq(q^{\mathsf{T}}U\Lambda^{-1}q)\right)$$

from where we solve for $\dot{U}q$, which inserted in $\ddot{q} = -\dot{U}q + U^2q$ yields

$$\ddot{q} = \left(-2\Lambda^{-1}\dot{M}\Lambda^{-1}q + \Lambda^{-1}qq^{\mathsf{T}}U^2q + Uqq^{\mathsf{T}}U\Lambda^{-1}q\right)(q^{\mathsf{T}}\Lambda^{-1}q)^{-1}.$$

Now, we replace in this formula $\dot{M}$ by its expression in (36) and we get the geodesic equations

$$\ddot{q} = -\frac{|\dot{q}|^2}{q^{\mathsf{T}}\Lambda^{-1}q}\Lambda^{-1}q. \tag{37}$$

The geodesic equations on the triaxial ellipsoid were solved by Jacobi. The complete solution is found in his course notes [27].

Remark 3 As a particular case, the geodesic equations on the sphere ($\Lambda = I_N$), are $\ddot{q} = -|\dot{q}|^2q$. ♢

Example 2B: $\mathscr{V} = 0$, geodesics on the Stiefel manifolds When $\mathscr{V} = 0$, (34) and (35) yield

$$\dot{Q} = -UQ, \quad \dot{P} = -UP + [QQ^{\mathsf{T}}, U\Lambda U]Q, \quad \dot{M} = [M, U] + \frac{1}{2}[U\Lambda U, QQ^{\mathsf{T}}], \tag{38}$$

where $M = \delta\ell/\delta U = \frac{1}{2}\left(QQ^{\mathsf{T}}U\Lambda + \Lambda UQQ^{\mathsf{T}}\right) \overset{(32)}{=} \frac{1}{2}\left(QP^{\mathsf{T}} - PQ^{\mathsf{T}}\right)$.

We now deduce from (38) the geodesic equations for the Stiefel manifolds (see Theorem 1, (2)). A direct computation yields

$$\dot{M}\Lambda^{-1}Q = \frac{1}{2}\Big(-UQQ^{\mathsf{T}}UQ + QQ^{\mathsf{T}}U^2Q + L(\dot{U}Q) - \Lambda U^2Q(Q^{\mathsf{T}}\Lambda^{-1}Q) + \Lambda UQ(Q^{\mathsf{T}}U\Lambda^{-1}Q)\Big), \tag{39}$$

where the linear operator L on the vector space of $N \times n$ matrices is defined by

$$L(X) := QQ^{\mathsf{T}}X + \Lambda XQ^{\mathsf{T}}\Lambda^{-1}Q. \tag{40}$$

Note that if $\Lambda = I_N$, then $L(X) = (I_N + QQ^{\mathsf{T}})X$.

We study the properties of the operator $L : \mathrm{Mat}(N \times n) \to \mathrm{Mat}(N \times n)$, where $\mathrm{Mat}(N \times n)$ denotes the real vector space of matrices with N rows and n columns. Recall that $\mathrm{Mat}(N \times n)$ has the natural inner product $\langle\langle A, B\rangle\rangle := \mathrm{Tr}\left(A^{\mathsf{T}}B\right)$. A direct computation shows that L is a linear symmetric operator relative to the inner product:

$$\Big\langle\Big\langle L(X), Y\Big\rangle\Big\rangle = \Big\langle\Big\langle X, L(Y)\Big\rangle\Big\rangle = \mathrm{Tr}\left(X^{\mathsf{T}}QQ^{\mathsf{T}}Y\right) + \mathrm{Tr}\left(X^{\mathsf{T}}\Lambda YQ^{\mathsf{T}}\Lambda^{-1}Q\right).$$

In particular,

$$\Big\langle\Big\langle L(X), X\Big\rangle\Big\rangle = \Big\langle Q^{\mathsf{T}}X, Q^{\mathsf{T}}X\Big\rangle + \mathrm{Tr}\left(X^{\mathsf{T}}\Lambda XQ^{\mathsf{T}}\Lambda^{-1}Q\right).$$

Note that $Q^{\mathsf{T}}\Lambda^{-1}Q$ is a symmetric positive definite matrix because Λ is a symmetric positive definite matrix and $Q \in V_n(\mathbb{R}^N)$. Therefore, there is a symmetric positive definite $n \times n$ matrix R such that $R^2 = Q^{\mathsf{T}}\Lambda^{-1}Q$. Hence the previous expression becomes

$$\Big\langle\Big\langle L(X), X\Big\rangle\Big\rangle = \Big\langle Q^{\mathsf{T}}X, Q^{\mathsf{T}}X\Big\rangle + \mathrm{Tr}\left((XR)^{\mathsf{T}}\Lambda(XR)\right)$$

and we note that each summand is ≥ 0. Hence $\langle\langle L(X), X\rangle\rangle = 0 \Rightarrow \mathrm{Tr}\left((XR)^{\mathsf{T}}\Lambda(XR)\right) = 0$. Since Λ is positive definite, we conclude that $XR = 0$ which implies that $X = 0$ because R is invertible. We conclude that $L : \mathrm{Mat}(N \times n) \to \mathrm{Mat}(N \times n)$ is a symmetric positive definite operator and hence invertible.

Returning to (39), we isolate $\left(\dot{U}Q\right)$, replace in this formula $\dot{M}$ by (38), and we get

$$L\left(\ddot{Q}\right) = L\left(-\dot{U}Q + U^2Q\right) = 2QQ^{\mathsf{T}}U^2Q \overset{(38)}{=} -2Q\dot{Q}^{\mathsf{T}}\dot{Q}, \tag{41}$$

which are the geodesic equations on the Stiefel manifold.

Remark 4 When $n = 1$, (41) coincide with (37). Indeed, in this case (41) becomes

$$q\left(q^{\mathsf{T}}\ddot{q}\right) + \Lambda\ddot{q}\left(q^{\mathsf{T}}\Lambda^{-1}q\right) = -2q\dot{q}^{\mathsf{T}}\dot{q}, \quad q \in S^{N-1}.$$

Since $q^{\mathsf{T}}q = 1$ we have $q\ddot{q}^{\mathsf{T}} + \dot{q}^{\mathsf{T}}\dot{q} = 0$, which then implies the geodesic equations on the ellipsoid (37). ◊

Example 2C: $n = 1$, $\Lambda = I_N$, $\mathscr{V}(q) = \frac{1}{2}Aq \cdot q$, $A := \mathrm{diag}(a_1, \ldots, a_N)$, the C. Neumann problem We now study the motion of a point on the sphere S^{N-1} under the influence of the quadratic potential $\frac{1}{2}Aq \cdot q$. For $N = 3$ the associated Hamilton equations were shown to be completely integrable by Carl Neumann (see [34]); for general N and a study of various geometric and dynamic aspects of this problem see [2, 3, 16, 31, 32, 36, 37].

Since $\dot{q} = -Uq$, the Lagrangian of this system is

$$\ell(U, q) = \frac{1}{2}\dot{q}^{\mathsf{T}}\dot{q} - \frac{1}{2}q^{\mathsf{T}}Aq = -\frac{1}{2}q^{\mathsf{T}}\left(U^2 + A\right)q \tag{42}$$

and hence

$$\frac{\delta\ell}{\delta U} = \frac{1}{2}\left(qq^{\mathsf{T}}U + Uqq^{\mathsf{T}}\right), \qquad \frac{\delta\ell}{\delta q} = -(U^2 + A)q + q(q^{\mathsf{T}}(U^2 + A)q).$$

Since $M := \frac{\delta\ell}{\delta U}$, (26) implies

$$\dot{M} = [M, U] + \frac{1}{2}[U^2 + A, qq^{\mathsf{T}}]. \tag{43}$$

On the other hand, using the definition of M, we get $\dot{M}q = \frac{1}{2}(qq^{\mathsf{T}}U^2q + \dot{U}q - U^2q)$ which yields the equations of motion for the Neumann system

$$\ddot{q} = -2\dot{M}q + qq^{\mathsf{T}}U^2q \overset{(43)}{=} -Aq + \left(Aq \cdot q - |\dot{q}|^2\right)q. \tag{44}$$

Example 2D: $\Lambda = I_N$, $\mathscr{V}(Q) = \frac{1}{2}\langle\langle AQ, Q\rangle\rangle$, $A := \operatorname{diag}(a_1, \ldots, a_N)$, the C. Neumann problem on Stiefel manifolds We now consider the motion of a point on the Stiefel manifold $V_n(\mathbb{R}^N)$ under the influence of the quadratic potential $\mathscr{V}(Q) = \frac{1}{2}\langle\langle AQ, Q\rangle\rangle$, where we can assume, without loss of generality, that $A = \operatorname{diag}(a_1, \ldots, a_N)$. We work in the generic case when $a_i \neq 0$ for all $i = 1, \ldots, N$.

Since $\dot{Q} = -UQ$, the Lagrangian of this system is

$$\ell(U, Q) = \frac{1}{2}\operatorname{Tr}\left(\dot{Q}^{\mathsf{T}}\dot{Q}\right) - \frac{1}{2}\operatorname{Tr}\left(Q^{\mathsf{T}}AQ\right) = -\frac{1}{2}\operatorname{Tr}\left(Q^{\mathsf{T}}\left(U^2 + A\right)Q\right) \tag{45}$$

and hence

$$\frac{\delta\ell}{\delta U} = \frac{1}{2}\left(QQ^{\mathsf{T}}U + UQQ^{\mathsf{T}}\right), \qquad \frac{\delta\ell}{\delta Q} = -(U^2 + A)Q + Q(Q^{\mathsf{T}}(U^2 + A)Q)$$

by (33). Since $M := \frac{\delta\ell}{\delta U}$, (26) implies

$$\dot{M} = [M, U] + \frac{1}{2}[U^2 + A, QQ^{\mathsf{T}}]. \tag{46}$$

On the other hand, using the definition of M, we get

$$\dot{M}Q = \frac{1}{2}\left(QQ^{\mathsf{T}}U^2Q + L(\dot{U}Q) - U^2Q\right), \tag{47}$$

where $L(X) := (I_N + QQ^{\mathsf{T}})X$, for $X \in \operatorname{Mat}(N \times n)$ (see (40)). Using (47), (46), and $2M = QQ^{\mathsf{T}}U + UQQ^{\mathsf{T}}$, we get $L\left(\ddot{Q}\right) = L\left(-\dot{U}Q + U^2Q\right) = -2Q\dot{Q}^{\mathsf{T}}\dot{Q} - AQ + QQ^{\mathsf{T}}AQ$, which yield the equations of motion for the Neumann system on $V_n(\mathbb{R}^N)$

$$\ddot{Q} = (I_N + QQ^{\mathsf{T}})^{-1}\left(-2Q\dot{Q}^{\mathsf{T}}\dot{Q} - AQ + QQ^{\mathsf{T}}AQ\right). \tag{48}$$

These equations for $A = 0$ coincide with (41) and for $n = 1$ with (44).

4 Clebsch Optimal Control Formulation for the Bloch-Iserles System

Given $N \in \mathfrak{so}(n)$, the Bloch-Iserles system [4, 12] is the ordinary differential equation on the space $\operatorname{sym}(n)$ of $n \times n$ symmetric matrices given by

$$\dot{X} = [X^2, N], \quad X(t) \in \operatorname{sym}(n). \tag{49}$$

Assume that N is invertible, $n = 2k$, and consider the symplectic group

$$Sp(2k, N^{-1}) := \left\{ Q \in GL(2k, \mathbb{R}) \mid Q^{\mathsf{T}} N^{-1} Q = N^{-1} \right\} \tag{50}$$

with Lie algebra $\mathfrak{sp}(2k, N^{-1}) = \{U \in \mathfrak{gl}(2k) \mid U^{\mathsf{T}} N^{-1} + N^{-1} U = 0\}$. The system (49) can be written as the Euler-Poincaré equation on $\mathfrak{sp}(2k, N^{-1})$ for the Lagrangian

$$\ell(U) = \frac{1}{2} \operatorname{Tr}((N^{-1}U)^2); \tag{51}$$

see [4]. Indeed, using the identification $\mathfrak{sp}(2k, N^{-1})^* := \operatorname{sym}(2k)$ with duality pairing $\langle\langle X, U \rangle\rangle = \operatorname{Tr}(XN^{-1}U)$ for $X \in \operatorname{sym}(2k)$ and $U \in \mathfrak{sp}(2k, N^{-1})$, we have $\delta\ell/\delta U = N^{-1}U$ and $\operatorname{ad}^*_U X = XN^{-1}UN - UX$, so the Euler-Poincaré equation $\frac{d}{dt}\frac{\delta \ell}{\delta U} = \operatorname{ad}^*_U \frac{\delta \ell}{\delta U}$ becomes $N^{-1}\dot{U} = N^{-1}UN^{-1}UN - UN^{-1}U$. Setting $X := N^{-1}U$, we recover (49). As a consequence, (49) describes left invariant geodesics on the Lie group (50).

When N is not invertible, then (49) describes left invariant geodesics on the Jacobi group and its generalizations; see [23].

4.1 Clebsch Optimal Control Formulation

Assume that N is invertible and consider the right action of the group $\operatorname{Sp}(2k, N^{-1})$ by multiplication on $\operatorname{GL}(2k, \mathbb{R})$. Consider the cost function $\ell : \mathfrak{sp}(2k, N^{-1}) \to \mathbb{R}$ given in (51). The associated Clebsch optimal control problem is

$$\min \int_0^T \ell(U) dt, \quad \text{subject to} \quad \dot{Q} = QU, \quad Q(0) = Q_0, \quad Q(T) = Q_T.$$

Conditions (2) read

$$\frac{\delta \ell}{\delta U} = \mathbf{J}(Q, P) = \frac{1}{2}(P^{\mathsf{T}} QN + (QN)^{\mathsf{T}} P), \quad \dot{Q} = QU, \quad \dot{P} = -PU^{\mathsf{T}}, \tag{52}$$

with respect to the duality pairing $\langle P, V \rangle := \operatorname{Tr}(P^{\mathsf{T}} V)$, for $V \in T_Q GL(2k, \mathbb{R})$ and $P \in T^*GL(2k, \mathbb{R})$. This optimal control problem falls into the setting studied in Sect. 2.3.

By Theorem 1, if Q, P satisfy the last two equations in (52), then $X = \frac{\delta \ell}{\delta U}$ verifies the Bloch-Iserles equations (49). Let's check this directly. We compute

$$\begin{aligned} 2\dot{X} = 2N^{-1}\dot{U} &= \frac{d}{dt}\left(P^{\mathsf{T}}QN - NQ^{\mathsf{T}}P\right) \\ &\overset{(52)}{=} -UP^{\mathsf{T}}QN + P^{\mathsf{T}}QUN - NU^{\mathsf{T}}Q^{\mathsf{T}}P + NQ^{\mathsf{T}}PU^{\mathsf{T}} \\ &= -U\left(P^{\mathsf{T}}QN - NQ^{\mathsf{T}}P\right) + \left(P^{\mathsf{T}}QN - NQ^{\mathsf{T}}P\right)N^{-1}UN \\ &= 2\left(XN^{-1}UN - UX\right) = 2\left[X^2, N\right] \end{aligned}$$

since $U = NX$, as stated.

This approach generalizes to the right action of $\mathrm{Sp}(2k, N^{-1})$ on $\mathfrak{gl}(2k, \mathbb{R})$ or, more generally, on the space $\mathrm{Mat}(n \times 2k)$ of rectangular $n \times 2k$ matrices.

Note that (49) is equivalent to the following Lax equation with parameter

$$\frac{d}{dt}(X + \lambda N) = \left[X + \lambda N, NX + XN + \lambda N^2\right]. \tag{53}$$

In this case, the Lax equation with parameter $\dot{L} = [L, B]$ has $L(Q, P) := P^{\mathsf{T}}QN - NQ^{\mathsf{T}}P + N\lambda$. For example, if $n = 1$, i.e., $\mathbf{q} \in \mathbb{R}^{2k}$ (seen as a row), then we have

$$\frac{\delta \ell}{\delta U} = \frac{1}{2}(\mathbf{p} \otimes \mathbf{q}N + \mathbf{q}N \otimes \mathbf{p}).$$

4.2 Symmetric Representation of the Bloch-Iserles System

Since $U \in \mathfrak{sp}(2k, N^{-1})$, the last two equations in system (52) are equivalent to

$$\dot{Q} = QU, \qquad \dot{P}N^{-1} = (PN^{-1})U,$$

which shows that if $U \in \mathfrak{sp}(2k, N^{-1})$ and the initial conditions $(Q(0), P(0)N^{-1}) \in Sp(2k, N^{-1}) \times Sp(2k, N^{-1})$, then $(Q(t), P(t)N^{-1}) \in Sp(2k, N^{-1}) \times Sp(2k, N^{-1})$.

Since $\delta\ell/\delta U = X = N^{-1}U$, the Hamiltonian $h : \mathrm{sym}(2k) \to \mathbb{R}$ has the expression

$$h(X) := \langle\langle X, U \rangle\rangle - \ell(U) = \frac{1}{2}\mathrm{Tr}(X^2).$$

Therefore, using (52), the Hamiltonian $H : T^*\mathfrak{gl}(2k, \mathbb{R}) \to \mathbb{R}$ is

$$H(Q, P) := h(\mathbf{J}(Q, P)) = \frac{1}{8} \operatorname{Tr}\left(\left(P^{\mathsf{T}} Q N - N Q^{\mathsf{T}} P\right)^2\right). \tag{54}$$

By Theorem 1, we get the following result.

Proposition 1 *Consider the canonical Hamiltonian system on $T^*\mathfrak{gl}(2k, \mathbb{R})$ with the symplectic structure*

$$\Omega_{\text{can}}((Q_1, P_1), (Q_2, P_2)) = \operatorname{Tr}(P_2^{\mathsf{T}} Q_1 - P_1^{\mathsf{T}} Q_2) \tag{55}$$

and Hamiltonian (54). Then its solutions are mapped by $\mathbf{J} : T^*\mathfrak{gl}(2k, \mathbb{R}) \to \operatorname{sym}(2k)$ to integral curves of the Bloch-Iserles system (49). The flow generated by (54) preserves the submanifold $\left\{(Q, P) \in \mathfrak{gl}(2k, \mathbb{R}) \times \mathfrak{gl}(2k, \mathbb{R}) \mid Q, P N^{-1} \in Sp(2k, N^{-1})\right\}$.

5 Clebsch Optimal Control Formulation for the Finite Toda Lattice

Consider a complex semisimple Lie algebra $\mathfrak{g}^{\mathbb{C}}$, its split normal real form $\mathfrak{g}$, and the decomposition $\mathfrak{g} = \mathfrak{b}_- \oplus \mathfrak{k}$, where $\mathfrak{k}$ is the compact normal Lie algebra and $\mathfrak{b}_-$ a Borel Lie subalgebra (we follow the notations of [10]).

Let us quickly recall how the full Toda equation can be viewed as the Euler-Poincaré equation on the Lie algebra $\mathfrak{b}_-$ for the Lagrangian $\ell(U) = \frac{1}{2}\kappa(U, U)$, with κ the Killing form. If we identify the dual Lie algebra as $(\mathfrak{b}_-)^* = \mathfrak{k}^{\perp}$ by using κ, we have $\delta\ell/\delta U = \pi_{\mathfrak{k}^\perp}(U)$ and $\operatorname{ad}^*_U \mu = -\pi_{\mathfrak{k}^\perp}\left(\left[U, \mu\right]\right)$, so the Euler-Poincaré equation reads

$$\pi_{\mathfrak{k}^\perp}(\dot{U}) = -\pi_{\mathfrak{k}^\perp}\left(\left[U, \pi_{\mathfrak{k}^\perp}(U)\right]\right) \tag{56}$$

Note that $(\pi_{\mathfrak{b}_-})|_{\mathfrak{k}^\perp} : \mathfrak{k}^{\perp} \to \mathfrak{b}_-$ is an isomorphism with inverse $(\pi_{\mathfrak{k}^\perp})|_{\mathfrak{b}_-} : \mathfrak{b}_- \to \mathfrak{k}^{\perp}$. We can rewrite the right hand side as

$$\begin{aligned}
\pi_{\mathfrak{k}^\perp}\left(\left[U, \pi_{\mathfrak{k}^\perp}(U)\right]\right) &= \pi_{\mathfrak{k}^\perp}\left(\left[\pi_{\mathfrak{b}_-}\pi_{\mathfrak{k}^\perp}(U), \pi_{\mathfrak{k}^\perp}(U)\right]\right) \\
&= \pi_{\mathfrak{k}^\perp}\left(\left[\pi_{\mathfrak{k}^\perp}(U) - \pi_{\mathfrak{k}}\pi_{\mathfrak{k}^\perp}(U), \pi_{\mathfrak{k}^\perp}(U)\right]\right) \\
&= -\pi_{\mathfrak{k}^\perp}\left(\left[\pi_{\mathfrak{k}}\pi_{\mathfrak{k}^\perp}(U), \pi_{\mathfrak{k}^\perp}(U)\right]\right) \\
&= -\left[\pi_{\mathfrak{k}}\pi_{\mathfrak{k}^\perp}(U), \pi_{\mathfrak{k}^\perp}(U)\right],
\end{aligned}$$

so defining $\mu := \pi_{\mathfrak{k}^\perp}(U)$, by using the isomorphism $(\pi_{\mathfrak{b}_-})|_{\mathfrak{k}^\perp} : \mathfrak{k}^\perp \to \mathfrak{b}_-$, we can rewrite (56) as

$$\dot{\mu} = [\pi_{\mathfrak{k}}(\mu), \mu],$$

which is the full Toda equation.

5.1 Clebsch Optimal Control Formulation for A_r-Toda Lattice

We first study to the A_r -Toda system. In this case, B_- is the group of lower triangular $(r+1) \times (r+1)$ matrices with determinant 1 and strictly positive diagonal elements; $\mathfrak{b}_-$ is the Lie algebra of lower triangular traceless matrices; $\mathfrak{k}$ is the Lie algebra of skew-symmetric matrices; $\mathfrak{b}_-^\perp$ consists of strictly lower triangular matrices; and $\mathfrak{k}^\perp$ consists of symmetric traceless matrices. Given $U \in \mathfrak{g} = \mathfrak{sl}(r+1, \mathbb{R})$, we have

$$\pi_{\mathfrak{b}_-}(U) = U_- + U_+^\mathsf{T} + U_0, \quad \pi_{\mathfrak{k}}(U) = U_+ - U_+^\mathsf{T},$$

where the indices $\pm$ and 0 on the matrices denote the strictly upper, lower, and diagonal part, respectively. For $X \in \mathfrak{g}^* = \mathfrak{sl}(r+1, \mathbb{R})$, we have

$$\pi_{\mathfrak{b}_-^\perp}(X) = X_- - X_+^\mathsf{T}, \quad \pi_{\mathfrak{k}^\perp}(X) = X_+^\mathsf{T} + X_0 + X_+.$$

We consider the action of B_- by multiplication on the right on $SL(r+1, \mathbb{R})$ and use the duality pairing between $TSL(r+1, \mathbb{R})$ and $T^*SL(r+1, \mathbb{R})$ given by the bi-invariant extension of the Killing form. For $P, V \in T_Q SL(r+1, \mathbb{R})$, we have $\langle P, V \rangle := \mathrm{Tr}(Q^{-1}PQ^{-1}V)$. With respect to this pairing, the momentum map is

$$\mathbf{J} : TSL(r+1, \mathbb{R}) \to \mathfrak{k}^\perp, \quad \mathbf{J}(Q, P) = \pi_{\mathfrak{k}^\perp}(Q^{-1}P).$$

The associated Clebsch optimal control problem, with cost function $\ell(U) = \frac{1}{2}\kappa(U, U)$, yields (see (2)),

$$\dot{Q} = QU, \quad \dot{P} = PU, \quad \pi_{\mathfrak{k}^\perp}(U) = \pi_{\mathfrak{k}^\perp}(Q^{-1}P). \tag{57}$$

The first two equations represent the *symmetric representation* of the A_r - Toda lattice. In particular, the solution curve $(Q(t), P(t))$ preserves $B_- \times B_-$ similarly to the rigid body case.

5.2 Clebsch Optimal Control Formulation for the Toda Lattice Associated to an Arbitrary Dynkin Diagram

For the general Toda system, we let B_- act on the right on G (the connected Lie group underlying the split normal real form). We identify T^*G with TG by using the bi-invariant duality pairing $\langle\,,\,\rangle_\kappa$ induced by κ.

In this case, the momentum map is given by

$$\mathbf{J} : T^*G = TG \to \mathfrak{k}^\perp, \quad \mathbf{J}(\alpha_Q) = \pi_{\mathfrak{k}^\perp}(TL_{Q^{-1}}\alpha_Q),$$

where we have $\alpha_Q \in T_QG = T_Q^*G$. Indeed,

$$\begin{aligned}\kappa(\mathbf{J}(\alpha_Q), U) &= \langle \alpha_Q, TL_Q U\rangle_\kappa = \left\langle TL_{Q^{-1}}\alpha_Q, U\right\rangle_\kappa = \kappa(TL_{Q^{-1}}\alpha_Q, U)\\ &= \kappa(\pi_{\mathfrak{k}^\perp}(TL_{Q^{-1}}\alpha_Q), U).\end{aligned}$$

For the non exceptional cases at least, the formulas can be written more explicitly since G is given by matrix groups; for A_r, B_r, C_r, D_r we have:

$$G = SL(r-1,\mathbb{R}), \quad G = SO(r+1,r), \quad G = Sp(2r,\mathbb{R}), \quad G = SO(r,r)$$

and the Killing form is given by a multiple of the trace: $\kappa(X,U) = c\,\mathrm{Tr}(XU)$. In this case, the momentum map reads $\mathbf{J}(Q,P) = \pi_{\mathfrak{k}^\perp}(Q^{-1}P)$. We note that since $P \in T_QG$, we have $Q^{-1}P \in \mathfrak{g}$, so $\pi_{\mathfrak{k}^\perp}(Q^{-1}P)$ is well-defined.

The associated Clebsch optimal control problem with cost function $\ell(U) = \frac{1}{2}\kappa(U,U)$ yields the same equations as in (57), understood now in the general sense of A_r, B_r, C_r, D_r. The first two equations being the symmetric representation of the Toda equations. From these conditions, one directly obtains:

$$\begin{aligned}\frac{d}{dt}\pi_{\mathfrak{k}^\perp}(Q^{-1}P) &= \pi_{\mathfrak{k}^\perp}(-Q^{-1}\dot{Q}Q^{-1}P + Q^{-1}\dot{P}) = \pi_{\mathfrak{k}^\perp}(-UQ^{-1}P + Q^{-1}PU)\\ &= -\pi_{\mathfrak{k}^\perp}([U, Q^{-1}P]) = -\pi_{\mathfrak{k}^\perp}([U, \pi_{\mathfrak{k}^\perp}(Q^{-1}P) + \pi_{\mathfrak{b}^\perp}(Q^{-1}P)])\\ &= -\pi_{\mathfrak{k}^\perp}([U, \pi_{\mathfrak{k}^\perp}(Q^{-1}P)] = -\pi_{\mathfrak{k}^\perp}([U, \pi_{\mathfrak{k}^\perp}(U)],\end{aligned}$$

which is the full Toda equation in Euler-Poincaré form (56).

As earlier, the solution curve $(Q(t), P(t))$ preserves the set $B_- \times B_-$.

6 Discrete Models

6.1 The Symmetric Representation of the Discrete Rigid Body

The Clebsch approach leads to a natural symmetric representation of the discrete rigid body equations of Moser and Veselov [33]. We now define the symmetric representation of the discrete rigid body equations as follows (see [7]):

$$Q_{k+1} = -U_k Q_k; \qquad P_{k+1} = -U_k P_k\,, \tag{58}$$

where $U_k \in SO(N)$ is defined by

$$\Lambda U_k - U_k^{\mathsf{T}} \Lambda = Q_k P_k^{\mathsf{T}} - P_k Q_k^{\mathsf{T}}. \tag{59}$$

We will write this as

$$J_D U_k = Q_k P_k^{\mathsf{T}} - P_k^{\mathsf{T}} Q_k^{\mathsf{T}}, \tag{60}$$

where $J_D : SO(N) \to \mathfrak{so}(N)$ (the discrete version of J) is defined by $J_D U = \Lambda U - U^{\mathsf{T}} \Lambda$. Notice that the derivative of J_D at the identity is J and hence, since J is invertible, J_D is a diffeomorphism from a neighborhood of the identity in $SO(N)$ to a neighborhood of 0 in $\mathfrak{so}(N)$. Using these equations, we have the algorithm $(Q_k, P_k) \mapsto (Q_{k+1}, P_{k+1})$ defined by: compute U_k from (59), compute Q_{k+1} and P_{k+1} using (58). Note that the update map for Q and P is done in parallel.

6.2 The Discrete Variational Problem in the Stiefel Case

The discrete variational problem on the Stiefel manifold is given by (see [9, 33])

$$\min_{Q_k} \sum_k \frac{1}{2} \langle \Lambda Q_{k+1}, Q_k \rangle, \tag{61}$$

subject to $Q_k^{\mathsf{T}} Q_k = I_n$, i.e., $Q_k \in V_n(\mathbb{R}^N)$. The extremal trajectories for this discrete variational problem are given by

$$\Lambda Q_{k+1} + \Lambda Q_{k-1} = Q_k B_k, \quad k \in \mathbb{Z}, \tag{62}$$

where $B_k = B_k^{\mathsf{T}}$ is a (symmetric) Lagrange multiplier matrix for the symmetric constraint $Q_k^{\mathsf{T}} Q_k = I_n$. Let us define $U_k := -Q_k Q_{k-1}^{\mathsf{T}}$ which implies that

$$Q_k = -U_k Q_{k-1},$$

Then (as in [9]) the following proposition gives the discrete extremal trajectories in terms of U_k and the discrete body momentum $M_k := \Lambda U_k - U_k^{\mathsf{T}}\Lambda$.

Proposition 2 *The extremal trajectories of the discrete variational problem (61) on the Stiefel manifold $V_n(\mathbb{R}^N)$ in terms of (M_k, U_k) are given by:*

$$M_{k+1} = U_k M_k U_k^{\mathsf{T}} + A_k, \tag{63}$$

where

$$A_k := U_k \Lambda\big(I_N - U_k U_k^{\mathsf{T}}\big) - \big(I_N - U_k U_k^{\mathsf{T}}\big)\Lambda U_k^{\mathsf{T}}. \tag{64}$$

6.3 Discrete Variational Problem for the Bloch-Iserles Problem

The natural optimization problem in this case is

$$\min_{U_k} \sum_k \frac{1}{2}\langle N^{-1}U_k, N^{-1}U_k\rangle, \tag{65}$$

subject to $Q_{k+1} = Q_k U_k$.

Here, as in the smooth case

$$\left\{ Q_k \in GL(2k, \mathbb{R}) \mid Q_k^{\mathsf{T}} N^{-1} Q_k = N^{-1} \right\}. \tag{66}$$

Thus we have

$$Q_k^{\mathsf{T}} N^{-1} Q_{k+1} = N^{-1} U_k \tag{67}$$

and hence the optimization problem may be reformulated as

$$\min_{Q_k} \sum_k \frac{1}{2}\langle Q_k^{\mathsf{T}} N^{-1} Q_{k+1}, Q_k^{\mathsf{T}} N^{-1} Q_{k+1}\rangle, \tag{68}$$

subject to

$$Q_k^{\mathsf{T}} N^{-1} Q_k = N^{-1}. \tag{69}$$

Choosing a skew symmetric matrix B_k of Lagrange multipliers we see that the relevant equations take the form

$$N^{-1} Q_{k+1} Q_{k+1} N^{-1} Q_k + N^{-1} Q_{k-1} Q_{k-1} N^{-1} Q_k + N^{-1} Q_k B_k = 0. \tag{70}$$

This gives a natural analogue of the Moser Veselov equations which we will analyze further in a future publication.

Acknowledgements AMB was partially supported by NSF grants DMS-1613819, AFSOR grant 9550-18-0028, INSPIRE 1343720 and the Simons Foundation. FGB was partially supported by ANR grant GEOMFLUID 14-CE23-0002-01. TSR was partially supported by National Natural Science Foundation of China grant number 11871334 and by Swiss NSF grant NCCR SwissMAP. We thank the referees for their useful comments.

References

1. Adams, M.R., Harnad, J., Previato E.: Isospectral Hamiltonian flows in finite and infinite dimensions. Commun. Math. Phys. **117**, 451–500 (1988)
2. Adler, M., van Moerbeke, P.: Completely integrable systems, Euclidean Lie algebras, and curves. Adv. Math. **38**(3), 267–317 (1980)
3. Adler, M., van Moerbeke, P.: Linearization of Hamiltonian systems, Jacobi varieties and representation theory. Adv. Math. **38**(3), 318–379 (1980)
4. Bloch, A.M., Brînzănescu, V., Iserles, A., Marsden, J.E., Ratiu, T.S.: A class of integrable flows on the space of symmetric matrices. Commun. Math. Phys. **290**, 399–435 (2009)
5. Bloch, A.M., Crouch, P.E.: Optimal control and the full Toda flow. Proc. CDC **36**, 1736–1740 (1997). IEEE
6. Bloch, A.M., Crouch, P.E., Marsden, J.E., Ratiu, T.S.: Discrete rigid body dynamics and optimal control. Proc. CDC **37**, 2249–2254 (1998)
7. Bloch, A.M., Crouch, P.E, Marsden, J.E., Ratiu, T.S.: The symmetric representation of the rigid body equations and their discretization. Nonlinearity **15**, 1309–1341 (2002)
8. Bloch, A.M., Crouch, P.E., Nordkivst, N., Sanyal, A.K.: Embedded geodesic problems and optimal control for matrix Lie groups. J. Geom. Mech. **3**, 197–223 (2011)
9. Bloch, A.M., Crouch, P.E., Sanyal, A.K.: A variational problem on Stiefel manifolds. Nonlinearity **19**(10), 2247–2276 (2006)
10. Bloch, A.M., Gay-Balmaz, F., Ratiu, T.S.: The geometric nature of the Flaschka transformation. Commun. Math. Phys. **352**(2), 457–517 (2017)
11. Bloch, A.M., Gay-Balmaz, F., Ratiu, T.S.: In progress (2017)
12. Bloch, A.M., Iserles, A.: On an isospectral Lie-Poisson system and its Lie algebra. Found. Comput. Math. **6**, 121–144 (2006)
13. Brînzănescu, V., Ratiu, T.S.: Algebraic complete integrability of the Bloch-Iserles system. Int. Math. Res. Not. IMRN. **14**, 5806–5817 (2015)
14. Deift, P., Li, L.C., Nanda, T., Tomei, C.: The Toda flow on a generic orbit is integrable. Comm. Pure Appl. Math. **39**(2), 183–232 (1986)
15. Deift, P., Li, L.C., Tomei, C.: Loop groups, discrete versions of some classical integrable systems, and rank 2 extensions. Mem. Amer. Math. Soc. **100**(479) (1992)
16. Devaney, R.L.: Transversal homoclinic orbits in an integrable system. Am. J. Math. **100**(3), 631–642 (1978)
17. Fedorov, Yu.N., Jovanović, B.: Geodesic flwos and Neumann systems on Stiefel variables: geometry and integrability. Math. Z. **270**, 659–698 (2012)
18. Flaschka, H.: The Toda lattice. I. Existence of integrals. Phys. Rev. B (3) **9**, 1924–1925 (1974a)
19. Flaschka, H.: On the Toda lattice. II. Inverse-scattering solution. Progr. Theoret. Phys. **51**, 703–716 (1974b)
20. Gay-Balmaz, F., Ratiu, T.S.: The geometric structure of complex fluids. Adv. Appl. Math. **42**(2), 176–275 (2008)

21. Gay-Balmaz, F., Ratiu, T.S. Clebsch optimal control formulation in mechanics. J. Geom. Mech. **3**(1), 41–79 (2011)
22. Gay-Balmaz, F., Tronci, C.: Reduction theory for symmetry breaking with applications to nematic systems. Phys. D **239**(20–22), 1929–1947 (2009)
23. Gay-Balmaz, F., Tronci, C.: Vlasov moment flows and geodesics on the Jacobi group. J. Math. Phys. **53**(12), 36 pp (2012)
24. Greub, W., Halperin, S., Vanstone, R.: Connections, Curvature, and Cohomology, vol. II. Academic, New York (1973)
25. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
26. Holm D.D., Marsden, J.E., Ratiu, T.S.: The Euler-Poincaré equations and semidirect products with applications to continuum theories. Adv. Math. **137**, 1–81 (1998)
27. Jacobi, C.G.J.: Vorlesungen über Dynamik. C.G.J. Jacobi's Gesammelte Werke, Supplementband herausgegeben von A. Clebsch, zweite revidirte Ausgabe, Druck und Verlag von G. Reimer, Berlin (1884)
28. Manakov, S.V.: Note on the integration of Euler's equations of the dynamics of an n-dimensional rigid body. Funct. Anal. Appl. **10**, 328–329 (1976)
29. Martinez, E.: Variational calculus on Lie algebroids. ESIAM Control Optim. Calc. Var. **14**, 356–380 (2008)
30. Mishchenko, A.S., Fomenko, A.T.: On the integration of the Euler equations on semisimple Lie algebras Sov. Math. Dokl. **17**, 1591–1593 (1976)
31. Moser, J.: Various Aspects of Integrable Hamiltonian Systems. In: Dynamical Systems (C.I.M.E. Summer School, Bressanone, 1978). Progress in Mathematics, vol. 8, pp. 233–289. Birkhäuser, Boston (1980)
32. Moser, J.: Geometry of Quadrics and Spectral Theory. In: Chern Symposium 1979 (Proceedings of International Symposium Berkeley, California, 1979), pp. 147–188. Springer, New York/Berlin, 1980
33. Moser, J., Veselov A.: Discrete versions of some classical integrable systems and factorization of matrix polynomials. Commun. Math. Phys. **139**, 217–243 (1991)
34. Neumann, C.: De problemate quodam mechanica, quod ad primam integralium ultraellipticorum classem revocatur. J. Reine Angew. Math. **56**, 54–66 (1859)
35. Ratiu, T.S.: The motion of the free n-dimensional rigid body. Indiana U. Math. J. **29**, 609–627 (1980)
36. Ratiu, T.S.: The C. Neumann problem as a completely integrable system on an adjoint orbit. Trans. Am. Math. Soc. **264**(2), 321–329 (1981)
37. Uhlenbeck, K.: Minimal 2-spheres and tori in S^k, informal preprint (1975)
38. Weinstein, A.: Lagrangian manifolds and groupoids. Fields Inst. Commun. **7**, 207–231 (1996)

The Geometry of Characters of Hopf Algebras

Geir Bogfjellmo and Alexander Schmeding

Abstract Character groups of Hopf algebras appear in a variety of mathematical contexts. For example, they arise in non-commutative geometry, renormalisation of quantum field theory, numerical analysis and the theory of regularity structures for stochastic partial differential equations. A Hopf algebra is a structure that is simultaneously a (unital, associative) algebra, and a (counital, coassociative) coalgebra that is also equipped with an antiautomorphism known as the antipode, satisfying a certain property. In the contexts of these applications, the Hopf algebras often encode combinatorial structures and serve as a bookkeeping device. Several species of "series expansions" can then be described as algebra morphisms from a Hopf algebra to a commutative algebra. Examples include ordinary Taylor series, B-series, arising in the study of ordinary differential equations, Fliess series, arising from control theory and rough paths, arising in the theory of stochastic ordinary equations and partial differential equations. These ideas are the fundamental link connecting Hopf algebras and their character groups to the topics of the Abel-symposium 2016 on "Computation and Combinatorics in Dynamics, Stochastics and Control". In this note we will explain some of these connections, review constructions for Lie group and topological structures for character groups and provide some new results for character groups.

Character groups of Hopf algebras appear in a variety of mathematical contexts. For example, they arise in non-commutative geometry, renormalisation of quantum field theory [14], numerical analysis [10] and the theory of regularity structures

G. Bogfjellmo
Matematiska vetenskaper, Chalmers tekniska högskola och Göteborgs universitet, Göteborg, Sweden
e-mail: geir.bogfjellmo@chalmers.se

A. Schmeding (✉)
Institutt for matematiske fag, NTNU Trondheim, Trondheim, Norway
e-mail: schmeding@tu-berlin.de

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_6

for stochastic partial differential equations [25]. A Hopf algebra is a structure that is simultaneously a (unital, associative) algebra, and a (counital, coassociative) coalgebra that is also equipped with an antiautomorphism known as the antipode, satisfying a certain property. In the contexts of these applications, the Hopf algebras often encode combinatorial structures and serve as a bookkeeping device.

Several species of "series expansions" can then be described as algebra morphisms from a Hopf algebra $\mathscr{H}$ to a commutative algebra B. Examples include ordinary Taylor series, B-series, arising in the study of ordinary differential equations, Fliess series, arising from control theory and rough paths, arising in the theory of stochastic ordinary equations and partial differential equations. An important fact about such algebraic objects is that, if B is commutative, the set of algebra morphisms $\mathrm{Alg}(\mathscr{H}, B)$, also called *characters*, forms a group with product given by convolution

$$a * b = m_B \circ (a \otimes b) \circ \Delta_{\mathscr{H}}.$$

These ideas are the fundamental link connecting Hopf algebras and their character groups to the topics of the Abelsymposium 2016 on "Computation and Combinatorics in Dynamics, Stochastics and Control". In this note we will explain some of these connections, review constructions for Lie group and topological structures for character groups and provide some new results for character groups.

Topological and manifold structures on these groups are important to applications in the various fields outlined above. In many places in the literature the character group is viewed as "an infinite dimensional Lie group" and one is interested in solving differential equations on these infinite-dimensional spaces (we refer to [6] for a more detailed discussion and further references). This is due to the fact that the character group admits an associated Lie algebra, the Lie algebra of infinitesimal characters[1]

$$\mathfrak{g}(\mathscr{H}, B) := \{\phi \in \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B) \mid \phi(xy) = \phi(x)\varepsilon_{\mathscr{H}}(y) + \varepsilon_{\mathscr{H}}(x)\phi(y),\ \forall x, y \in \mathscr{H}\},$$

whose Lie bracket is given by the commutator bracket with respect to convolution. As was shown in [5], character groups of a large class of Hopf algebras are infinite-dimensional Lie groups. Note however, that in ibid. it was also shown that not every character group can be endowed with an infinite-dimensional Lie group structure. In this note we extend these results to a larger class Hopf algebras. To this end, recall that a topological algebra is a *continuous inverse algebra* (or CIA for short) if the set of invertible elements is open and inversion is continuous on this set (e.g. a Banach algebra). Then we prove the following theorem.

[1] Note that this Lie algebra is precisely the one appearing in the famous Milnor-Moore theorem in Hopf algebra theory [41].

Theorem A *Let $\mathscr{H} = \oplus_{n\in\mathbb{N}_0}\mathscr{H}_n$ be a graded Hopf algebra such that* $\dim \mathscr{H}_0 < \infty$ *and B be a commutative CIA. Then $\mathscr{G}(\mathscr{H}, B)$ is an infinite-dimensional Lie group whose Lie algebra is $\mathfrak{g}(\mathscr{H}, B)$.*

As already mentioned, in applications one is interested in solving differential equations on character groups (see e.g. [42] and compare [6]). These differential equations turn out to be a special class of equations appearing in infinite-dimensional Lie theory in the guise of regularity for Lie groups. To understand this and our results, we recall this concept now for the readers convenience.

Regularity (in the sense of Milnor) Let G be a Lie group modelled on a locally convex space, with identity element e, and $r \in \mathbb{N}_0 \cup \{\infty\}$. We use the tangent map of the left translation $\lambda_g : G \to G, x \mapsto gx$ by $g \in G$ to define $g.v := T_e\lambda_g(v) \in T_gG$ for $v \in T_e(G) =: \mathbf{L}(G)$. Following [20], G is called C^r *-semiregular* if for each C^r-curve $\gamma : [0, 1] \to \mathbf{L}(G)$ the initial value problem

$$\begin{cases} \eta'(t) &= \eta(t).\gamma(t) \\ \eta(0) &= e \end{cases}$$

has a (necessarily unique) C^{r+1}-solution $\mathrm{Evol}(\gamma) := \eta : [0, 1] \to G$. If furthermore the map

$$\mathrm{evol} : C^r([0, 1], \mathbf{L}(G)) \to G, \quad \gamma \mapsto \mathrm{Evol}(\gamma)(1)$$

is smooth, G is called C^r*-regular*.[2] If G is C^r-regular and $r \leq s$, then G is also C^s-regular. A C^∞-regular Lie group G is called *regular (in the sense of Milnor)* – a property first defined in [40]. Every finite-dimensional Lie group is C^0-regular (cf. [43]). In the context of this paper our results on regularity for character groups of Hopf algebras subsume the following theorem.

Theorem B *Let $\mathscr{H} = \oplus_{n\in\mathbb{N}_0}\mathscr{H}_n$ be a graded Hopf algebra such that* $\dim \mathscr{H}_0 < \infty$ *and B be a sequentially complete commutative CIA. Then $\mathscr{G}(\mathscr{H}, B)$ is C^0-regular.*

Recently, also an even stronger notion regularity called measurable regularity has been considered [19]. For a Lie group this stronger condition induces many Lie theoretic properties (e.g. validity of the Trotter product formula). In this setting, L^1-regularity means that one can solve the above differential equations for absolutely continuous functions (whose derivatives are L^1-functions). A detailed discussion of these properties can be found in [19]. However, we will sketch in Remark 19 a proof for the following proposition.

[2] Here we consider $C^r([0, 1], \mathbf{L}(G))$ as a locally convex vector space with the pointwise operations and the topology of uniform convergence of the function and its derivatives on compact sets.

Proposition C *Let $\mathscr{H} = \oplus_{n\in\mathbb{N}_0}\mathscr{H}_n$ be a graded Hopf algebra with* $\dim \mathscr{H}_0 < \infty$ *which is of countable dimension, e.g. $\mathscr{H}$ is of finite type. Then for any commutative Banach algebra B, the group $\mathscr{G}(\mathscr{H}, B)$ is L^1-regular.*

One example of a Hopf algebra whose group of characters represent a series expansion is the Connes–Kreimer Hopf algebra or Hopf algebra of rooted trees $\mathscr{H}_{\mathrm{CK}}$.

Brouder [10] established a very concrete link between $\mathscr{H}_{\mathrm{CK}}$ and B-series. B-series, due to Butcher [12], constitute an algebraic structure for the study of integrators for ordinary differential equations. In this context, the group of characters $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ is known as the Butcher group. The original idea was to isolate the numerical integrator from the concrete differential equation, and even from the surrounding space (assuming only that it is affine), thus enabling a study of the integrator *an sich*.

Another example connecting character groups to series expansions arises in the theory of regularity structures for stochastic partial differential equations (SPDEs) [11, 25]. In this theory one studies singular SPDEs, such as the continuous parabolic Anderson model (PAM, cf. the monograph [33]) formally given by

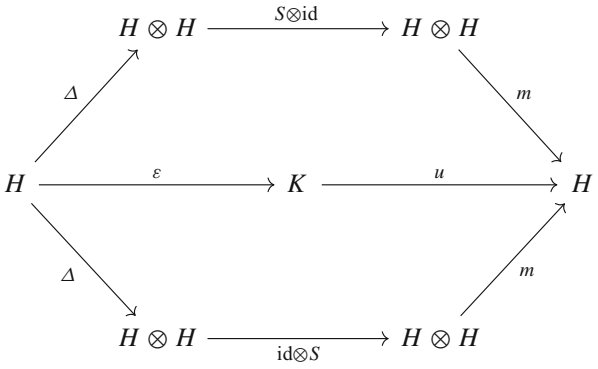
$$\left(\frac{\partial}{\partial t} - \Delta\right) u(t,x) = u(t,x)\zeta(x) \quad (t,x) \in]0,\infty[\times\mathbb{R}^2, \quad \zeta \text{ spatial white noise.}$$

We remark that due to the distributional nature of the noise, the product and thus the equation is ill-posed in the classical sense (see [25, p. 5]). To make sense of the equation, one wants to describe a potential solution by "local Taylor expansion" with respect to reference objects built from the noise terms. The analysis of this "Taylor expansion" is quite involved, since products of noise terms are not defined. However, it is possible to obtains Hopf algebras which describe the combinatorics involved. Their $\mathbb{R}$-valued character group $\mathscr{G}$ is then part of a so called regularity structure $(\mathscr{A}, \mathscr{T}, \mathscr{G})$ ([11, Definition 5.1]) used in the renormalisation of the singular SPDE. See Example 25 for a discussion of the Lie group structure for these groups.

1 Foundations: Character Groups and Graded Algebra

In this section we recall basic concepts and explain the notation used throughout the article. Whenever in the following we talk about algebras (or coalgebras or bialgebras) we will assume that the algebra (coalgebra, bialgebra) is unital (counital or unital and counital in the bialgebra case). Further $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ will always denote either the field of real or complex numbers (though many of the following concepts make sense over general base fields).

Definition 1 A *Hopf algebra* (over $\mathbb{K}$) $\mathscr{H}$ is a $\mathbb{K}-$bialgebra $(\mathscr{H}, m, \mathbb{1}_{\mathscr{H}}, \Delta, \varepsilon)$ equipped with an antiautomorphism S, called the *antipode*, such that the diagram



commutes.

In the diagram $u : \mathbb{K} \to \mathscr{H}, k \mapsto k\mathbb{1}_{\mathscr{H}}$ is the unit map of $\mathscr{H}$, i.e. the map which sends scalars to multiples of the unit $\mathbb{1}_{\mathscr{H}} \in \mathscr{H}$. We refer to [13, 35, 37, 46] for basic information on Hopf algebras.

Let B be a commutative algebra. The set of linear maps $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ forms a new algebra with the convolution product

$$\phi \star \psi = m_B \circ (\phi \otimes \psi)\Delta_{\mathscr{H}},$$

and unit $u_B \circ \varepsilon_{\mathscr{H}}$(where u_B is the unit map of B).

Recall that the invertible elements or *units* of an algebra A form a group, which we denote $A^{\times}$.

Definition 2 A linear map $\phi : \mathscr{H} \to B$ is called

1. a (B-valued) *character* if $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in \mathscr{H}$. The set of all characters is denoted $\mathscr{G}(\mathscr{H}, B)$.
2. a (B-valued) *infinitesimal character* if $\phi(ab) = \varepsilon_{\mathscr{H}}(b)\phi(a) + \varepsilon_{\mathscr{H}}(a)\phi(b)$ for all $a, b \in \mathscr{H}$. The set of all infinitesimal characters is denoted $\mathfrak{g}(\mathscr{H}, B)$.

Lemma 3 ([37, Proposition 21 and 22])

1. *$\mathscr{G}(\mathscr{H}, B)$ is a subgroup of the group of units* $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)^{\times}$. *On $\mathscr{G}(\mathscr{H}, B)$, the inverse is given by*

$$\phi^{\star -1} = \phi \circ S$$

2. *$\mathfrak{g}(\mathscr{H}, B)$ is a Lie subalgebra of the commutator Lie algebra* $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B), [\cdot, \cdot]$, *where the bracket is given by*

$$[\phi, \psi] = \phi \star \psi - \psi \star \phi$$

An algebraic property of characters and infinitesimal characters is that the algebraic exponential

$$\exp^\star(\phi) = \sum_{n=0}^{\infty} \frac{1}{n!}\phi^{\star n}$$

is a map from $\mathfrak{g}(\mathscr{H}, B)$ to $\mathscr{G}(\mathscr{H}, B)$. [37, Proposition 22].

In order to study the topological aspects of characters and infinitesimal characters of Hopf algebras, we need to assume at this step that B is a topological algebra, i.e., that B is topological vector space and that the product in B is a continuous bilinear function

$$\mu_B \colon B \times B \to B$$

We can then endow the space $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ with the *topology of pointwise convergence*. The sets $\mathfrak{g}(\mathscr{H}, B)$ and $\mathscr{G}(\mathscr{H}, B)$ are then closed subsets of $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$, and carry the induced topologies.

Proposition 4 *Let $\mathscr{H}$ be a Hopf algebra, and B a commutative, topological algebra. Endow* $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ *with the topology of pointwise convergence. Then*

- $(\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B), \star)$ *is a topological algebra,*
- $\mathscr{G}(\mathscr{H}, B)$ *is a topological group,*
- $\mathfrak{g}(\mathscr{H}, B)$ *is a topological Lie algebra.*

Proof It is sufficient to prove that $\star$ is continuous. Since $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ is endowed with the topology of pointwise convergence, it suffices to test convergence when evaluating at an element $h \in \mathscr{H}$. Using Sweedler notation, we get $\phi \star \psi(h) = \sum_{(h)} \phi(h_{(1)}) \star \psi(h_{(2)})$ where the multiplication is carried out in B. As point evaluations are continuous on $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$, and multiplication is continuous in B, $\star$ is continuous. □

The definition of $\star$ does not depend on the algebra structure of $\mathscr{H}$, only the coalgebra structure. We therefore get as a corollary:

Corollary 5 *Let C be a coalgebra, and B a commutative, topological algebra. Then* $(\mathrm{Hom}_{\mathbb{K}}(C, B), \star)$*, equipped with the topology of pointwise convergence, is a topological algebra.*

In Sect. 2 we will be able to state more about the topology and geometry of groups of characters, under further assumptions on $\mathscr{H}$ and B. In particular, we are interested in cases where $\mathscr{G}(\mathscr{H}, B)$ is an (infinite dimensional) Lie group, or a projective limit of finite dimensional Lie groups, i.e. a *pro-Lie* group. Both of these classes of topological groups can to some extent claim to be the generalization of finite dimensional Lie groups, and have been studied extensively for this reason (see e.g. [27, 29, 43]).

For many arguments later on gradings will be important. We recall the following basic definitions

Definition 6 Recall that a graded Hopf algebra $\mathscr{H} = \bigoplus_{n\in\mathbb{N}_0} \mathscr{H}_0$ is a Hopf algebra together with a grading as algebra and coalgebra (i.e. $\mathscr{H}_n \cdot \mathscr{H}_m \subseteq \mathscr{H}_{n+m}$ and $\Delta(\mathscr{H}_n) \subseteq \bigoplus_{k+l=n} \mathscr{H}_k \otimes \mathscr{H}_l$). In particular, $\mathscr{H}_0$ becomes a Hopf subalgebra of $\mathscr{H}$.

Note that for a graded Hopf algebra $\mathscr{H}$, identifying a mapping $f : \mathscr{H} \to B$ with its components on the grading induces a natural isomorphisms of topological vector spaces (with respect to the topologies of pointwise convergence)

$$\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B) = \mathrm{Hom}_{\mathbb{K}}(\bigoplus_{n\in\mathbb{N}_0} \mathscr{H}_n, B) \cong \prod_{n\in\mathbb{N}_0} \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}_n, B).$$

Hence $A = \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ becomes a densely graded topological vector space (see [5, Appendix B]). We denote by $A_n := \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}_n, B)$ the parts of the dense grading. Note that A_0 becomes a locally convex subalgebra of A by definition of the grading.

2 Geometry of Groups of Characters

In this section, we review results on geometry, topology and Lie theory for character groups of Hopf algebras $\mathscr{G}(\mathscr{H}, B)$. Further, in Sect. 2.1 we prove a new result which establishes a Lie group structure for character groups of non-connected Hopf algebras.

In general, the existence of a Lie group structure on the character group of a Hopf algebra $\mathscr{H}$ depends on structural properties of the underlying Hopf algebra (e.g. we need graded and connected Hopf algebras), the table below provides an overview of the topological and differentiable structures related to these additional assumptions (See Fig. 1).

In general, the character group need not admit a Lie group structure as was shown in [5, Example 4.11]. There we exhibited a character group of the group algebra of an abelian group of infinite rank which can not be a Lie group.

Remark 7 If the target algebra B is a weakly complete algebra, e.g. a finite dimensional algebra, the character group $\mathscr{G}(\mathscr{H}, B)$ is always a projective limit of finite-dimensional Lie groups. In [5, Theorem 5.6] we have proved that for an arbitrary Hopf algebra $\mathscr{H}$ and B a weakly complete algebra, $\mathscr{G}(\mathscr{H}, B)$ is a special

Hopf algebra $\mathscr{H}$	commutative algebra B	Structure on $\mathscr{G}(\mathscr{H}, B)$
arbitrary	weakly complete	pro-Lie group (cf. Remark 7)
graded and dim $\mathscr{H}_0 < \infty$	continuous inverse algebra	∞-dim. Lie group (Section 2.1)
graded and connected	locally convex algebra	∞-dim. Lie group (Section 2.2)

Fig. 1 Overview of topological and Lie group structures on character groups of Hopf algebras

kind of topological group, a so called pro-Lie group (see the survey [29]). A pro-Lie group is closely connected to its pro-Lie algebra which turns out to be isomorphic to $\mathfrak{g}(\mathscr{H}, B)$ for the pro-Lie group $\mathscr{G}(\mathscr{H}, B)$. Although pro-Lie groups do not in general admit a differentiable structure, a surprising amount of Lie theoretic properties carries over to pro-Lie groups (we refer to the monograph [27] for a detailed account).

Often the character group of a Hopf algebra will have more structure than a topological group. As we will see in the next example character groups often carry Lie group structures.

Example 8 Let G be a compact Lie group. Then we consider the set $\mathscr{R}(G)$ of *representative functions*, i.e. continuous functions $f : G \to \mathbb{R}$ such that the set of right translates $R_x f : G \to \mathbb{R}$, $R_x f(y) = f(yx)$ generates a finite-dimensional subspace of $C(G, \mathbb{R})$, cf. [28, Chapter 3] or [13, Section 3] for more information and examples.

Using the group structure of G and the algebra structure induced by $C(G, \mathbb{R})$ (pointwise multiplication), $\mathscr{R}(G)$ becomes a Hopf algebra (see [23, pp. 42–43]). Following Remark 7, we know that $\mathscr{G}(\mathscr{R}(G), \mathbb{R})$ becomes a topological group.

It follows from Tannaka-Kreĭn duality that as compact groups $\mathscr{G}(\mathscr{R}(G), \mathbb{R}) \cong G$, whence $\mathscr{G}(\mathscr{R}(G), \mathbb{R})$ inherits a posteriori a Lie group structure [23, Theorem 1.30 and 1.31].[3] Observe that the Lie group structure induced on $\mathscr{G}(\mathscr{R}(G), \mathbb{R})$ via Tannaka-Kreĭn duality coincides with the ones discussed in Sects. 2.1 and 2.2 (when these results are applicable to the Hopf algebra of representative functions).

Example 9 (The Butcher group) Let $\mathscr{T}$ denote the set of rooted trees, and $\mathscr{H}_{\mathrm{CK}} = \langle\!\langle \mathscr{T} \rangle\!\rangle$ the free commutative algebra generated by $\mathscr{T}$. The Grossman–Larson coproduct is defined on trees as

$$\Delta(\tau) = \tau \otimes \mathbb{1} + \sum_{\sigma} (\tau \setminus \sigma) \otimes \sigma$$

where the sum goes over all connected subsets σ of τ containing the root. Together with the algebra structure, The Grossman–Larson coproduct defines a graded, connected bialgebra structure on $\mathscr{H}_{\mathrm{CK}}$, and therefore also a Hopf algebra structure.

The characters $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ are the algebra morphisms $\mathrm{Hom}_{\mathrm{Alg}}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$. Clearly, we can identify

$$\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R}) \simeq \mathbb{R}^{\mathscr{T}}$$

[3] This is only a glimpse at Tannaka-Kreĭn duality, which admits a generalisation to compact topological groups (using complex representative functions, see [26, Chapter 7, §30] and cf. [23, p. 46] for additional information in the Lie group case). Also we recommend [28, Chapter 6] for more information on compact Lie groups.

In numerical analysis, the character group $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ is known as the Butcher group [12, 24]. This group is closely connected to a class of numerical integrators for ordinary differential equations. Namely, we let $\dot{y} = f(y)$ be an autonomous ordinary differential equation on an affine space E. Many numerical integrators [4] can be expanded in terms of the *elementary differentials* of the vector field f, i.e. as a series

$$y_{n+1} = y_n + a(\bullet)hf(y_n) + a(\overset{\bullet}{\bullet})h^2 f'f(y_n) + \cdots \tag{1}$$

The elementary differentials are in a natural one-to-one correlation with $\mathscr{T}$, and the series (1) thus defines an element in $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$. The crucial observation is that, (after a suitable scaling,) the convolution product in $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ corresponds to the composition of numerical integrators.

In the following, it will be established that $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ is a $\mathbb{R}$-analytic, C^0-regular Fréchet Lie group as well as a pro-Lie group. See [5, 8] for further details on the Lie group structure of $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$.

However, in some sense, the Butcher group $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ is too big to properly analyze numerical integrators. For every numerical integrator, the coefficients $a: \mathscr{T} \to \mathbb{R}$ satisfy a growth bound $|a(\tau)| \leq CK^{|\tau|}$. Elements of $\mathscr{G}(\mathscr{H}_{\mathrm{CK}}, \mathbb{R})$ satisfying such growth bounds form a subgroup, and even a Lie subgroup. However, the modelling space now becomes a Silva space.[5] This group is studied in the article [7].

In the next section we begin with establishing general results on the infinite-dimensional Lie group structure of Hopf algebra character groups. These manifolds will in general be modelled on spaces which are more general then Banach spaces. Thus the usual differential calculus has to be replaced by the so called Bastiani calculus (see [2], i.e. differentiability means existence and continuity of directional derivatives). For the readers convenience, Appendix 3 contains a brief recollection of this calculus.

2.1 Character Groups for $\mathscr{H}$ Graded with Finite Dimensional $\mathscr{H}_0$ and B a Continuous Inverse Algebra

In this section we consider graded but not necessarily connected Hopf algebras. In general, character groups of non-connected Hopf algebras do not admit a Lie group structure. Recall from example from [5, Example 4.11 (b)] that the character group

[4] To be exact: The class of integrators depending only on the affine structure (cf. [38]).

[5] Silva spaces arise as special inductive limits of Banach spaces, see [15] for more information. They are also often called (DFS)-space in the literature, as they can be characterised as the duals of Fréchet-Schwartz spaces.

of the group algebra of an infinite-group does in general not admit a Lie group structure. However, if the Hopf algebra is not too far removed from being connected (i.e. the 0-degree subspace $\mathscr{H}_0$ is finite-dimensional) and the target algebra is at least a continuous inverse algebra, we can prove that the character group $\mathscr{G}(\mathscr{H}, B)$ is an infinite-dimensional Lie group. This result is new and generalises [5] where only character groups of graded and connected Hopf algebras were treated (albeit the target algebra in the graded connected case may be a locally convex algebra).

2.1 Let $(A, \cdot)$ be a (real or complex) locally convex algebra (i.e. the locally convex topology of A makes the algebra product jointly continuous). We call $(A, \cdot)$ *continuous inverse algebra* (or *CIA* for short) if its unit group $A^\times$ is open and inversion $A^\times \to A, a \mapsto a^{-1}$ is continuous.

The class of locally convex algebras which are CIAs are of particular interest to infinite-dimensional Lie theory as their unit groups are in a natural way (analytic) Lie groups (see [16, 21]).

Before we establish the structural result, we have to construct an auxiliary Lie group structure in which we can realise the character group as subgroup. To construct the auxiliary Lie group, we can dispense with the Hopf algebra structure and consider only (graded) coalgebras at the moment. The authors are indebted to K.–H. Neeb for providing a key argument in the proof of the following Lemma.

Lemma 10 *Let (C, Δ) be a finite-dimensional coalgebra and B be a CIA. Then $(\mathrm{Hom}_{\mathbb{K}}(C, B), \star)$ is a CIA.*

Proof Consider the algebra $(A := \mathrm{Hom}_{\mathbb{K}}(C, \mathbb{K}), \star_A)$, where $\star_A$ is the convolution product. Then the algebraic tensor product $T := B \otimes_{\mathbb{K}} A$ with the product $(b \otimes \varphi) \cdot (c \otimes \psi) := bc \otimes \varphi \star_A \psi$ is a locally convex algebra. Since C is a finite-dimensional coalgebra, A is a finite-dimensional algebra. Due to an argument which was communicated to the authors by K.–H. Neeb, the tensor product of a finite-dimensional algebra and a CIA is again a CIA.[6] Thus it suffices to prove that the linear mapping defined on the elementary tensors via

$$\kappa \colon T \to \mathrm{Hom}_{\mathbb{K}}(C, B),\ b \otimes \varphi \mapsto (x \mapsto \varphi(x)b)$$

is an isomorphism of unital algebras. Since A is finite-dimensional, it is easy to see that κ is an isomorphism of locally convex spaces. Thus it suffices to prove that κ is a algebra morphism. To this end let ε be the counit of C. We recall that $\mathbb{1}_A = \varepsilon$, $\mathbb{1}_{\mathrm{Hom}_{\mathbb{K}}(C,B)} = (x \mapsto \varepsilon(x) \cdot \mathbb{1}_B)$ and $\mathbb{1}_T = \mathbb{1}_B \otimes \mathbb{1}_1 = \mathbb{1}_B \otimes \varepsilon$ are the units in A, $\mathrm{Hom}_{\mathbb{K}}(C, B)$ and T, respectively. Now $\kappa(\mathbb{1}_T) = (x \mapsto \varepsilon(x)\mathbb{1}_B) = \mathbb{1}_{\mathrm{Hom}_{\mathbb{K}}(C,B)}$, whence κ preserves the unit.

[6] We are not aware of a literature reference of this fact apart from the forthcoming book by Glöckner and Neeb [22]. To roughly sketch the argument from [22]: Using the regular representation of A one embeds $B \otimes_{\mathbb{K}} A$ in the matrix algebra $M_n(B)$ (where $n = \dim A$). Now as A is a commutant in $\mathrm{End}_{\mathbb{K}}(A)$, the same holds for $B \otimes A$ in $M_n(B)$. The commutant of a set in a CIA is again a CIA, whence the assertion follows as matrix algebras over a CIA are again CIAs (cf. [45, Corollary 1.2]).

As the elementary tensors span T, it suffices to establish multiplicativity of κ on elementary tensors $b_1 \otimes \varphi, b_2 \otimes \psi \in T$. For $c \in C$ we use Sweedler notation to write $\Delta(c) = \sum_{(c)} c_1 \otimes c_2$. Then

$$\begin{aligned}\kappa((b_1 \otimes \varphi) \cdot (b_2 \otimes \psi))(c) &= \kappa(b_1 b_2 \otimes \varphi \star_A \psi)(c) = \varphi \star_A \psi(x) b_1 b_2 \\ &= \sum_{(c)} \varphi(c_1)\psi(c_2) b_1 b_2 = \sum_{(c)} (\varphi(c_1) b_1)(\psi(c_2) b_2) \\ &= \sum_{(c)} \kappa(b_1 \otimes \varphi)(c_1)\kappa(b_2 \otimes \psi)(c_2) \\ &= \kappa(b_1 \otimes \varphi) \star \kappa(b_2 \otimes \psi)(c)\end{aligned}$$

shows that the mappings agree on each $c \in C$, whence κ is multiplicative. Summing up, $\mathrm{Hom}_{\mathbb{K}}(C, B)$ is a CIA as it is isomorphic to the CIA $B \otimes A$. □

Proposition 11 ($A^\times$ is a regular Lie group) *Let $C = \bigoplus_{n \in \mathbb{N}_0} C_n$ be a graded coalgebra with $\dim C_0 < \infty$ and B be a CIA. Then $A = (\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B), \star)$ is a CIA whose unit group $A^\times$ is Baker–Campbell–Hausdorff–Lie group (BCH–Lie group)[7] with Lie algebra $(A, [ie])$, where $[ie]$ denotes the commutator bracket with respect to $\star$.*

If in addition B is Mackey complete, then A is Mackey complete and the Lie group $A^\times$ is C^1-regular. If B is even sequentially complete, so is A and the Lie group $A^\times$ is C^0-regular. In both cases the associated evolution map is even $\mathbb{K}$-analytic and the Lie group exponential map is given by the exponential series.

Proof Recall from [5, Lemma 1.6 (c) and Lemma B.7] that the locally convex algebra A is a Mackey complete CIA since A_0 is such a CIA by Lemma 10 (both CIAs are even sequentially complete if B is so). Now the assertions concerning the Lie group structure of the unit group are due to Glöckner (see [16]).

To see that the Lie group $A^\times$ is C^k-regular ($k = 1$ for Mackey complete and $k = 0$ of sequentially complete CIA B), we note that the regularity of the unit group follows from the so called (GN)-property (cf. [5] and see Definition 12 below) and (Mackey) completeness of A. Having already established completeness, we recall from [5, Lemma 1.10] that A has the (GN)-property if A_0 has the (GN)-property. Below in Lemma 13 we establish that A_0 has the (GN)-property if B has the (GN)-property. Now B is a commutative Mackey complete CIA, whence B has the (GN)-property by [21, Remark 1.2 and the proof of Corollary 1.3]. Summing up, A has the (GN)-property and thus [21, Proposition 4.4] asserts that $A^\times$ is C^k-regular with analytic evolution map. □

[7] BCH-Lie groups derive their name from the fact that there is a neighborhood in their Lie algebra on which the Baker–Campbell–Hausdorff series converges and yields an analytic map. See Definition 35.

Before we can establish the (GN)-property for A_0 as in the proof of Proposition 11 we need to briefly recall this condition.

Definition 12 ((GN)-property) A locally convex algebra A is said to satisfy the *(GN)-property*, if for every continuous seminorm p on A, there exists a continuous seminorm q and a number $M \geq 0$ such that for all $n \in \mathbb{N}$, we have the estimate:

$$\left\| \mu_A^{(n)} \right\|_{p,q} := \sup\{ p(\mu_A^{(n)}(x_1, \dots, x_n)) \mid q(x_i) \leq 1,\ 1 \leq i \leq n \} \leq M^n. \tag{2}$$

Here, $\mu_A^{(n)}$ is the n-linear map $\mu_A^{(n)} \colon \underbrace{A \times \cdots \times A}_{n \text{ times}} \to A, (a_1, \dots, a_n) \mapsto a_1 \cdots a_n$.

Lemma 13 *Let C be a finite-dimensional coalgebra and B be a CIA with the (GN)-property. Then the following holds:*

1. *$(A := \mathrm{Hom}_{\mathbb{K}}(C, B), \star)$ has the (GN)-property.*
2. *If $(B, \|\cdot\|)$ is even a Banach algebra, then A is a Banach algebra.*

Proof Choose a basis $e_i, 1 \leq i \leq d$ for $\mathscr{H}_0$. Identifying linear maps with the coefficients on the basis, we obtain $A = \mathrm{Hom}_{\mathbb{K}}(C, B) \cong B^d$ (isomorphism of topological vector spaces). For every continuous seminorm p on B, we obtain a corresponding seminorm $p_\infty \colon A \to \mathbb{R}, \phi \mapsto \max_{1 \leq i \leq d} p(\phi(e_i))$ and these seminorms form a generating set of seminorms for the locally convex algebra A. Let us now write the coproduct of the basis elements as

$$\Delta(e_i) = \sum_{j,k} v_i^{jk} e_j \otimes e_k, \quad 1 \leq i \leq d \quad \text{for } v_i^{jk} \in \mathbb{K}.$$

To establish the properties of A we need an estimate of the structural constants, whence we a constant $K := d^2 \max_{i,j,k}\{|v_i^{jk}|, 1\}$

1. It suffices to establish the (GN)-property for a set of seminorms generating the topology of A. Hence it suffices to consider seminorms q_∞ induced by a continuous seminorm q on B. Since B has the (GN)-property there are a continuous seminorm p on B and a constant $M \geq 0$ which satisfy (2) with respect to the chosen q. We will now construct a constant such that (2) holds for the seminorms q_∞ and p_∞ taking the rôle of q and p.
   Observe that $q_\infty(\psi) \leq 1$ implies that $q(\psi(e_i)) \leq 1$ for each $1 \leq i \leq d$. Thus by choice of the constants a trivial computation shows that the constant KM satisfies (2) for q_∞, p_∞ and $n = 1$. We will show that this constant satisfies the inequality also for all $n > 1$ and begin for the readers convenience with the case

$n = 2$. Thus for $\psi_1, \psi_2 \in A$ with $q_\infty(\psi_l) \leq 1, l = 1, 2$ and $1 \leq k \leq d$ we have

$$p(\psi_1 \star \psi_2(e_i)) = p\left(\sum_{j,k} \nu_i^{jk} \psi_1(e_j)\psi_2(e_k)\right) \leq \sum_{j,k} |\nu_i^{jk}| \underbrace{p(\psi_1(e_i)\psi_2(e_j))}_{\leq M^2}$$

$$\leq \underbrace{K}_{\geq 1} M^2 \leq (KM)^2$$

As the estimate neither depends on i nor on the choice of ψ_1, ψ_2 (we only used $q_\infty(\psi_l) \leq 1$), KM satisfies $\left\|\mu_A^{(2)}\right\|_{p_\infty, q_\infty} \leq (KM)^2$. Now for general $n \geq 2$ we choose $\psi_l \in A$ with $q_\infty(\psi_l) \leq 1$ and $1 \leq l \leq n$. As convolution is associative, $\psi_1 \star \cdots \star \psi_n$ is obtained from applying $\psi_1 \otimes \cdots \otimes \psi_n$ to the iterated coproduct $\Delta^n := \mathrm{id}_C \otimes \Delta^{n-1}$, $\Delta^1 := \Delta$ and subjecting the result to the n-fold multiplication map $B \otimes B \otimes \cdots \otimes B \to B$ of the algebra B. Hence one obtains the formula

$$p(\psi_1 \star \psi_2 \star \cdots \star \psi_n(e_i))$$

$$\leq \underbrace{\sum_{j_1,k_1} \sum_{j_2,k_2} \cdots \sum_{j_{n-1},k_{n-1}}}_{\#\text{of terms} = d^2 \cdot d^2 \cdots d^2 = (d^2)^{n-1}} \underbrace{|\nu^i_{j_1,k_1}||\nu^{k_1}_{j_2,k_2}| \cdots |\nu^{k_{n-2}}_{j_{n-1},k_{n-2}}|}_{\leq (\max_{i,j,k}\{|\nu_i^{jk}|,1\})^{n-1}} p\left(\psi_1(e_{k_1}) \prod_{2 \leq r \leq n} \psi_r(e_{k_r})\right)$$

$$\leq K^{n-1} M^n \leq (KM)^n.$$

Again the estimate does neither depend on e_i nor on the choice of $\psi_1, \ldots, \psi_n$, whence we see that one can take KM in general as a constant which satisfies (2) for q_∞ and p_∞. We conclude that A has the (GN)-property if B has the (GN)-property.

2. Let now $(B, \|\cdot\|)$ be a Banach algebra, then $A \cong B^d$ is a Banach space with the norm $\|\phi\|_\infty := \max_{1 \leq i \leq d} \|\phi(e_i)\|$. To prove that A admits a submultiplicative norm, define the norm $\|\alpha\|_K := K \|\alpha\|_\infty$ (for K the constant chosen above). By construction $\|\cdot\|_K$ is equivalent to $\|\cdot\|_\infty$ and we will prove that $\|\cdot\|_K$ is submultiplicative. Consider $\alpha, \beta \in A$ and compute the norm of $\alpha \star \beta$ on a basis element

$$\|\alpha \star \beta(e_i)\| = \|m_B \circ (\alpha \otimes \beta) \circ \Delta(e_i)\| \leq \sum_{j,k} |\nu_i^{jk}| \, \|\alpha(e_j)\beta(e_k)\|$$

$$\leq \sum_{j,k} |\nu_i^{jk}| \, \|\alpha(e_j)\| \, \|\beta(e_k)\| \leq K \|\alpha\|_\infty \|\beta\|_\infty.$$

In passing from the first to the second row we have used that the norm on B is submultiplicative. Summing up, we obtain $\|\alpha \star \beta\|_K \leq \|\alpha\|_K \|\beta\|_K$ whence A is a Banach algebra. □

In case the Hopf algebra $\mathscr{H}$ is only of countable dimension, e.g. a Hopf algebra of finite type, and B is a Banach algebra, the unit group $A^\times$ satisfies an even stronger regularity condition.

Lemma 14 *Let $C = \bigoplus_{n\in\mathbb{N}_0}$ be a graded coalgebra with $\dim C_0 < \infty$ and B be a Banach algebra. Assume in addition that C is of countable dimension, then the Lie group $A^\times$ from Proposition 11 is L^1-regular.*

Proof If C is of countable dimension, then $A = \mathrm{Hom}_{\mathbb{K}}(C, B)$ is a Fréchet space (as it is isomorphic as a locally convex space to a countable product of Banach spaces). Now [19, Proposition 7.5] asserts that the unit group $A^\times$ of a continuous inverse algebra which is a Fréchet space will be L^1-regular if A is locally m-convex. However, by the fundamental theorem for coalgebras [39, Theorem 4.12], $C = \lim_{\rightarrow} \Sigma_n$ is the direct limit of finite dimensional subcoalgebras. Dualising, Lemma 13 shows that $A = \lim_{\leftarrow} \mathrm{Hom}_{\mathbb{K}}(\Sigma_n, B)$ is the projective limit of Banach algebras. As Banach algebras are locally m-convex and the projective limit of a system of locally m-convex algebras is again locally m-convex (cf. e.g. [36, Chapter III, Lemma 1.1]), the statement of the Lemma follows. □

To establish the following theorem we will show now that the Lie group $A^\times$ induces a Lie group structure on the character group of the Hopf algebra. Note that the character group is a closed subgroup of $A^\times$, but, contrary to the situation for finite dimensional Lie groups, closed subgroups do not automatically inherit a Lie group structure (see [43, Remark IV.3.17] for a counter example).

Theorem 15 *Let $\mathscr{H} = \bigoplus_{n\in\mathbb{N}_0} \mathscr{H}_n$ be a graded Hopf algebra with $\dim \mathscr{H}_0 < \infty$. Then for any commutative CIA B, the group $\mathscr{G}(\mathscr{H}, B)$ of B-valued characters of $\mathscr{H}$ is a ($\mathbb{K}-$analytic) Lie group. Furthermore, we observe the following.*

(i) *The Lie algebra $\mathbf{L}(\mathscr{G}(\mathscr{H}, B))$ of $\mathscr{G}(\mathscr{H}, B)$ is the Lie algebra $\mathfrak{g}(\mathscr{H}, B)$ of infinitesimal characters with the commutator bracket $[\phi, \psi] = \phi \star \psi - \psi \star \phi$.*
(ii) *$\mathscr{G}(\mathscr{H}, B)$ is a BCH–Lie group which is locally exponential, i.e. the Lie group exponential map $\exp\colon \mathfrak{g}(\mathscr{H}, B) \to \mathscr{G}(\mathscr{H}, B)$, $x \mapsto \sum_{n=0}^{\infty} \frac{x^{\star n}}{n!}$ is a local $\mathbb{K}$-analytic diffeomorphism.*

Proof Recall from Propositions 11 and 4 that $\mathscr{G}(\mathscr{H}, B)$ is a closed subgroup of the locally exponential Lie group $(A^\times, \star)$. We will now establish the Lie group structure using a well-known criterion for locally exponential Lie groups: Let $\exp_A$ be the Lie group exponential of $A^\times$ and consider the subset

$$\mathbf{L}^e(\mathscr{G}(\mathscr{H}, B)) := \{x \in \mathbf{L}(A^\times) = A \mid \exp_A(\mathbb{R}x) \subseteq \mathscr{G}(\mathscr{H}, B)\}.$$

We establish in Lemma 37 that $\mathfrak{g}(\mathscr{H}, B)$ is mapped by $\exp_A$ to $\mathscr{G}(\mathscr{H}, B)$, whence $\mathfrak{g}(\mathscr{H}, B) \subseteq \mathbf{L}^e(\mathscr{G}(\mathscr{H}, B))$. To see that $\mathbf{L}^e(\mathscr{G}(\mathscr{H}, B))$ only contains infinitesimal characters, recall from Lemma 37 that there is an open 0-neighborhood $\Omega \subseteq A$ such that $\exp_A$ maps $\mathfrak{g}(\mathscr{H}, B) \cap \Omega$ bijectively to $\exp_A(\Omega) \cap \mathscr{G}(\mathscr{H}, B)$. If $x \in \mathbf{L}^e(\mathscr{G}(\mathscr{H}, B))$ then we can pick $t > 0$ so small that $tx \in \Omega$. By definition of $\mathbf{L}^e(\mathscr{G}(\mathscr{H}, B))$ we see that then $\exp_A(tx) \in \mathscr{G}(\mathscr{H}, B) \cap \exp_A(\Omega)$. Therefore, we must have

$tx \in \Omega \cap \mathfrak{g}(\mathscr{H}, B)$, whence $x \in \mathfrak{g}(\mathscr{H}, B)$. This entails that $\mathbf{L}^e(\mathscr{G}(\mathscr{H}, B)) = \mathfrak{g}(\mathscr{H}, B)$ and then [43, Theorem IV.3.3] implies that $\mathscr{G}(\mathscr{H}, B)$ is a locally exponential closed Lie subgroup of $(A^\times, \star)$ whose Lie algebra is the Lie algebra of infinitesimal characters $\mathfrak{g}(\mathscr{H}, B)$. Moreover, since $(A^\times, \star)$ is a BCH–Lie group, so is $\mathscr{G}(H, B)$ (cf. [43, Definition IV.1.9]). □

Note that the Lie group $\mathscr{G}(\mathscr{H}, B)$ constructed in Theorem 15 will be modelled on a Fréchet space if $\mathscr{H}$ is of countable dimension (e.g. if $\mathscr{H}$ is of finite type) and B is in addition a Fréchet algebra. If $\mathscr{H}$ is even finite-dimensional and B is a Banach algebra, then $\mathscr{G}(\mathscr{H}, B)$ will even be modelled on a Banach space.

Example 16

1. The characters of a Hopf algebra of finite-type, i.e. the components $\mathscr{H}_n$ in the grading $\mathscr{H} = \oplus_{n\in\mathbb{N}_0}\mathscr{H}_n$ are finite-dimensional, are infinite-dimensional Lie groups by Theorem 15. Most natural examples of Hopf algebras appearing in combinatorial contexts are of finite-type.
2. Every finite-dimensional Hopf algebra $\mathscr{H}$ can be endowed with the trivial grading $\mathscr{H}_0 := \mathscr{H}$. Thus Theorem 15 implies that character groups (with values in a commutative CIA) of finite-dimensional Hopf algebras (cf. [3] for a survey) are infinite-dimensional Lie groups.
3. Graded and connected Hopf algebras (see next section) appear in the Connes-Kreimer theory of perturbative renormalisation of quantum field theories. However, recently in [32] it has been argued that instead of the graded and connected Hopf algebra of Feynman graphs considered traditionally (see e.g. the exposition in [14]) a non connected extension of this Hopf algebra should be used. The generalisation of the Hopf algebra then turns out to be a Hopf algebra with $\dim \mathscr{H}_0 < \infty$, whence its character groups with values in a Banach algebra turn out to be infinite-dimensional Lie groups.

These classes of examples could in general not be treated by the methods developed in [5].

Remark 17 Recall that by definition of a graded Hopf algebra $\mathscr{H} = \bigoplus_{n\in\mathbb{N}_0}\mathscr{H}_n$, the Hopf algebra structure turns $\mathscr{H}_0$ into a sub-Hopf algebra. Therefore, we always obtain two Lie groups $\mathscr{G}(\mathscr{H}, B)$ and $\mathscr{G}(\mathscr{H}_0, B)$ if $\dim \mathscr{H}_0 < \infty$. It is easy to see that the restriction map $q : \mathscr{G}(\mathscr{H}, B) \to \mathscr{G}(\mathscr{H}_0, B)$, $\phi \mapsto \phi|_{\mathscr{H}_0}$ is a morphism of Lie groups with respect to the Lie group structures from Theorem 15. Its kernel is the normal subgroup

$$\ker q = \{\phi \in \mathscr{G}(\mathscr{H}, B) \mid \phi|_{\mathscr{H}_0} = \mathbb{1}|_{\mathscr{H}_0}\},$$

i.e. the group of characters which coincide with the unit on degree 0.

We now turn to the question whether the Lie group constructed in Theorem 15 is a *regular* Lie group.

Theorem 18 *Let $\mathscr{H}$ be a graded Hopf algebra with $\dim \mathscr{H}_0 < \infty$ and B be a Banach algebra. Then the Lie group $\mathscr{G}(\mathscr{H}, B)$ is C^0-regular and the associated evolution map is even a $\mathbb{K}$-analytic map.*

Proof In Theorem 15 the Lie group structure on the character group was identified as a closed subgroup of the Lie group $A^\times$. By definition of the character group (cf. Definition (2)), $\mathscr{G}(\mathscr{H}, B)$ can be described as

$$\mathscr{G}(\mathscr{H}, B) = \{\phi \in A^\times \mid \phi \circ m_{\mathscr{H}} = m_B \circ (\phi \otimes \phi)\}$$

As discussed in Remark 36 Equation (3), the map $(m_{\mathscr{H}})^* \colon A^\times \to A^\times_\otimes, \phi \mapsto \phi \circ m_{\mathscr{H}}$ is a Lie group morphism. Now we consider the map $\theta \colon A \to A_\otimes, \phi \mapsto m_B \circ (\phi \otimes \phi)$. Observe that $\theta = \beta \circ \mathrm{diag}$, where β is the continuous bilinear map (4) and $\mathrm{diag} \colon A \to A \times A, \phi \mapsto (\phi, \phi)$. Since β is smooth (as a continuous bilinear map), and the diagonal map is clearly smooth, θ is smooth. Further, (5) shows that $\theta(\phi \star \psi) = \theta(\phi) \star_{At} \theta(\psi)$. Thus θ restricts to a Lie group morphism $\theta_{A^\times} \colon A^\times \to At^\times$.

Summing up, we see that $\mathscr{G}(\mathscr{H}, B) = \{\phi \in A^\times \mid (m_{\mathscr{H}})^*(\phi) = \theta_{A^\times}(\phi)\}$. Since by Proposition 11 the Lie group $A^\times$ is C^0-regular, the closed subgroup $\mathscr{G}(\mathscr{H}, B)$ is also C^0-regular by [20, Theorem G]. □

Remark 19 In Theorem 18 we have established that the character group $\mathscr{G}(\mathscr{H}, B)$ inherits the regularity properties of the ambient group of units. If the Hopf algebra $\mathscr{H}$ is in addition of countable dimension, e.g. a Hopf algebra of finite type, then Lemma 14 asserts that the ambient group of units is even L^1-regular. Unfortunately, Theorem [20, Theorem G] only deals with regularity of type C^k for $k \in \mathbb{N}_0 \cup \{\infty\}$. However, since $\mathscr{G}(\mathscr{H}, B)$ is a closed Lie subgroup of $\mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)^\times$, it is easy to see that the proof of [20, Theorem G] carries over without any changes to the L^1-case.[8] Hence, we can adapt the proof of Theorem 18 to obtain the following statement:

Corollary *Let $\mathscr{H}$ be a graded Hopf algebra of countable dimension with $\dim \mathscr{H}_0 < \infty$ and B be a Banach algebra. Then the Lie group $\mathscr{G}(\mathscr{H}, B)$ is L^1-regular.*

Note that the results on L^1-regularity of character groups considerably strengthen the results which have been available for regularity of character groups (see [5]).

[8] One only has to observe that a function into the Lie subgroup is absolutely continuous if and only if it is absolutely continuous as a function into the larger group. On the author's request, a suitable version of [20, Theorem G] for L^1-regularity will be made available in a future version of [19].

2.2 Character Groups for a Graded and Connected Hopf Algebra $\mathscr{H}$ and B a Locally Convex Algebra

In many interesting cases, the Hopf algebra is even connected graded (i.e. $\mathscr{H}_0$ is one-dimensional). In this case, we can weaken the assumption on B to be only locally convex.

Theorem 20 ([5, Theorem 2.7]) *Let $\mathscr{H}$ be a graded and connected Hopf algebra and B be a locally convex algebra. Then the manifold structure induced by the global parametrisation* $\exp\colon \mathfrak{g}(\mathscr{H}, B) \to \mathscr{G}(\mathscr{H}, B)$, $x \mapsto \sum_{n\in\mathbb{N}_0} \frac{x^{\star n}}{n!}$ *turns $\mathscr{G}(\mathscr{H}, B)$ into a $\mathbb{K}$-analytic Lie group.*

The Lie algebra associated to this Lie group is $\mathfrak{g}(\mathscr{H}, B)$. Moreover, the Lie group exponential map is given by the real analytic diffeomorphism exp.

Remark 21 Note that Theorem 20 yields more information for the graded and connected case then the new results discussed in Sect. 2.1: In the graded and connected case, the Lie group exponential is a global diffeomorphism, whereas the theorem for the non-connected case only establishes that exp induces a diffeomorphism around the unit.

Note that the connectedness of the Hopf algebra is the important requirement here. Indeed, we can drop the assumption of an integer grading and generalise to the grading by a suitable monoid.

Definition 22 Let M be a submonoid of $(\mathbb{R}, +)$ with $0 \in M$. We call M an *index monoid* if every initial segment $I_m := \{n \in M \mid n \leq m\}$ is finite. Note that this forces the monoid to be at most countable.

As in the $\mathbb{N}_0$ graded case, we say a Hopf algebra $\mathscr{H}$ is *graded by an index monoid* M, if $\mathscr{H} = \bigoplus_{m\in M} \mathscr{H}_m$ and the algebra, coalgebra and antipode respect the grading in the usual sense.

Example 23 The monoid $(\mathbb{N}_0, +)$ is an index monoid.

A source for (more interesting) index monoids is Hairer's theory of regularity structures for locally subcritical semilinear stochastic partial differential equations [25, Section 8.1] (cf. in particular [25, Lemma 8.10] and see Example 25 below).

Note that the crucial property of an index monoid is the finiteness of initial segments. This property allows one to define the functional calculus used in the proof of [5, Theorem B] to this slightly more general setting. Changing only trivial details in the proof of loc.cit., one immediately deduces that the statement of Theorem 20 generalises from an $\mathbb{N}_0$-grading to the more general situation

Corollary 24 *Let $\mathscr{H} = \bigoplus_{m\in M} \mathscr{H}_m$ be a graded and connected Hopf algebra graded by an index monoid M, B a sequentially complete locally convex algebra. Then the manifold structure induced by* $\exp\colon \mathfrak{g}(\mathscr{H}, B) \to \mathscr{G}(\mathscr{H}, B)$, $x \mapsto \sum_{n\in\mathbb{N}_0} \frac{x^{\star n}}{n!}$ *turns $\mathscr{G}(\mathscr{H}, B)$ into a $\mathbb{K}$-analytic Lie group. This Lie group is C^0-regular, its Lie algebra is $\mathfrak{g}(\mathscr{H}, B)$ and the Lie group exponential map is the real analytic diffeomorphism* exp.

2.2.1 Application to Character Groups in the Renormalisation of SPDEs

Hopf algebras graded by index monoids appear in Hairer's theory of regularity structures for locally subcritical semilinear SPDEs [25, Section 8.1]. Character groups of these Hopf algebras (and their quotients) then appear as structure group in Hairer's theory (cf. [25, Theorem 8.24] and [11]) of regularity structures. Recall that a regularity structure (A, T, G) in the sense of Hairer (cf. [25, Definition 2.1]) is given by

- an index set $A \subseteq \mathbb{R}$ with $0 \in A$, which is bounded from below and locally finite,
- a *model space* $T = \bigoplus_{\alpha \in A} T_\alpha$ which is a graded vector space with $T_0 \cong \mathbb{R}$ (denote by 1 its unit element) and T_α a Banach space for every $\alpha \in A$,
- a *structure group* G of linear operators acting on T such that, for every $\Gamma \in G$, $\alpha \in A$ and $a \in T_\alpha$ one has

$$\Gamma a - a \in \bigoplus_{\beta < \alpha} T_\beta \text{ and } \Gamma 1 = 1.$$

We sketch now briefly the theory developed in [11], where for a class of examples singular SPDEs the structure group was recovered as the character group of a connected Hopf algebra graded by an index monoid.

Example 25 The construction outlined in [11] builds first a general Hopf algebra of decorated trees in the category of bigraded spaces. Note that this bigrading does not induce a suitable $\mathbb{N}_0$-grading (or grading by index monoid) for our purposes. This Hopf algebra encodes the combinatorics of Taylor expansions in the SPDE setting and it needs to be tailored to the specific SPDE. This is achieved by choosing another ingredient, a so called *subcritical and complete normal rule*, i.e. a device depending on the SDE in question which selects a certain sub-Hopf algebra (see [11, Section 5] for details). Basically, the rule collects all admissible terms (= abstract decorated trees) which appear in the local Taylor expansion of the singular SPDE.[9]

Using the rule, we can select an algebra of decorated trees $\mathscr{T}^{ex}_+$ admissible with respect to the rule. Here the "+" denotes that we only select trees which are positive with respect to a certain grading $|\cdot|_+$ (cf. [11, Remark 5.3 and Definition 5.29]). Then [11, Proposition 5.34] shows that $(\mathscr{T}^{ex}_+, |\cdot|_+)$ is a graded and connected Hopf algebra of decorated trees. Note however, that the grading $|\cdot|_+$ is in general not integer valued, i.e. $\mathscr{T}^{ex}_+$ is graded by a submonoid M of $[0, \infty[$.

Since we are working with a normal rule which is complete and subcritical, the submonoid M satisfies $|\{\gamma \in M \mid \gamma < c\}| < \infty$ for each $c \in \mathbb{R}$, i.e. M is an index monoid. The reason for this is that by construction $\mathscr{T}^{ex}_+$ is generated by tree products of trees which strongly conform to the rule [11, Lemma 5.25 and Definition 5.29]. As the rule is complete and subcritical, there are only finitely many trees τ with $|\tau|_+ < c$ (for $c \in \mathbb{R}$) which strongly conform to the rule. Now the tree product is

[9] See [11, Section 5.5] for some explicit examples of this procedure, e.g. for the KPZ equation.

the Hopf algebra product (i.e. the product respects the grading), whence the property for M follows.

Now by [11, Proposition 5.39] the graded space $\mathcal{T}^{ex} = (\langle B_0 \rangle, |\cdot|_+)$ (suitably generated by a certain subset of the strongly conforming trees) together with the index set $A^{ex} = \{|\tau|_+ \mid \tau \in B_0\}$ and the character group $\mathcal{G}_+^{ex} := \mathcal{G}(\mathcal{T}_+^{ex}, \mathbb{R})$ form a regularity structure $(A^{ex}, \mathcal{T}^{ex}, \mathcal{G}_+^{ex})$. In conclusion, Corollary 24 yields an infinite-dimensional Lie group structure for the structure group $\mathcal{G}_+^{ex}$ of certain subcritical singular SPDEs.

Remark 26 In the full renormalisation procedure outlined in [11] two groups are involved in the renormalisation. Apart from the structure group outlined above, also a so called *renormalisation group* $\mathcal{G}_-^{ex}$ is used (cf. [11, Section 5]). This group $\mathcal{G}_-^{ex}$ is (in the locally subcritical case) a finite-dimensional group arising as the character group of another Hopf algebra. However, it turns out that the actions of $\mathcal{G}_+^{ex}$ and $\mathcal{G}_-^{ex}$ interact in an interesting way (induced by a cointeraction of underlying Hopf algebras, think semidirect product). We hope to return to these actions and use the (infinite-dimensional) Lie group structure to study them in future work.

3 Appendix: Infinite-Dimensional Calculus

In this section basic facts on the differential calculus in infinite-dimensional spaces are recalled. The general setting for our calculus are locally convex spaces (see [30, 44]).

Definition 27 Let E be a topological vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. E is called *locally convex space* if there is a family $\{p_i \mid i \in I\}$ of continuous seminorms for some index set I, such that

(i) the topology is the initial topology with respect to $\{\mathrm{pr}_{p_i} : E \to E_{p_i} \mid i \in I\}$, i.e. the E-valued map f is continuous if and only if $\mathrm{pr}_i \circ f$ is continuous for each $i \in I$, where $E_{p_i} := E/p_i^{-1}(0)$ is the normed space associated to the p_i and $\mathrm{pr}_i : E \to E_{p_i}$ is the canonical projection,
(ii) if $x \in E$ with $p_i(x) = 0$ for all $i \in I$, then $x = 0$ (i.e. E is Hausdorff).

Many familiar results from finite-dimensional calculus carry over to infinite dimensions if we assume that all spaces are locally convex.

As we are working beyond the realm of Banach spaces, the usual notion of Fréchet-differentiability cannot be used.[10] Moreover, there are several inequivalent notions of differentiability on locally convex spaces (see again [31]).

[10]The problem here is that the bounded linear operators do not admit a good topological structure if the spaces are not normable. In particular, the chain rule will not hold for Fréchet-differentiability in general for these spaces (cf. [31]).

We base our investigation on the so called Bastiani calculus, [2]. The notion of differentiability we adopt is natural and quite simple, as the derivative is defined via directional derivatives.

Definition 28 Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $r \in \mathbb{N} \cup \{\infty\}$ and E, F locally convex $\mathbb{K}$-vector spaces and $U \subseteq E$ open. Moreover we let $f\colon U \to F$ be a map. If it exists, we define for $(x, h) \in U \times E$ the directional derivative

$$df(x,h) := D_h f(x) := \lim_{\mathbb{K}^\times \ni t \to 0} t^{-1}(f(x+th) - f(x)).$$

We say that f is $C^r_{\mathbb{K}}$ if the iterated directional derivatives

$$d^{(k)} f(x, y_1, \dots, y_k) := (D_{y_k} D_{y_{k-1}} \cdots D_{y_1} f)(x)$$

exist for all $k \in \mathbb{N}_0$ such that $k \leq r$, $x \in U$ and $y_1, \dots, y_k \in E$ and define continuous maps $d^{(k)} f\colon U \times E^k \to F$. If it is clear which $\mathbb{K}$ is meant, we simply write C^r for $C^r_{\mathbb{K}}$. If f is $C^\infty_{\mathbb{C}}$, we say that f is *holomorphic* and if f is $C^\infty_{\mathbb{R}}$ we say that f is *smooth*.

For more information on our setting of differential calculus we refer the reader to [17, 31]. Another popular choice for infinite-dimensional calculus is the so called "convenient setting" of global analysis outlined in [34]. On Fréchet spaces (i.e. complete metrisable locally convex spaces) our notion of differentiability coincides with differentiability in the sense of convenient calculus. Note that differentiable maps in our setting are continuous by default (which is in general not true in the convenient setting). We encounter analytic mappings between infinite-dimensional spaces, as a preparation for this, note first:

Remark 29 A map $f\colon U \to F$ is of class $C^\infty_{\mathbb{C}}$ if and only if it is *complex analytic* i.e., if f is continuous and locally given by a series of continuous homogeneous polynomials (cf. [4, Proposition 7.4 and 7.7]). We then also say that f is of class $C^\omega_{\mathbb{C}}$.

To introduce real analyticity, we have to generalise a suitable characterisation from the finite-dimensional case: A map $\mathbb{R} \to \mathbb{R}$ is real analytic if it extends to a complex analytic map $\mathbb{C} \supseteq U \to \mathbb{C}$ on an open $\mathbb{R}$-neighbourhood U in $\mathbb{C}$. We proceed analogously for locally convex spaces by replacing $\mathbb{C}$ with a suitable complexification.

Definition 30 (Complexification of a locally convex space) Let E be a real locally convex topological vector space. Endow $E_{\mathbb{C}} := E \times E$ with the following operation

$$(x + iy).(u, v) := (xu - yv, xv + yu) \quad \text{for } x, y \in \mathbb{R}, u, v \in E.$$

The complex vector space $E_{\mathbb{C}}$ with the product topology is called the *complexification* of E. We identify E with the closed real subspace $E \times \{0\}$ of $E_{\mathbb{C}}$.

Definition 31 Let E, F be real locally convex spaces and $f\colon U \to F$ defined on an open subset U. Following [40] and [17], we call f *real analytic* (or $C^\omega_{\mathbb{R}}$) if f extends to a $C^\infty_{\mathbb{C}}$-map $\tilde{f}\colon \tilde{U} \to F_{\mathbb{C}}$ on an open neighbourhood $\tilde{U}$ of U in the complexification $E_{\mathbb{C}}$.[11]

Note that many of the usual results of differential calculus carry over to our setting. In particular, maps on connected domains whose derivative vanishes are constant as a version of the fundamental theorem of calculus holds. Moreover, the chain rule holds in the following form:

Lemma 32 (Chain Rule [17, Propositions 1.12, 1.15, 2.7 and 2.9]) *Fix $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$ and $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ together with $C^k_{\mathbb{K}}$-maps $f\colon E \supseteq U \to F$ and $g\colon H \supseteq V \to E$ defined on open subsets of locally convex spaces. Assume that $g(V) \subseteq U$. Then $f \circ g$ is of class $C^k_{\mathbb{K}}$ and the first derivative of $f \circ g$ is given by*

$$d(f \circ g)(x; v) = df(g(x); dg(x, v)) \quad \text{for all } x \in U,\ v \in H.$$

The differential calculus developed so far extends easily to maps which are defined on non-open sets. This situation occurs frequently in the context of differential equations on closed intervals (see [1] for an overview).

Having the chain rule at our disposal we can define manifolds and related constructions which are modelled on locally convex spaces.

Definition 33 Fix a Hausdorff topological space M and a locally convex space E over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. An ($E$-)manifold chart (U_κ, κ) on M is an open set $U_\kappa \subseteq M$ together with a homeomorphism $\kappa\colon U_\kappa \to V_\kappa \subseteq E$ onto an open subset of E. Two such charts are called C^r-compatible for $r \in \mathbb{N}_0 \cup \{\infty, \omega\}$ if the change of charts map $\nu^{-1} \circ \kappa\colon \kappa(U_\kappa \cap U_\nu) \to \nu(U_\kappa \cap U_\nu)$ is a C^r-diffeomorphism. A C^r-atlas of M is a set of pairwise C^r-compatible manifold charts, whose domains cover M. Two such C^r-atlases are equivalent if their union is again a C^r-atlas.

A *locally convex C^r-manifold* M modelled on E is a Hausdorff space M with an equivalence class of C^r-atlases of (E-)manifold charts.

Direct products of locally convex manifolds, tangent spaces and tangent bundles as well as C^r-maps of manifolds may be defined as in the finite dimensional setting (see [43, I.3]). The advantage of this construction is that we can now give a very simple answer to the question, what an infinite-dimensional Lie group is:

Definition 34 A (locally convex) *Lie group* is a group G equipped with a $C^\infty_{\mathbb{K}}$-manifold structure modelled on a locally convex space, such that the group operations are smooth. If the manifold structure and the group operations are in addition ($\mathbb{K}$-) analytic, then G is called a ($\mathbb{K}$-) *analytic* Lie group.

[11] If E and F are Fréchet spaces, real analytic maps in the sense just defined coincide with maps which are continuous and can be locally expanded into a power series. See [18, Proposition 4.1].

We recommend [43] for a survey on the theory of locally convex Lie groups. However, the Lie groups constructed in this article have strong structural properties as they belong to the class of Baker–Campbell–Hausdorff–Lie groups.

Definition 35 (Baker–Campbell–Hausdorff (BCH-)Lie groups and Lie algebras)

1. A Lie algebra $\mathfrak{g}$ is called *Baker–Campbell–Hausdorff–Lie algebra* (BCH–Lie algebra) if there exists an open 0-neighbourhood $U \subseteq \mathfrak{g}$ such that for $x, y \in U$ the *BCH-series* $\sum_{n=1}^{\infty} H_n(x, y)$ converges and defines an analytic map $U \times U \to \mathfrak{g}$. The H_n are defined as $H_1(x, y) = x + y$, $H_2(x, y) = \frac{1}{2}[x, y]$ and for $n \geq 3$ by linear combinations of iterated brackets, see [43, Definition IV.1.5.] or [9, Chapter 2, §6].
2. A locally convex Lie group G is called *BCH–Lie group* if it satisfies one of the following equivalent conditions (cf. [43, Theorem IV.1.8])
   (i) G is a $\mathbb{K}$-analytic Lie group whose Lie group exponential function is $\mathbb{K}$-analytic and a local diffeomorphism in 0.
   (ii) The exponential map of G is a local diffeomorphism in 0 and $\mathbf{L}(G)$ is a BCH–Lie algebra.

BCH–Lie groups share many of the structural properties of Banach Lie groups while not necessarily being Banach Lie groups themselves.

4 Appendix: Characters and the Exponential Map

Fix for the rest of this section a $\mathbb{K}$-Hopf algebra $\mathscr{H} = (\mathscr{H}, m_{\mathscr{H}}, u_{\mathscr{H}}, \Delta_{\mathscr{H}}, \varepsilon_{\mathscr{H}}, S_{\mathscr{H}})$ and a commutative continuous inverse algebra B. Furthermore, we assume that the Hopf algebra $\mathscr{H}$ is graded, i.e. $\mathscr{H} = \bigoplus_{n\in\mathbb{N}_0} \mathscr{H}$ and $\dim \mathscr{H}_0 < \infty$. The aim of this section is to prove that the Lie group exponential map $\exp_{A^\times}$ of $A := \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B)$ restricts to a bijection from the infinitesimal characters to the characters.

Remark 36 (Cocomposition with Hopf multiplication) Let $\mathscr{H} \otimes \mathscr{H}$ be the tensor Hopf algebra (cf. [35, p. 8]). We regard the tensor product $\mathscr{H} \otimes \mathscr{H}$ as a graded and connected Hopf algebra with the tensor grading, i.e. $\mathscr{H} \otimes \mathscr{H} = \bigoplus_{n\in\mathbb{N}_0} (\mathscr{H} \otimes \mathscr{H})_n$ where for all $n \in \mathbb{N}_0$ the nth degree is defined as $(\mathscr{H} \otimes \mathscr{H})_n = \bigoplus_{i+j=n} \mathscr{H}_i \otimes \mathscr{H}_j$.

Since $\dim \mathscr{H}_0 < \infty$ we see that $\dim (\mathscr{H} \otimes \mathscr{H})_0 = \dim \mathscr{H}_0 \otimes \mathscr{H}_0 < \infty$ Thus with respect to the topology of pointwise convergence and the convolution product, the algebras

$$A := \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B) \qquad A_{\otimes} := \mathrm{Hom}_{\mathbb{K}}(\mathscr{H} \otimes \mathscr{H}, B)$$

become continuous inverse algebras (see Definition 6 and Lemma 10). This structure turns

$$\cdot \circ m_{\mathscr{H}} \colon \mathrm{Hom}_{\mathbb{K}}(\mathscr{H}, B \to \mathrm{Hom}_{\mathbb{K}}(\mathscr{H} \otimes \mathscr{H}, B), \quad \phi \mapsto \phi \circ m_{\mathscr{H}}$$

into a continuous algebra homomorphism. Hence its restriction

$$(m_{\mathscr{H}})^*\colon A^\times \to A^\times_\otimes, \quad \phi \mapsto \phi \circ m_{\mathscr{H}} \tag{3}$$

is a Lie group morphism with $\mathbf{L}((m_{\mathscr{H}})^*) := T_e(m_{\mathscr{H}})^* = \cdot \circ m_{\mathscr{H}}$.

Lemma 37 *The Lie group exponential* $\exp_A\colon \mathbf{L}(A^\times) = A \to A^\times,\ x \mapsto \sum_{n=0}^\infty \frac{x^{\star n}}{n!}$ *maps* $\mathfrak{g}(\mathscr{H}, B)$ *to* $\mathscr{G}(\mathscr{H}, B)$. *Further, there is a* 0*-neighborhood* $\Omega \subseteq A$ *such that* $\exp_A$ *maps* $\mathfrak{g}(\mathscr{H}, B) \cap \Omega$ *bijectively onto* $\mathscr{G}(\mathscr{H}, B) \cap \exp_A(\Omega)$. [12]

Proof We denote by $\star_{A_\otimes}$ the convolution product of the CIA $A_\otimes$ and let $\exp_{A_\otimes}$ be the Lie group exponential of $A^\times_\otimes$. Let $m_B\colon B \otimes B \to B,\ b_1 \otimes b_2 \mapsto b_1 \cdot b_2$ be the multiplication in B. Define the continuous bilinear map (cf. [5, proof of Lemma B.10] for detailed arguments)

$$\beta\colon A \times A \to A_\otimes, \quad (\phi, \psi) \mapsto \phi \diamond \psi := m_B \circ (\phi \otimes \psi). \tag{4}$$

We may use β to rewrite the convolution in A as $\star_A = \beta \circ \Delta$ and obtain

$$(\phi_1 \diamond \psi_1) \star_{A_\otimes} (\phi_2 \diamond \psi_2) = (\phi_1 \star_A \phi_2) \diamond (\psi_1 \star_A \psi_2). \tag{5}$$

Recall, that $1_A := u_B \circ \varepsilon_{\mathscr{H}}$ is the neutral element of the algebra A. From equation (5), it follows at once, that the continuous linear maps

$$\beta(\cdot, 1_A)\colon A \to A_\otimes, \quad \phi \mapsto \phi \diamond 1_A \quad \text{and} \quad \beta(1_A, \cdot)\colon A \to A_\otimes, \quad \phi \mapsto 1_A \diamond \phi$$

are algebra homomorphisms which restrict to Lie group morphisms

$$\beta^\rho\colon A^\times \to A^\times_\otimes, \quad \phi \mapsto \beta(\phi, 1_A) \text{ and } \beta^\lambda\colon A^\times \to A^\times_\otimes, \quad \phi \mapsto \beta(1_A, \phi) \tag{6}$$

with $\mathbf{L}(\beta^\rho) = \beta(\cdot, 1_A)$ and $\mathbf{L}(\beta^\lambda) = \beta(1_A, \cdot)$. Let $\phi \in A$ be given and recall from (5) that $(\phi \diamond 1_A) \star_{A_\otimes} (1_A \diamond \phi) = \phi \diamond \phi = (1_A \diamond \phi) \star_{A_\otimes} (\phi \diamond 1_A)$. As a consequence we obtain

$$\exp_{A_\otimes}(\phi \diamond 1_A + 1_A \diamond \phi) = \exp_{A_\otimes}(\phi \diamond 1_A) \star_{A_\otimes} \exp_{A_\otimes}(1_A \diamond \phi). \tag{7}$$

since every Lie group exponential function transforms addition into multiplication for commuting elements.

[12] Note that apart from the locallity and several key arguments, the proof follows the general idea of the similar statement [5, Lemma B.10]. For the readers convenience we repeat the arguments to exhibit how properties of the Lie group exponential replace the functional calculus used in [5].

Note that it suffices to check multiplicativity of $\exp_A(\phi)$ as $\exp_A(\phi)(1_{\mathscr{H}}) = 1_B$ is automatically satisfied. For an infinitesimal character $\phi \in A$ we have by definition $\phi \circ m_{\mathscr{H}} = \phi \diamond 1_A + 1_A \diamond \phi$. Using now the naturality of the Lie group exponentials (i.e. for a Lie group morphism $f\colon G \to H$ we have $\exp_H \circ \mathbf{L}(f) = f \circ \exp_G$), we derive the following:

$$
\begin{aligned}
\phi \in \mathfrak{g}(\mathscr{H}, B) &\overset{\text{Def}}{\iff} \phi \circ m_{\mathscr{H}} = \phi \diamond 1_A + 1_A \diamond \phi \\
&\overset{(!)}{\implies} \exp_{A_\otimes}(\phi \circ m_{\mathscr{H}}) = \exp_{A_\otimes}(\phi \diamond 1_A + 1_A \diamond \phi) \\
&\overset{(7)}{\iff} \exp_{A_\otimes}(\phi \circ m_{\mathscr{H}}) = \exp_{A_\otimes}(\phi \diamond 1_A) \star_{A_\otimes} \exp_{A_\otimes}(1_A \diamond \phi) \\
&\overset{(6)}{\iff} \exp_{A_\otimes}(\phi \circ m_{\mathscr{H}}) = \big(\exp_A(\phi) \diamond 1_A\big) \star_{A_\otimes} \big(1_A \diamond \exp_A(\phi)\big) \\
&\overset{(5)}{\iff} \exp_{A_\otimes}(\phi \circ m_{\mathscr{H}}) = \big(\exp_A(\phi) \star_A 1_A\big) \diamond \big(1_A \star_A \exp_A(\phi)\big) \\
&\iff \exp_{A_\otimes}(\phi \circ m_{\mathscr{H}}) = \exp_A(\phi) \diamond \exp_A(\phi) \\
&\overset{(3)}{\iff} \exp_A(\phi) \circ m_{\mathscr{H}} = \exp_A(\phi) \diamond \exp_A(\phi) \\
&\overset{\text{Def}}{\iff} \exp_A(\phi) \in \mathscr{G}(\mathscr{H}, B).
\end{aligned}
$$

This shows that infinitesimal characters are mapped by the Lie group exponential to elements in the character group.

Now we observe that in general the implication from the first to the second line will not be an equivalence (as the Lie group exponential is not a global diffeomorphism unlike the connected Hopf algebra case discussed in [5, Lemma B.10]). We exploit now that $A^\times$ and $A_\otimes^\times$ are locally exponential Lie groups, whence locally around 0 in A and $A_\otimes$ the Lie group exponentials induce diffeomorphisms. Hence there are open neighborhoods of 0 and the units of $A^\times$ and $A_\otimes^\times$, such that $\exp_{A^\times}\colon A \supseteq V \to W \subseteq A^\times$ and $\exp_{A_\otimes^\times}\colon A_\otimes \supseteq V_\otimes \to W_\otimes \subseteq A_\otimes^\times$ are diffeomorphisms. Since $A_\otimes$ is a locally convex space, there is an open 0-neighborhood $\Omega_\otimes \subseteq A_\otimes$ such that $\Omega_\otimes + \Omega_\otimes \subseteq V_\otimes$. By continuity of β and $\cdot \otimes m_{\mathscr{H}}$, we obtain an open 0-neighborhood

$$
\Omega := V \cap (\cdot \otimes m_{\mathscr{H}})^{-1}(V_\otimes) \cap \beta(\cdot, 1_A)^{-1}(\Omega_\otimes) \cap \beta(1_A, \cdot)^{-1}(\Omega_\otimes) \subseteq A.
$$

Now by construction elements in $\mathfrak{g}(\mathscr{H}, B) \cap \Omega$ are mapped by $\cdot \circ m_{\mathscr{H}}$ into $V_\times$ and by $\beta(\cdot, 1_A) + \beta(1_A, \cdot)$ into $\Omega_\otimes + \Omega_\otimes \subseteq V_\otimes$. Since $\exp_{A_\otimes}$ induces a diffeomorphism on $V_\otimes$, the implication (!) becomes an equivalence for elements in $\mathfrak{g}(\mathscr{H}, B) \cap \Omega$. We have thus established that $\exp_A$ maps $\mathfrak{g}(\mathscr{H}, B) \cap \Omega$ bijectively to $\exp_A(\Omega) \cap \mathscr{G}(\mathscr{H}, B)$. □

Acknowledgements This research was partially supported by the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691070 and by the Knut and Alice Wallenberg Foundation grant agreement KAW 2014.0354. We are indebted to K.-H. Neeb and R. Dahmen for discussions which led to Lemma 10. Further, we would like to thank L. Zambotti and Y. Bruned for explaining their results on character groups in the renormalisation of SPDEs. Finally, we thank K.H. Hofmann for encouraging and useful comments and apologize to him for leaving out [28] at first.

References

1. Alzaareer, H., Schmeding, A.: Differentiable mappings on products with different degrees of differentiability in the two factors. Expo. Math. **33**(2), 184–222 (2015). https://doi.org/10.1016/j.exmath.2014.07.002
2. Bastiani, A.: Applications différentiables et variétés différentiables de dimension infinie. J. Anal. Math. **13**, 1–114 (1964)
3. Beattie, M.: A survey of Hopf algebras of low dimension. Acta Appl. Math. **108**(1), 19–31 (2009). https://doi.org/10.1007/s10440-008-9367-3
4. Bertram, W., Glöckner, H., Neeb, K.H.: Differential calculus over general base fields and rings. Expo. Math. **22**(3), 213–282 (2004). https://doi.org/10.1016/S0723-0869(04)80006-9
5. Bogfjellmo, G., Dahmen, R., Schmeding, A.: Character groups of Hopf algebras as infinite-dimensional Lie groups. Ann. Inst. Fourier (Grenoble) **66**(5), 2101–2155 (2016)
6. Bogfjellmo, G., Dahmen, R., Schmeding, A.: Overview of (pro-)Lie group structures on Hopf algebra character groups. In: Ebrahimi-Fard, K., Barbero Linan, M. (eds.) Discrete Mechanics, Geometric Integration and Lie-Butcher Series. Springer Proceedings in Mathematics and Statistics, vol. 267, pp. 287–314. Springer, Cham (2018)
7. Bogfjellmo, G., Schmeding, A.: The tame Butcher group. J. Lie Theor. **26**, 1107–1144 (2016)
8. Bogfjellmo, G., Schmeding, A.: The Lie group structure of the Butcher group. Found. Comput. Math. **17**(1), 127–159 (2017). https://doi.org/10.1007/s10208-015-9285-5
9. Bourbaki, N.: Lie groups and Lie algebras. Chapters 1–3. Elements of Mathematics (Berlin). Springer, Berlin (1998). Translated from the French, Reprint of the 1989 English translation
10. Brouder, C.: Trees, renormalization and differential equations. BIT Num. Anal. **44**, 425–438 (2004)
11. Bruned, Y., Hairer, M., Zambotti, L.: Algebraic Renormalisation of Regularity Structures (2016). http://arxiv.org/abs/1610.08468v1
12. Butcher, J.C.: An algebraic theory of integration methods. Math. Comput. **26**, 79–106 (1972)
13. Cartier, P.: A primer of Hopf algebras. In: Frontiers in Number Theory, Physics, and Geometry, vol. II, pp. 537–615. Springer, Berlin (2007)
14. Connes, A., Marcolli, M.: Noncommutative Geometry, Quantum Fields and Motives, American Mathematical Society Colloquium Publications, vol. 55. American Mathematical Society/Hindustan Book Agency, Providence/New Delhi (2008)
15. Floret, K.: Lokalkonvexe Sequenzen mit kompakten Abbildungen. J. Reine Angew. Math. **247**, 155–195 (1971)
16. Glöckner, H.: Algebras whose groups of units are Lie groups. Stud. Math. **153**(2), 147–177 (2002). http://dx.doi.org/10.4064/sm153-2-4
17. Glöckner, H.: Infinite-dimensional Lie groups without completeness restrictions. In: Geometry and Analysis on Finite- and Infinite-Dimensional Lie Groups (Będlewo, 2000), Banach Center Publication, vol. 55, pp. 43–59. Institute of Mathematics of the Polish Academy of Sciences, Warsaw (2002). https://doi.org/10.4064/bc55-0-3
18. Glöckner, H.: Instructive examples of smooth, complex differentiable and complex analytic mappings into locally convex spaces. J. Math. Kyoto Univ. **47**(3), 631–642 (2007). http://dx.doi.org/10.1215/kjm/1250281028

19. Glöckner, H.: Measurable Regularity Properties of Infinite-Dimensional Lie Groups (2015). http://arxiv.org/abs/1601.02568v1
20. Glöckner, H.: Regularity Properties of Infinite-Dimensional Lie Groups, and Semiregularity (2015). http://arxiv.org/abs/1208.0715v3
21. Glöckner, H., Neeb, K.H.: When unit groups of continuous inverse algebras are regular Lie groups. Stud. Math. **211**(2), 95–109 (2012). http://dx.doi.org/10.4064/sm211-2-1
22. Glöckner, H., Neeb, K.H.: Infinite-dimensional Lie Groups. General Theory and Main Examples (2018). Unpublished
23. Gracia-Bondí a, J.M., Várilly, J.C., Figueroa, H.: Elements of Noncommutative Geometry. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser Boston, Inc., Boston (2001) https://doi.org/10.1007/978-1-4612-0005-5
24. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Springer Series in Computational Mathematics, vol. 31. Springer, New York (2006)
25. Hairer, M.: A theory of regularity structures. Invent. Math. **198**(2), 269–504 (2014). http://dx.doi.org/10.1007/s00222-014-0505-4
26. Hewitt, E., Ross, K.A.: Abstract Harmonic Analysis, vol. II: Structure and Analysis for Compact Groups. Analysis on Locally Compact Abelian Groups. Die Grundlehren der mathematischen Wissenschaften, Band 152. Springer, New York/Berlin (1970)
27. Hofmann, K.H., Morris, S.A.: The Lie theory of connected pro-Lie groups. EMS Tracts in Mathematics, vol. 2. EMS, Zürich (2007). https://doi.org/10.4171/032
28. Hofmann, K.H., Morris, S.A.: The Structure of Compact Groups. De Gruyter Studies in Mathematics, vol. 25. De Gruyter, Berlin (2013). https://doi.org/10.1515/9783110296792. A primer for the student—a handbook for the expert, Third edition, revised and augmented
29. Hofmann, K.H., Morris, S.A.: Pro-Lie groups: A survey with open problems. Axioms **4**, 294–312 (2015). https://doi.org/10.3390/axioms4030294
30. Jarchow, H.: Locally Convex Spaces. B.G. Teubner, Stuttgart (1981). Mathematische Leitfäden. [Mathematical Textbooks]
31. Keller, H.: Differential Calculus in Locally Convex Spaces. Lecture Notes in Mathematics 417. Springer, Berlin (1974)
32. Kock, J.: Perturbative renormalisation for not-quite-connected bialgebras. Lett. Math. Phys. **105**(10), 1413–1425 (2015). https://doi.org/10.1007/s11005-015-0785-7
33. König, W.: The Parabolic Anderson Model. Pathways in Mathematics. Birkhäuser/Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33596-4. Random walk in random potential
34. Kriegl, A., Michor, P.W.: The Convenient Setting of Global Analysis. Mathematical Surveys and Monographs, vol. 53. AMS, Providence (1997)
35. Majid, S.: Foundations of Quantum Group Theory. Cambridge University Press, Cambridge (1995). https://doi.org/10.1017/CBO9780511613104
36. Mallios, A.: Topological Algebras. Selected Topics. North-Holland Mathematics Studies, vol. 124. North-Holland, Amsterdam (1986). Notas de Matemática [Mathematical Notes], 109
37. Manchon, D.: Hopf algebras in renormalisation. In: Handbook of Algebra, vol. 5, pp. 365–427. Elsevier/North-Holland, Amsterdam (2008). https://doi.org/10.1016/S1570-7954(07)05007-3
38. McLachlan, R.I., Modin, K., Munthe-Kaas, H., Verdier, O.: B-series methods are exactly the affine equivariant methods. Numer. Math. **133**(3), 599–622 (2016). http://dx.doi.org/10.1007/s00211-015-0753-2
39. Michaelis, W.: Coassociative coalgebras. In: Handbook of Algebra, vol. 3, pp. 587–788. North-Holland, Amsterdam (2003). http://dx.doi.org/10.1016/S1570-7954(03)80072-4
40. Milnor, J.: Remarks on infinite-dimensional Lie groups. In: Relativity, Groups and Topology, II (Les Houches, 1983), pp. 1007–1057. North-Holland, Amsterdam (1984)
41. Milnor, J.W., Moore, J.C.: On the structure of Hopf algebras. Ann. Math. (2) **81**, 211–264 (1965). http://dx.doi.org/10.2307/1970615
42. Murua, A., Sanz-Serna, J.M.: Computing normal forms and formal invariants of dynamical systems by means of word series. Nonlinear Anal. **138**, 326–345 (2016). http://dx.doi.org/10.1016/j.na.2015.10.013

43. Neeb, K.H.: Towards a Lie theory of locally convex groups. Japan J. Math. **1**(2), 291–468 (2006). https://doi.org/10.1007/s11537-006-0606-y
44. Schaefer, H.H.: Topological Vector Spaces. Springer, New York/Berlin (1971). Third printing corrected, Graduate Texts in Mathematics, vol. 3
45. Swan, R.G.: Topological examples of projective modules. Trans. Am. Math. Soc. **230**, 201–234 (1977). http://dx.doi.org/10.2307/1997717
46. Sweedler, M.E.: Hopf Algebras. Mathematics Lecture Note Series. W. A. Benjamin, Inc., New York (1969)

Shape Analysis on Homogeneous Spaces: A Generalised SRVT Framework

Elena Celledoni, Sølve Eidnes, and Alexander Schmeding

Abstract Shape analysis is ubiquitous in problems of pattern and object recognition and has developed considerably in the last decade. The use of shapes is natural in applications where one wants to compare curves independently of their parametrisation. One computationally efficient approach to shape analysis is based on the Square Root Velocity Transform (SRVT). In this paper we propose a generalised SRVT framework for shapes on homogeneous manifolds. The method opens up for a variety of possibilities based on different choices of Lie group action and giving rise to different Riemannian metrics.

1 Shapes on Homogeneous Manifolds

Shapes are unparametrised curves, evolving on a vector space, on a Lie group or on a manifold. Shape spaces and spaces of curves are infinite dimensional Riemannian manifolds, whose Riemannian metrics are the essential tool to compare and analyse shapes. By combining infinite dimensional differential geometry, analysis and computational mathematics, shape analysis provides a powerful approach to a variety of applications.

In this paper, we are concerned with the approach to shape analysis based on the Square Root Velocity Transform (SRVT), [27]. This method is effective and computationally efficient. On vector spaces, the SRVT maps parametrised curves to appropriately scaled tangent vector fields along them. The transformed curves are compared computing geodesics in the L^2 metric, and the scaling can be chosen suitably to yield reparametrisation invariance, [6, 27]. Notably, applying a (reparametrisation invariant) L^2 metric directly on the original parametrised curves is not an option as it leads to vanishing geodesic distance on parametrised curves and on the quotient shape space [4, 21]. As an alternative, higher order Sobolev type

E. Celledoni (✉) · S. Eidnes · A. Schmeding
NTNU Trondheim, Institutt for matematiske fag, Trondheim, Norway
e-mail: elena.celledoni@ntnu.no; solve.eidnes@ntnu.no; schmeding@tu-berlin.de

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_7

metrics were proposed [22], even though they can be computationally demanding, since computing geodesics in this infinite dimensional Riemannian setting amounts in general to solving numerically partial differential equations. These geodesics are used in practice for finding distances between curves and for interpolation between curves. The SRVT approach, on the other hand, is quite practical because it allows the use of the L^2 metric on the transformed curves: distances between curves are just L^2 distances of the transformed curves, and geodesics between curves are "*straight lines*" between the transformed curves. It is also possible to prove that this algorithmic approach corresponds (at least locally) to a particular Sobolev type metric, see [6, 9].

In the present paper we propose a generalisation of the SRVT, from vector spaces and Lie groups, [6, 27], to homogeneous manifolds. This problem has been previously considered for manifold valued curves in [18, 29], but our approach is different, the main idea is to take advantage of the Lie group acting transitively on the homogeneous manifold. The Lie group action allows us to transport derivatives of curves to our choice of base point in the homogeneous manifold. Then this information is lifted to a curve in the Lie algebra. It is natural to require that the lifted curve does not depend on the representative of the class used to pull back the curve to the base point.

The main contribution of this paper is the definition of a generalised square root velocity transform framework using transitive Lie group actions for curves on homogeneous spaces. Different choices of Lie group actions will give rise to different metrics on the infinite dimensional manifold of curves on the homogeneous space, with different properties. These different metrics, their geodesics and associated geometric tools for shape analysis can all be implemented in the computationally advantageous SRVT framework.

We extend previous results for Lie group valued curves and shapes [9], to the homogeneous manifold setting. Using ideas from the literature on differential equations on manifolds [10], we describe the main tools necessary for the definition of the SRVT and discuss the minimal requirements guaranteeing that the SRVT is well defined, Sect. 2. On a general homogeneous manifold, the SRVT is obtained using a right inverse of the composition of the Lie group action with the evolution operator of the Lie group. If the homogeneous manifold is reductive, there is an explicit way to construct this right inverse (based on a canonical 1-form for the reductive space, cf. 3.3–3.4), see also [20]. We prove smoothness of the defined SRVT in Sect. 2.1. Detailed examples on matrix Lie groups are provided in Sect. 4.

A Riemannian metric on the manifold of curves on the homogeneous space is obtained by pulling back the L^2 inner product of curves on the Lie algebra through the SRVT, Theorem 11. To ensure that the distance function obtained on the space of parametrised curves descends to a distance function on the shape space, it is necessary to prove equivariance with respect to the group of orientation preserving diffeomorphisms (reparametrization invariance), these results are presented in Sect. 2.3.

For the case of reductive homogeneous spaces, fixed the Lie group action, two different approaches are considered: one obtained pulling back the curves to the Lie

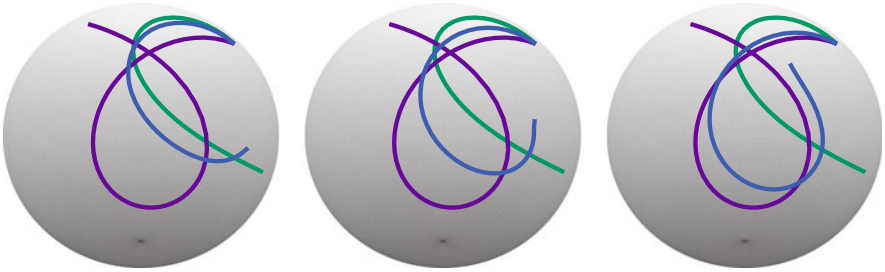


Fig. 1 The blue curve shows the deformation of the green curve into the purple one along a geodesic $\gamma : [0, 1] \to \mathrm{Imm}(I, S^2)$ plotted for the three times $\left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$ from left to right

algebra $\mathfrak{g}$ (Proposition 18) and one obtained pulling back the curves to the reductive subspace $\mathfrak{m} \subset \mathfrak{g}$ (Sect. 3.1.1). The resulting distances are both reparametrization invariant, see Lemmata 13 and 22. For the second approach it follows similarly to what shown in [9] that the geodesic distance is globally defined by the L^2 distance, Proposition 20. We conjecture that also for general homogeneous manifolds, at least locally, the geodesic distance of the pullback metric is given by the L^2 distance of the curves transformed by the SRVT, see end of Sect. 2.2. To illustrate the performance of the proposed approaches we compute geodesics between curves on the 2-sphere (viewed as a homogeneous space with respect to the canonical SO(3)-action), see Fig. 1 for an example. Numerical experiments show that the two algorithms perform differently when applied to curves on the sphere (Sect. 5).

This work appeared on the arXiv on the 5th of April 2017, later a related but different work from colleagues at Florida State University was completed and posted on the arXiv on the 9th of June 2017. The latter work has now appeared in [26], see also the follow up [25]. Moreover, loc.cit. treats quotients by compact subgroups focuses on the existence of optimal reparametrisations.

1.1 Preliminaries and Notation

Fix a Lie group G with identity element e and Lie algebra $\mathfrak{g}$.[1] Denote by $R_g : G \to G$ and $L_g : G \to G$ the right resp. left multiplication by $g \in G$. Let H be a closed Lie subgroup of G and $\mathcal{M} := G/H$ the quotient with the manifold structure turning $\pi : G \to G/H, g \mapsto gH$ into a submersion (see [12, Theorem G (b)]). Then $\mathcal{M}$ becomes a homogeneous space for G with respect to the (transitive) left action:

$$\Lambda : G \times \mathcal{M} \to \mathcal{M}, \quad (g, kH) \mapsto (gk)H.$$

[1] In this paper we assume all Lie groups and Lie algebras to be finite dimensional. Note however, that many of our techniques carry over to Lie groups modelled on Hilbert spaces, [9].

For $c_0 \in \mathcal{M}$ we write $\Lambda(g, c_0) = \Lambda_{c_0}(g) = g.c_0 = \Lambda^g(c_0)$, i.e. $\Lambda_{c_0} : G \to \mathcal{M}$ (the orbit map of the orbit through c_0) and $\Lambda^g : \mathcal{M} \to \mathcal{M}$.

1.1.1 We will consider smooth curves on $\mathcal{M}$ and describe them using the Lie group action. Namely for $c\colon [0,1] \to \mathcal{M}$ we choose a smooth lift $g\colon [0,1] \to G$ of c, i.e.:

$$c(t) = g(t).c_0, \quad c_0 \in \mathcal{M}, \quad t \in [0,1] \quad \text{(the dot denotes the action of } G \text{ on } G/H).$$

In general, there are many different choices for a smooth lifts g.[2] For brevity we will in the following write $I := [0,1]$.

Later on we consider smooth functions on infinite-dimensional manifolds beyond the realm of Banach manifolds. Hence the standard definition for smooth maps (i.e. the derivative as a (continuous) map to a space of continuous operators) breaks down. We base our investigation on the so called Bastiani calculus (see [3]): A map $f\colon E \supseteq U \to F$ between Fréchet spaces is smooth if all iterated directional derivatives exist and glue together to continuous maps.[3]

1.1.2 Let M be a (possibly infinite-dimensional) manifold. By $C^\infty(I, M)$ we denote smooth functions from I to M. Recall that the topology on these spaces, the compact-open C^∞-topology, allows one to control a function and its derivatives. This topology turns $C^\infty(I, M)$ into an infinite-dimensional manifold (see e.g. [15, Section 42]).

Denote by $\mathrm{Imm}(I, M) \subseteq C^\infty(I, M)$ the set of smooth immersions (i.e. smooth curves $c\colon I \to M$ with $\dot{c}(t) \neq 0$) and recall from [15, 41.10] that $\mathrm{Imm}(I, M)$ is an open subset of $C^\infty(I, M)$.

1.1.3 We further denote by Evol the evolution operator, which is defined as

$$\mathrm{Evol} : C^\infty(I, \mathfrak{g}) \to \{g \in C^\infty(I, G) : g(0) = e\} =: C^\infty_*(I, G)$$

$$\mathrm{Evol}(q)(t) := g(t) \qquad \text{where} \qquad \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} g = R_{g(t)*}(q(t)), \\ g(0) = e \end{cases}$$

[2] Every homogeneous space G/H is a principal H-bundle, whence there are smooth horizontal lifts of smooth curves (depending on some choice of connection, cf. e.g. [23, Chapter 5.1]).

[3] In the setting of manifolds on Fréchet spaces (with which we deal here) our setting of calculus is equivalent to the so called convenient calculus (see [15]). Convenient calculus defines a map f to be smooth if it "maps smooth curves to smooth curves", i.e. $f \circ c$ is smooth for any smooth curve c. This yields a calculus on infinite-dimensional spaces where smoothness does not necessarily imply continuity (though this does not happen on Fréchet spaces), we refer to [15] for a detailed exposition. Note that both calculi can handle smooth maps on intervals $[a, b]$, see e.g. [13, 1.1] and [15, Chapter 24].

and $R_{g*} = T_e R_g$ is the tangent of the right translation. Recall from [13, Theorem A] that Evol is a diffeomorphism with inverse the *right logarithmic derivative*

$$\delta^r : C^\infty_*(I, G) \to C^\infty(I, \mathfrak{g}), \quad \delta^r g := R^{-1}_{g^*}(\dot{g}).$$

1.1.4 We fix a Riemannian metric $(\langle \cdot, \cdot \rangle_g)_{g \in G}$ on G which is right H-invariant (i.e. the maps $R_h, h \in H$ are Riemannian isometries). Since $\mathcal{M} = G/H$ is constructed using the right H-action on G, an H-right invariant metric descends to a Riemannian metric on $\mathcal{M}$. We refer to [11, Proposition 2.28] for details and will always endow the quotient with this canonical metric to relate the Riemannian geometries.

Hence H-right invariance should be seen as a minimal requirement for the metric on G. Note that a natural way to obtain (right) invariant metrics is to transport a Hilbert space inner product from the Lie algebra by (right) translation in the group. This method yields a G-right invariant metric and we will usually work with such a metric induced by $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}$. Albeit it is very natural, G-invariance does not immediately add any benefits. In the following table we record properties of H, the Riemannian metric and of the canonical G-action on the quotient.

1.1.5 Let $f : M \to N$ be a smooth map and denote postcomposition by

$$\theta_f : C^\infty(I, M) \to C^\infty(I, N), \qquad c \mapsto f \circ c.$$

Note that θ_f is smooth as a map between (infinite-dimensional) manifolds.

1.1.6 (The SRVT on Lie groups) For a Lie group G with Lie algebra $\mathfrak{g}$, consider an immersion $c : I \to G$. The square root velocity transform of c is

$$\mathcal{R} : \mathrm{Imm}(I, G) \to C^\infty(I, \mathfrak{g} \setminus \{0\}), \qquad \mathcal{R}(c) = \frac{\delta^r(c)}{\sqrt{\|\dot{c}\|}} = \frac{\left(R^{-1}_{c(t)}\right)_*(\dot{c})}{\sqrt{\|\dot{c}\|}}, \tag{1}$$

where the norm $\|\cdot\|$ is induced by a right G-invariant Riemannian metric, [9]. The SRVT consists of the composition of three maps:

- *differentiation* $D : C^\infty(I, G) \to C^\infty(I, TG)$, $D(c) := \dot{c}$,
- *transport* $\alpha : C^\infty(I, TG) \to C^\infty(I, \mathfrak{g}), \quad \gamma \mapsto (R^{-1}_{\pi_{TG} \circ \gamma})_*(\gamma)$ and
- *scaling* $\mathrm{sc} : C^\infty(I, \mathfrak{g} \setminus \{0\}) \to C^\infty(I, \mathfrak{g} \setminus \{0\}), \quad q \mapsto \left(t \mapsto \frac{q(t)}{\sqrt{\|q(t)\|}}\right)$.

The scaling by the square root of the norm of the velocity is crucial to obtain a parametrisation invariant Riemannian metric, see [9] and Lemma 13.

2 Definition of the SRVT for Homogeneous Manifolds

Our aim is to construct the SRVT for curves with values in the homogeneous manifold $\mathcal{M}$. It was crucial in our investigation of the Lie group case [9] that the right-logarithmic derivative inverts the evolution operator, see 1.1.3. To mimic this behaviour we introduce a version of the evolution for homogeneous manifolds.

Definition 1 Fix $c_0 \in \mathcal{M}$ and denote by $C^\infty_{c_0}(I, \mathcal{M})$ all smooth curves $c\colon I \to \mathcal{M}$ with $c(0) = c_0$. Then we define

$$\rho_{c_0}\colon C^\infty(I, \mathfrak{g}) \to C^\infty_{c_0}(I, \mathcal{M}), \quad \rho_{c_0}(q) = \Lambda_{c_0}(\mathrm{Evol}(q)(t)) = \Lambda(\mathrm{Evol}(q)(t), c_0).$$

Remark 2 Fix $q \in C^\infty(I, \mathfrak{g})$ and $c_0 \in \mathcal{M}$ and denote by $g(t) = \mathrm{Evol}(q)(t)$. Then

$$\rho_{c_0}(q) := c(t) \qquad \text{where} \qquad \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}c(t) = T_e\Lambda_{c(t)}(q(t)), \\ c(0) \ = c_0. \end{cases}$$

Proof In fact

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{c_0}(q)(t) = T_{g(t)}\Lambda_{c_0}\left(\frac{\mathrm{d}}{\mathrm{d}t}g(t)\right) = T_{g(t)}\Lambda_{c_0}((R_{g(t)})_*(q(t))) = T_{g(t)}\Lambda_{c_0} \circ (R_{g(t)})_*(q(t))$$

$$= T_e(\Lambda_{c_0} \circ R_{g(t)})(q(t)) = T_e\Lambda_{\Lambda_{c_0}(g(t))}(q(t)) = T_e\Lambda_{\rho_{c_0}(q)(t)}(q(t)),$$

with $T_{g(t)}\Lambda_{c_0}\colon T_{g(t)}G \to T_{\Lambda_{c_0}(g(t))}\mathcal{M} = T_{\rho_{c_0}(q)(t)}\mathcal{M}$, $\qquad T_e\Lambda_{c(t)}\colon \mathfrak{g} \to T_{c(t)}\mathcal{M}$. □

Hence we can interpret ρ_{c_0} as a version of the evolution operator Evol for homogeneous manifolds.

Example 3 Consider the two dimensional unit sphere $\mathcal{M} = S^2$ in $\mathbb{R}^3$. Consider the action of SO(3) on S^2 by matrix-vector multiplication: $\Lambda : \mathrm{SO}(3) \times S^2 \to S^2$, $\Lambda(Q, u) = Q \cdot u$. Assume $c_0 := e_1$ the first canonical vector in $\mathbb{R}^3$, then given a curve in the Lie algebra of skew-symmetric matrices $q(t) \in \mathfrak{so}(3)$, $\rho_{e_1}(q(t)) = y(t)$, where $y(t)$ satisfies $\dot{y} = q(t)y$ with $y(0) = e_1$.

We want to construct a section of the submersion ρ_{c_0} to mimic the construction for Lie groups, see also [10, Proposition 2.2]. As we have seen in the Lie group case, the SRVT factorises into a derivation map, a map transporting the derivative to the Lie algebra and a scaling in the Lie algebra. For homogeneous spaces, we can make sense of this procedure if we can replace the transport from the Lie group case by a map which transports derivatives from the tangent bundle of the homogeneous manifold to the Lie algebra. Thus we search for a map

$\alpha\colon C^\infty(I, T\mathcal{M}) \to C^\infty(I, \mathfrak{g})$ such that the following diagram commutes:

$$C^\infty_{c_0}(I,\mathcal{M}) \xrightarrow{D} C^\infty(I, T\mathcal{M}) \xrightarrow{\alpha} C^\infty(I,\mathfrak{g}) \xrightarrow{\rho_{c_0}} C^\infty_{c_0}(I,\mathcal{M}), \qquad \mathrm{id}_{C^\infty_{c_0}(I,\mathcal{M})}$$

Moreover, in the Lie group case we see that the mapping $\alpha \circ D$ maps the submanifold of immersions into the subset $C^\infty(I, \mathfrak{g} \setminus \{0\})$. We will require this property in general, as derivatives of immersions should vanish nowhere and this property should be preserved by the transport α. The next definition details necessary properties of α.

Definition 4 (Square root velocity transform) Let $c_0 \in \mathcal{M}$ be fixed and define the closed submanifold[4] $\mathcal{P}_{c_0} := \{c \in \mathrm{Imm}(I, \mathcal{M}) \mid c(0) = c_0\} = \mathrm{Imm}(I, \mathcal{M}) \cap C^\infty_{c_0}(I, \mathcal{M})$ of $C^\infty(I, \mathcal{M})$. Assume there is a smooth $\alpha\colon C^\infty(I, T\mathcal{M}) \to C^\infty(I, \mathfrak{g})$, such that

$$\rho_{c_0} \circ \alpha \circ D = \mathrm{id}_{C^\infty_{c_0}(I,\mathcal{M})} \text{ and} \tag{2}$$

$$\alpha \circ D(\mathcal{P}_{c_0}) \subseteq C^\infty(I, \mathfrak{g} \setminus \{0\}). \tag{3}$$

Then we define the *square root velocity transform* on $\mathcal{M}$ at c_0, with respect to α as

$$\mathcal{R}\colon \mathcal{P}_{c_0} \to C^\infty(I, \mathfrak{g} \setminus \{0\}), \quad \mathcal{R}(c) := \frac{\alpha(\dot{c})}{\sqrt{\|\alpha(\dot{c})\|}},$$

where $\|\cdot\|$ is the norm induced by the right invariant Riemannian metric on the Lie algebra. We will see in Lemma 9 that $\mathcal{R}$ is smooth.

The SRVT allows us to transport curves (via α) from the homogeneous manifold to curves with values in a fixed vector space (i.e. the Lie algebra $\mathfrak{g}$). *The crucial property here is that $\alpha \circ D$ is a right-inverse of ρ_{c_0}*, and we note that our construction depends strongly on the choice of the map ρ_{c_0}.

Example 5 Let G be a Lie group and $H = \{e\}$ the trivial subgroup (with e the Lie group identity). Then $G = G/\{e\}$ is a homogeneous manifold and $\rho_e = \mathrm{Evol}$. Taking $\alpha(v) = (R_g^{-1})_*(v)$, we reproduce the definition of the SRVT on Lie groups 1.1.6. However, contrary to Evol, ρ_{c_0} is not invertible if the subgroup H (with $\mathcal{M} = G/H$) is non-trivial, but we might still be able to find a right inverse.

Example 6 We have $T_u S^2 := \{v \in \mathbb{R}^3 \mid v \cdot u = 0\}$ where we have denoted with "$\cdot$" the Euclidean inner product in $\mathbb{R}^3$. Then we can write

$$v = (vu^T - uv^T)u, \qquad \forall v \in T_u S^2$$

[4] As $\mathrm{Imm}(I, \mathcal{M}) \subseteq C^\infty(I, \mathcal{M})$ is open and the evaluation map $\mathrm{ev}_0\colon \mathrm{Imm}(I, \mathcal{M}) \to \mathcal{M}$ is a submersion, $\mathcal{P}_{c_0} = \mathrm{ev}_0^{-1}(c_0)$ is a closed submanifold of $\mathrm{Imm}(I, \mathcal{M})$ (cf. [12]).

and we can define the map

$$\alpha : v \in T_u S^2 \mapsto vu^T - uv^T \in \mathfrak{so}(3).$$

For c a curve evolving on S^2 with $c(0) = e_1$, we have $\rho_{e_1}(\alpha(\dot{c})) = c$, so $\alpha \circ D$ is the right inverse of ρ_{e_1}. The SRVT is then

$$\mathcal{R}(c) = \frac{\dot{c}c^T - c\dot{c}^T}{\sqrt{\|\dot{c}c^T - c\dot{c}^T\|}},$$

and $\|\cdot\|$ is the norm deduced by the usual Frobenius inner product of matrices (the scaled negative Killing form in $\mathfrak{so}(3)$ see table in example 16). See Sects. 4 and 5, for further details and more examples.

The definition of α and the SRVT in Definition 4 depend on the initial point $c_0 \in \mathcal{M}$. In many cases our choices of α satisfy (2) for every $c_0 \in \mathcal{M}$, i.e. α satisfies

$$\rho(c(0), \alpha(\dot{c})) := \rho_{c(0)}(\alpha(\dot{c})) = c \quad \text{for all } c \in C^\infty(I, \mathcal{M}).$$

Further, the SRVT also depends on the choice of the left-action $\Lambda\colon G \times \mathcal{M} \to \mathcal{M}$. A different action will yield a different SRVT. For example, there are several ways to interpret a Lie group as a homogeneous manifold with respect to different group actions. One of these recovers exactly the SRVT from [9] (see Example 5). See [20, Section 5.1] for more information on Lie groups as homogeneous spaces, e.g. by using the Cartan-Schouten action.

Remark 7 Fix $c \in C^\infty(I, \mathcal{M})$ to obtain a smooth map $\Lambda_c\colon C^\infty(I, G) \to C^\infty(I, \mathcal{M})$, $f \mapsto (t \mapsto \Lambda(f, c)(t))$ [19, Corollary 11.10 1. and Theorem 11.4]. Further we recall from [15, Theorem 42.17] that $C^\infty(I, T\mathcal{M}) \cong TC^\infty(I, \mathcal{M})$. Identifying the tangent space over the constant $e\colon I \to G$ (taking everything to the unit) we obtain

$$T_e\Lambda_c\colon C^\infty(I, \mathfrak{g}) \to T_cC^\infty(I, \mathcal{M}), \quad q \mapsto \left(t \mapsto T_e\Lambda_{c(t)}(q(t))\right).$$

If $T_e\Lambda_c$ was invertible (which it will not be in general), we could use it to define α.

2.1 Smoothness of the SRVT

One of the most important properties of the square root velocity transform is that it allows us to transport curves from the manifold to curves in the Lie algebra, and this operation is smooth and invertible. The details are summarised in the following two

lemmata. Following [9, Lemma 3.9], we consider the smooth scaling maps

$$\begin{aligned} \mathrm{sc}\colon C^\infty(I,\mathfrak{g}\setminus\{0\}) &\to C^\infty(I,\mathfrak{g}\setminus\{0\}), \quad q \mapsto \left(t \mapsto \frac{q(t)}{\sqrt{\|q(t)\|}}\right), \\ \mathrm{sc}^{-1}\colon C^\infty(I,\mathfrak{g}\setminus\{0\}) &\to C^\infty(I,\mathfrak{g}\setminus\{0\}), \quad q \mapsto (t \mapsto q(t)\|q(t)\|). \end{aligned} \tag{4}$$

Lemma 8 *Fix $c_0 \in \mathcal{M}$, then*

1. *$C^\infty_{c_0}(I,\mathcal{M})$ is a closed and split submanifold[5] of $C^\infty(I,\mathcal{M})$,*
2. *$\rho_{c_0}\colon C^\infty(I,\mathfrak{g}) \to C^\infty_{c_0}(I,\mathcal{M})$ is a smooth surjective submersion.*

Proof

1. Note that $C^\infty_{c_0}(I,\mathcal{M})$ is the preimage of c_0 under the evaluation map

$$\mathrm{ev}_0\colon C^\infty(I,\mathcal{M}) \to \mathcal{M}, \quad c \mapsto c(0).$$

   One can show, similarly to the proof of [9, Proposition 4.1] that ev_0 is a submersion. Hence, [12, Theorem C] implies that $C^\infty_{c_0}(I,\mathcal{M})$ is a closed submanifold of $C^\infty(I,\mathcal{M})$.
2. Recall that $\rho_{c_0} = \theta_{\Lambda_{c_0}} \circ \mathrm{Evol}$ with $\theta_{\Lambda_{c_0}}\colon C^\infty(I,G) \to C^\infty(I,\mathcal{M})$, $f \mapsto \Lambda_{c_0} \circ f$.
   As $\mathcal{M}$ is a homogeneous space, $\pi\colon G \to \mathcal{M}$ is a surjective submersion. Hence [23, Chapter 5.1] implies that $\theta_\pi\colon C^\infty(I,G) \to C^\infty(I,\mathcal{M})$ is surjective. Further, the Stacey-Roberts Lemma [2, Lemma 2.4] asserts that θ_π is a submersion. Picking $g \in \pi^{-1}(c_0)$, we can also write $\theta_{\Lambda_{c_0}}(f) = \pi \circ R_g \circ f = \theta_\pi(\theta_{R_g}(f))$. Thus $\theta_{\Lambda_{c_0}} = \theta_\pi \circ \theta_{R_g}$ is a surjective submersion and

$$\theta^{-1}_{\Lambda_{c_0}}(C^\infty_{c_0}(I,\mathcal{M})) = C^\infty_*(I,G) = \{c \in C^\infty(I,G) \mid c(0) = e\}.$$

   By [13, Theorem C], $\theta_{\Lambda_{c_0}}$ restricts to a smooth surjective submersion $C^\infty_*(I,G) \to C^\infty_{c_0}(I,\mathcal{M})$. Finally, since $\mathrm{Evol}\colon C^\infty(I,\mathfrak{g}) \to C^\infty_*(I,G)$ is a diffeomorphism (cf. 1.1.3), $\rho_{c_0} = \theta_{\Lambda_{c_0}} \circ \mathrm{Evol}$ is a smooth surjective submersion. □

Lemma 9 *Fix $c_0 \in \mathcal{M}$ and let α be as in Definition 4. Then the square root velocity transform $\mathcal{R} = \mathrm{sc} \circ \alpha \circ D$ constructed from α is a smooth immersion $\mathcal{R}\colon \mathcal{P}_{c_0} \to C^\infty(I,\mathfrak{g}\setminus\{0\})$.*

Proof The map $D\colon C^\infty(I,\mathcal{M}) \to C^\infty(I,T\mathcal{M})$, $c \mapsto \dot{c}$ is smooth by Lemma 25. Hence on $\mathcal{P}_{c_0}$, the restriction of D is smooth. As a composition of smooth maps, $\mathcal{R} = \mathrm{sc} \circ \alpha \circ D|_{\mathcal{P}_{c_0}}$ is also smooth.

[5] A submanifold N of a (possibly infinite-dimensional) manifold M is called *split* if it is modeled on a closed subvectorspace F of the model space E of M, such that F is complemented, i.e. $E = F \oplus G$ as topological vector spaces (see [12, Section 1]).

Since $\mathrm{sc}\colon C^\infty(I,\mathfrak{g}\setminus\{0\}) \to C^\infty(I,\mathfrak{g}\setminus\{0\})$ is a diffeomorphism, it suffices to prove that $\alpha \circ D|_{\mathcal{P}_{c_0}}$ is an immersion. As we are dealing with infinite-dimensional manifolds, it is not sufficient to prove that the derivative of $\alpha \circ D|_{\mathcal{P}_{c_0}}$ is injective (which is evident from (2)). Instead we have to construct immersion charts for $x \in \mathcal{P}_{c_0}$, i.e. charts in which $\alpha \circ D$ is conjugate to an inclusion of vector spaces.[6]

To construct these charts, recall from (2) that $f := \alpha \circ D|_{\mathcal{P}_{c_0}}$ is a right-inverse to ρ_{c_0}. In Lemma 8 we established that ρ_{c_0} is a surjective submersion which restricts to a submersion $\rho_{c_0}^{-1}(\mathcal{P}_{c_0}) \to \mathcal{P}_{c_0}$ by [12, Theorem C]. Fix $x \in \mathcal{P}_{c_0}$ and use the submersion charts for ρ_{c_0}. By [12, Lemma 1.2] there are open neighborhoods $x \in U_x \subseteq \mathcal{P}_{c_0}$ and $f(x) \in U_{f(x)} \subseteq \rho_{c_0}^{-1}(\mathcal{P}_{c_0})$ together with a smooth manifold N and a diffeomorphism $\theta\colon U_x \times N \to U_{f(x)}$ such that $\rho_{c_0} \circ \theta(u,n) = u$. Thus $\theta^{-1} \circ f|_{U_x} = (\mathrm{id}_{U_x}, f_2)$ for a smooth map $f_2\colon U_x \to U_{f_x}$. Hence $\theta^{-1} \circ f|_{U_x}$ induces a diffeomorphism onto the split submanifold $\Gamma(f_2) := \{(y, f_2(y)) \mid y \in U_x\} \subseteq U_x \times U_{f_x}$. Following [12, Lemma 1.13], we see that $f = \alpha \circ D|_{U_x}$ is an immersion. As x was arbitrary, the SRVT $\mathcal{R}$ is an immersion. □

Exploiting that $\mathcal{R}$ is an immersion, we transport Riemannian structures and distances from $C^\infty(I,\mathfrak{g}\setminus\{0\})$ to $\mathcal{P}_{c_0}$ by pullback. Note that the image of the SRVT for a homogeneous space is in general only an immersed submanifold of $C^\infty(I,\mathfrak{g}\setminus\{0\})$. For reductive homogeneous spaces, a certain SRVT will always yield a smooth embedding (see Lemma 19). We investigate now the Riemannian structure on $\mathcal{P}_{c_0}$.

2.2 The Riemannian Geometry of the SRVT

As a first step, we construct a Riemannian metric using the L^2 metric on $C^\infty(I,\mathfrak{g})$.

Definition 10 Endow $C^\infty(I,\mathfrak{g})$ with the L^2 inner product

$$\langle f, g\rangle_{L^2} = \int_0^1 \langle f(t), g(t)\rangle \mathrm{d}t,$$

where $\langle\cdot,\cdot\rangle$ is induced by the right H-invariant Riemannian metric of G on $\mathfrak{g}$.

The L^2 inner product induces a weak Riemannian metric. The L^2-geodesics are straight lines, i.e. a curve $c(t) \in C^\infty(I,\mathfrak{g})$ is a L^2-geodesic if and only if for every t, $s \mapsto c(t)(s)$ is a straight line in the vector space $\mathfrak{g}$. In Lemma 9 the square root velocity transform was identified as an immersion, which we now turn into a Riemannian immersion by pulling back the L^2 metric. Arguing as in the proof of [9, Theorem 3.11] one obtains the following formula for this pullback metric.

[6] See [12] for more information on immersions between infinite-dimensional manifolds.

Theorem 11 *Let $c \in \mathcal{P}_{c_0}$ and consider $v, w \in T_c\mathcal{P}_{c_0}$, i.e. $v, w \colon I \to T\mathcal{M}$ are curves with $v(t), w(t) \in T_{c(t)}\mathcal{M}$. The pullback of the L^2 metric on $C^\infty(I, \mathfrak{g} \setminus \{0\})$ under the SRVT to the manifold of immersions $\mathcal{P}_{c_0}$ is given by:*

$$G^{\mathcal{R}}_c(v, w) = \int_I \frac{1}{4} \langle D_s v, u_c \rangle \langle D_s w, u_c \rangle + \left\langle D_s v - u_c \langle D_s v, u_c \rangle , D_s w - u_c \langle D_s w, u_c \rangle \right\rangle \mathrm{d}s, \tag{5}$$

where $D_s v := T_c(\alpha \circ D)(v)/\|\alpha(\dot{c})\|$, $u_c := \alpha(\dot{c})/\|\alpha(\dot{c})\|$ is the (transported) unit tangent vector of c, and $\mathrm{d}s = \|\alpha(\dot{c}(t))\| \, \mathrm{d}t$. The pullback of the L^2 norm is given by

$$G^{\mathcal{R}}_c(v, v) = \int_I \frac{1}{4} \langle D_s v, u_c \rangle^2 + \left\| D_s v - u_c \langle D_s v, u_c \rangle \right\|^2 \mathrm{d}s.$$

The formula for the pullback metric in Theorem 11 depends on α and its derivative. However, notice that we always obtain a first order Sobolev metric which measures the derivative $D_s v$ of the vector field over a curve c.

The distance on $\mathcal{P}_{c_0}$ will now be defined as the geodesic distance of the first order Sobolev metric $G^{\mathcal{R}}$, i.e. of the pullback of an L^2 metric. Thus we just need to pull the L^2 geodesic distance on $\mathcal{R}(\mathcal{P}_{c_0})$ back using the SRVT. But, in general, the geodesic distance of two curves on the submanifold $\mathcal{R}(\mathcal{P}_{c_0})$ with respect to the L^2 metric will not be the L^2 distance of the curves (see e.g. [8, Section 2]). The question is now, under which conditions is the geodesic distance at least locally given by the L^2 distance. Note first that the image of the SRVT will in general not be an open submanifold of $C^\infty(I, \mathfrak{g})$ (this was the key argument to derive the geodesic distance in [9, Theorem 3.16]). As a consequence we were unable to derive a general result describing the links between the geodesic distance by $G^{\mathcal{R}}$ on $\mathcal{P}_{c_0}$ and the SRVT algorithmic approach for homogeneous manifolds. Nonetheless, we conjecture that at least locally the geodesic distance should be given by the L^2 distance (note that $\rho_{c_0}^{-1}(\mathcal{P}_{c_0})$ is an open set, whence the geodesic distance is locally given by the L^2 distance). On the other hand, for reductive homogeneous spaces (discussed in Sect. 3), an auxiliary map can be used to obtain a geodesic distance which globally coincides with the transformed L^2 distance.

2.3 Equivariance of the Riemannian Metric

Often in applications, one is interested in a metric on the shape space

$$\mathcal{S}_{c_0} := \mathcal{P}_{c_0} / \operatorname{Diff}^+(I) = \operatorname{Imm}_{c_0}(I, \mathcal{M}) / \operatorname{Diff}^+(I),$$

where $\mathrm{Diff}^+(I)$ is the group of orientation preserving diffeomorphisms of I acting on $\mathcal{P}_{c_0}$ from the right (cf. [5]). To assure that the distance function $d_{\mathcal{P}_{c_0}}$ descends to a distance function on the shape space, we need to require that it is invariant with respect to the group action.

Definition 12 Let $d\colon \mathcal{P}_{c_0} \times \mathcal{P}_{c_0} \to [0, \infty]$ be a metric. Then d is *reparametrisation invariant* if

$$d(f, h) = d(f \circ \varphi, g \circ \varphi) \quad \forall \varphi \in \mathrm{Diff}^+(I). \tag{6}$$

In other words d is invariant with respect to the diagonal (right) action of $\mathrm{Diff}^+(I)$ on $\mathcal{P}_{c_0} \times \mathcal{P}_{c_0}$.

Let $[f], [g] \in \mathcal{S}$ be equivalence classes and pick arbitrary representatives $f \in [f]$ and $g \in [g]$. If d is a reparametrisation invariant, we define a metric on $\mathcal{S}$ as

$$d_{\mathcal{S}}([f], [g]) := \inf_{\varphi \in \mathrm{Diff}^+(I)} d(f, g \circ \varphi). \tag{7}$$

Since d is reparametrisation invariant, the definition of $d_{\mathcal{S}}$ makes sense (cf. [9, Lemma 3.4]). To obtain a metric on $\mathcal{S}$, we need reparametrisation invairance of

$$d_{\mathcal{P}_{c_0}}\colon \mathcal{P}_{c_0} \times \mathcal{P}_{c_0} \to \mathcal{R}, \qquad d_{\mathcal{P}_{c_0}}(f, g) := \sqrt{\int_0^1 \left\| \mathcal{R}(f)(t) - \mathcal{R}(g)(t) \right\|^2 \mathrm{d}t}.$$

Lemma 13 *Let $\mathcal{R}$ be the square root velocity transform with respect to $c_0 \in \mathcal{M}$ and $\alpha\colon C^\infty(I, T\mathcal{M}) \cong TC^\infty(I, \mathcal{M}) \to C^\infty(I, \mathfrak{g})$. Then $d_{\mathcal{P}_{c_0}}$ is reparametrisation invariant if α is a $C^\infty(I, \mathfrak{g})$-valued 1-form on $C^\infty(I, \mathcal{M})$, e.g. if $\alpha = \theta_\omega$ for a $\mathfrak{g}$-valued 1-form on $\mathcal{M}$.*

Proof Consider $\varphi \in \mathrm{Diff}^+(I)$ and $f, g \in \mathcal{P}_{c_0}$. Then a computation yields

$$\mathcal{R}(f \circ \varphi) = \frac{\alpha(\dot{f} \circ \varphi \cdot \dot{\varphi})}{\sqrt{\left\| \alpha(\dot{f} \circ \varphi \cdot \dot{\varphi}) \right\|}} = \frac{\alpha(\dot{f} \circ \varphi) \cdot \dot{\varphi}}{\sqrt{\left\| \alpha(\dot{f} \circ \varphi) \cdot \dot{\varphi} \right\|}} = (\mathcal{R}(f) \circ \varphi) \cdot \sqrt{\dot{\varphi}},$$

where we have used that α is fibre-wise linear as a 1-form. Thus we can now compute

$$d_{\mathcal{P}_{c_0}}(f \circ \varphi, g \circ \varphi) = \sqrt{\int_I \left\| \mathcal{R}(f) \circ \varphi(t) - \mathcal{R}(g) \circ \varphi(t) \right\|^2 \dot{\varphi}(t) \mathrm{d}t} = d_{\mathcal{P}_{c_0}}(f, g).$$

□

The condition on α from Lemma 13 is satisfied in all examples of the SRVT considered in the present paper. For example, for a reductive homogeneous case (see Sect. 3), we can always choose α as the pushforward of a $\mathfrak{g}$-valued 1-form.

3 SRVT for Curves in Reductive Homogeneous Spaces

A fundamental problem in our approach to shape spaces with values in homogeneous spaces is that we need to somehow lift curves from the homogeneous space to the Lie group. Ideally, this lifting process should be compatible with the Riemannian metrics on the spaces. Note that for our purposes it suffices to lift the derivatives of smooth curves to curves in the Lie algebra of the Lie group. Hence we need a suitable Lie algebra valued 1-form, which turns out to exist for reductive homogeneous spaces, cf. e.g. [16, Chapter X] (see also [20] for a recent account)

3.1 Recall that $\mathrm{Ad}(g) := T_e\mathrm{conj}_g$, where $\mathrm{conj}_g = L_g \circ R_{g^{-1}}$ denotes conjugation $\mathrm{conj}_g \colon G \to G$. Suppose $\mathfrak{m}$ is a subspace of $\mathfrak{g}$ such that $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$.

Let $\omega_e \colon T_{eH}\mathcal{M} \to \mathfrak{m}$ be the inverse of $T_e\pi|_{\mathfrak{m}} \colon \mathfrak{g} \supseteq \mathfrak{m} \to T_{eH}\mathcal{M}$. Identify $\mathfrak{g} = T_eG$ and observe that $T_e\pi : \mathfrak{g} \to T_{eH}\mathcal{M}$ induces an isomorphism $T_e\pi|_{\mathfrak{m}} \colon \mathfrak{m} \to T_{eH}\mathcal{M}$.

By definition $\pi \circ R_h = \pi$ holds for all $h \in H$. Now the group actions of G on itself by left and right multiplication commute and we observe that

$$\text{for all } g \in G \quad \pi \circ L_g = \Lambda^g \circ \pi \quad \text{and} \quad T_e\pi \circ \mathrm{Ad}(h) = T\Lambda^h \circ T_e\pi \text{ for } h \in \mathfrak{h}. \tag{8}$$

3.2 We will from now on assume that $\mathcal{M}$ is a reductive homogeneous manifold. This means that the subalgebra $\mathfrak{h}$ admits a *reductive complement*, i.e. a vector subspace $\mathfrak{m} \subseteq \mathfrak{g}$ such that

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m} \text{ and } \mathrm{Ad}(h).\mathfrak{m} \subseteq \mathfrak{m} \text{ for all } h \in H.$$

If it exists, a reductive complement will in general not be unique. However, we choose and fix a reductive complement $\mathfrak{m}$ for $\mathfrak{h}$.

3.3 As a reductive complement, $\mathfrak{m}$ is closed with respect to the adjoint action of H. Hence one deduces (cf. [20, Lemma 4.6] for a proof) that ω_e is H-invariant with respect to the adjoint action, i.e.

$$\omega_e(T\Lambda^h(v)) = \mathrm{Ad}(h).\omega_e(v) \quad \text{for all } v \in T_{eH}\mathcal{M} \text{ and } h \in H.$$

Thus the following map is well-defined:

$$\omega \colon T\mathcal{M} \to \mathfrak{g}, \quad v \mapsto \mathrm{Ad}(g).\omega_e(T\Lambda^{g^{-1}}(v)) \quad \text{for all } v \in T_{gH}\mathcal{M}.$$

From the definition it is clear that ω is a smooth $\mathfrak{g}$-valued 1-form on $\mathcal{M}$. Moreover, ω is even G-equivariant with respect to the canonical and adjoint action:

$$\omega(T\Lambda^k(v)) = \mathrm{Ad}(k).\omega(v) \quad \text{for all } v \in T\mathcal{M} \text{ and } k \in G. \tag{9}$$

Note that ω depends by construction on our choice of reductive complement $\mathfrak{m}$. However, we will suppress this dependence in the notation. As noted in [20, Section 4.2], the 1-forms ω correspond bijectively to reductive structures on G/H.[7]

3.4 Let ω be the 1-form constructed in 3.3. Then we define the map

$$\theta_\omega \colon C^\infty(I, T\mathcal{M}) \to C^\infty(I, \mathfrak{g}), \quad f \mapsto \omega \circ f.$$

Note first that θ_ω is smooth by [15, Theorem 42.13]. We will prove that θ_ω indeed satisfies (2) and (3), whence $\alpha = \theta_\omega$ yields an SRVT as in 4.

To motivate the computations, let us investigate an important special case.

Example 14 Similarly to example 5, let G be a Banach Lie group and $H = \{e\}$ the trivial subgroup. Then $G = G/\{e\}$ can be viewed as a reductive homogeneous manifold with $\mathfrak{m} = \mathfrak{g}$, $\pi = \mathrm{id}_G$ and $\omega_e = \mathrm{id}_\mathfrak{g}$. From the definition of ω we obtain $\omega(v) = \mathrm{Ad}(g).(L^{g^{-1}})_*(v) = (R_g^{-1})_*(v) = \kappa^r(v)$, where κ^r denotes the right Maurer-Cartan form, [15, Section 38] or [20, Section 5.1]. In particular, for $c \colon I \to G$ we have $\theta(c) = \kappa^r(\dot{c}) = \delta^r(c)$ (right logarithmic derivative). As we have $\mathrm{Evol} \circ \delta^r(c) = c$ for a curve starting at e.

The SRVT for reductive spaces coincides thus with the SRVT for Lie group valued shape spaces as outlined in 1.1.6.

Albeit Example 14 is quite trivial as a homogeneous space, it highlights a general principle of the construction for reductive homogeneous spaces.

Remark 15 We here provide an alternative interpretation for $\theta_\omega \circ D$: A smooth curve $c \colon I \to \mathcal{M}$ admits a smooth horizontal lift $\tilde{c} \colon I \to G$ depending on a choice of connection for the principal bundle $G \to \mathcal{M}$ [23, Chapter 5.1]. For a reductive homogeneous manifold we construct a horizontal lift $\tilde{c}$ using the canonical invariant connection (depending on the reductive complement, see [16, X.2]). Now we take the (right) Darboux derivative (aka right logarithmic derivative) of $\tilde{c} \colon I \to G$ (see [24, 3.§5]). Then unraveling the definitions similar to Examples 5 and 14, one can show that $\delta^r(\tilde{c}) = \theta_\omega \circ D(c)$ holds for the 1-form θ_ω as in 3.4. Thus for a reductive homogeneous space the proposed SRVT can be viewed (up to scaling) as the Darboux derivative of a horizontal lift of a curve in $\mathcal{M}$. Note that this interpretation justifies again to view ρ_{c_0} as a generalised version of the evolution operator Evol (which inverts the right logarithmic derivative, see Remark 2).

A rich source for reductive homogeneous spaces are quotients of semisimple Lie groups. We recall now some of the main examples.

Example 16 Let G be a semisimple Lie group and H a Lie subgroup of G which is also semisimple. Then the homogeneous space $\mathcal{M} = G/H$ is reductive. A reductive

[7] Note that there might be different reductive structures on a homogeneous manifold. We refer to [20, Section 5.1] for examples and further references.

complement of $\mathfrak{h}$ in $\mathfrak{g}$ is the orthogonal complement $\mathfrak{h}^{\perp}$ with respect to the Cartan-Killing form on $\mathfrak{g}$ (recall that the Killing form of a semisimple Lie algebra is non-degenerate by Cartan's criterion [17, I.§7 Theorem 1.45]). For example, this occurs for $G = \mathrm{SL}(n)$ and $H = \mathrm{SL}(n-p)$ or $G = \mathrm{SO}(n)$ and $H = \mathrm{SO}(n-p)$ (where $1 \leq p < n$), since by [17, I.§8 and I.§18] the following properties hold:

Lie group G	Compact?	Semisimple?	Killing form $B(X, Y)$ on $\mathfrak{g}$
$\mathrm{SO}(n)$	Yes	Yes (for $n \geq 3$)	$(n-2)\mathrm{Tr}(XY)$
$\mathrm{SL}(n)$	No	Yes	$2n\mathrm{Tr}(XY)$
$\mathrm{GL}(n)$	No	No	$2n\mathrm{Tr}(XY) - 2\mathrm{Tr}(X)\mathrm{Tr}(Y)$

Here Tr denotes the trace of a matrix. All main examples in this paper are reductive.

Proposition 17 *Let $\mathcal{M} = G/H$ be a reductive homogeneous space, $c_0 \in \mathcal{M}$, ω and θ_ω as in 3.4. Consider $D\colon C^\infty_{c_0}(I, \mathcal{M}) \to C^\infty(I, T\mathcal{M})$, $c \mapsto \dot{c}$. Then*

$$\rho_{c_0} \circ \theta_\omega \circ D = \mathrm{id}_{C^\infty_{c_0}(I,\mathcal{M})}\,.$$

Proof As a shorthand write $\theta := \theta_\omega \circ D$. We establish in Lemma 26 the identity

$$\mathrm{id}_{C^\infty_{eH}(I,\mathcal{M})} = \rho_{eH} \circ \theta = \Lambda_{eH} \circ \mathrm{Evol} \circ \theta = \pi \circ \mathrm{Evol} \circ \theta\,. \tag{10}$$

Let now $c \in C^\infty_{c_0}(I, \mathcal{M})$ with $c_0 = g_0 H$. Then we obtain $\Lambda^{g_0^{-1}} \circ c \in C^\infty_{eH}(I, \mathcal{M})$ and

$$\begin{aligned}\rho_{c_0} \circ \theta(c) &= (\Lambda_{c_0} \circ \mathrm{Evol}) \circ \theta_\omega(\dot{c}) = \Lambda_{c_0} \circ \mathrm{Evol} \circ \omega(T\Lambda^{g_0} T\Lambda^{g_0^{-1}} \dot{c}) \\ &\overset{(9)}{=} \Lambda_{c_0} \circ \mathrm{Evol}(\mathrm{Ad}(g_0).\omega(T\Lambda^{g_0^{-1}} \dot{c})) = \Lambda_{c_0} \circ \mathrm{Evol}(\mathrm{Ad}(g_0).\theta(\Lambda^{g_0^{-1}} \circ c)).\end{aligned}$$

Recall from [13, 1.16] that for a Lie group morphism φ one has the identity $\mathrm{Evol} \circ \mathbf{L}(\varphi) = \varphi \circ \mathrm{Evol}$. By definition, $\mathrm{Ad}(g) = \mathbf{L}(\mathrm{conj}_g) := T_e \mathrm{conj}_g$, where $\mathrm{conj}_g = L_g \circ R_{g^{-1}}$ denotes the conjugation morphism. Insert this into the above equation:

$$\begin{aligned}\rho_{c_0} \circ \theta(c) &= \Lambda_{c_0} \circ \mathrm{Evol} \circ \theta(c) = \Lambda_{c_0} \circ L_{g_0} \circ R_{g_0^{-1}} \circ \mathrm{Evol}(\theta(\Lambda^{g_0^{-1}} \circ c)) \\ &= \pi \circ L_{g_0} \mathrm{Evol}(\theta(\Lambda^{g_0^{-1}} \circ c)) = \Lambda^{g_0} \circ \pi \circ \mathrm{Evol}(\theta(\Lambda^{g_0^{-1}} \circ c)) \\ &\overset{(10)}{=} \Lambda^{g_0} \circ \Lambda^{g_0^{-1}} \circ c = c.\end{aligned}$$

In passing to the second line we used that left and right multiplication maps commute and that $\Lambda_{c_0}(R_{g_0^{-1}}(k)) = \Lambda_{c_0}(kg_0^{-1}) = kg_0^{-1}c_0 = kg_0^{-1}g_0H = \pi(k)$. □

Proposition 18 *Let $\mathcal{M} = G/H$ be a reductive homogeneous space, $c_0 \in \mathcal{M}$, ω and θ_ω as in 3.4. Then θ_ω satisfies (2) and (3), whence for a reductive homogeneous space we can define the SRVT as*

$$\mathcal{R}(c) := \frac{\theta_\omega(\dot{c})}{\sqrt{\|\theta_\omega(\dot{c})\|}} \quad \text{for } c \in \operatorname{Imm}(I, \mathcal{M})$$

Proof In Proposition 17 we have already established (2). To see that (3) also holds for θ_ω, observe first that for $v \in T_{gH}\mathcal{M}$, we have $\omega(v) = \operatorname{Ad}(g).\omega_e(TA^{g^{-1}}(v))$. Since $\omega_e \circ TA^{g^{-1}} : T_{gH}\mathcal{M} \to \mathfrak{m}$ and $\operatorname{Ad}(g) : \mathfrak{g} \to \mathfrak{g}$ are linear isomorphisms, we see that $\omega(v) = 0$ if and only if $v = 0_{gH}$. As θ_ω is post-composition by ω, θ_ω satisfies (3). □

3.1 Riemannian Geometry and the Reductive SRVT

In the reductive space case, it is easier to describe the image of the square root velocity transform. It turns out that the image is a split submanifold with a global chart. Using this chart, we can also obtain information on the geodesic distance.

The idea is to transform the image of the SRVT such that it becomes $C^\infty(I, \mathfrak{m} \setminus \{0\})$, where $\mathfrak{m}$ is again the reductive complement. Pick $g_0 \in \pi^{-1}(c_0)$ and use the adjoint action of G and the evolution $\operatorname{Evol} : C^\infty(I, \mathfrak{g}) \to C^\infty(I, G)$ to define

$$\Psi_{g_0}(q) := -\operatorname{Ad}(g_0 \operatorname{Evol}(q)^{-1}).q \quad \text{for } q \in C^\infty(I, \mathfrak{g})$$

where the dot denotes pointwise application of the linear map $\operatorname{Ad}(\operatorname{Evol}(q)^{-1})$. Then Ψ_{g_0} is a diffeomorphism with inverse $\Psi_{g_0^{-1}}$ (see Lemma 28). We will now see that $\Psi_{g_0^{-1}}$ maps the image of the SRVT to $C^\infty(I, \mathfrak{m} \setminus \{0\})$.

Lemma 19 *Choose $c_0 \in \mathcal{M}$ in the reductive homogeneous space $\mathcal{M}$, and let ω and θ_ω, D be as in Proposition 17. Then $\operatorname{Im} \theta_\omega \circ D$ is a split submanifold of $C^\infty(I, \mathfrak{g}\setminus\{0\})$ modelled on $C^\infty(I, \mathfrak{m})$ and $\theta_\omega \circ D$ is a smooth embedding. In particular, $\mathcal{R}(\mathcal{P}_{c_0}) = \Psi_{g_0}(C^\infty(I, \mathfrak{m} \setminus \{0\}))$ is a split submanifold of $C^\infty(I, \mathfrak{g} \setminus \{0\})$ and $\mathcal{R}$ is a smooth embedding.*

Proof As $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, we have $C^\infty(I, \mathfrak{g}) = C^\infty(I, \mathfrak{h} \oplus \mathfrak{m}) \cong C^\infty(I, \mathfrak{h}) \oplus C^\infty(I, \mathfrak{m})$. Thus $C^\infty(I, \mathfrak{m} \setminus \{0\})$ is a closed and split submanifold of $C^\infty(I, \mathfrak{g} \setminus \{0\})$. Fix $g_0 \in G$ with $\pi(g_0) = c_0$ and note that Ψ_{g_0} restricts to a diffeomorphism $C^\infty(I, \mathfrak{g} \setminus \{0\}) \to C^\infty(I, \mathfrak{g} \setminus \{0\})$ by Lemma 28. Now as $\Psi_{g_0}(C^\infty(I, \mathfrak{m} \setminus \{0\})) = \operatorname{Im} \theta_\omega \circ D$ (cf. Lemma 29), the image $\operatorname{Im} \theta_\omega \circ D$ is a closed and split submanifold of $C^\infty(I, \mathfrak{g} \setminus \{0\})$. Further, we deduce from Lemma 29 that $\rho_{c_0}|_{\operatorname{Im}\theta_\omega \circ D}$ is smooth with $\theta_\omega \circ D \circ \rho_{c_0}|_{\operatorname{Im}\theta_\omega \circ D} = \operatorname{id}_{\operatorname{Im}\theta_\omega \circ D}$. As also $\rho_{c_0} \circ \theta_\omega = \operatorname{id}_{\operatorname{Imm}_{c_0}(I,\mathcal{M})}$, we see that θ_ω is a diffeomorphism onto its image. Thus $\theta_\omega \circ D$ is indeed a smooth embedding.

Since the scaling maps are diffeomorphisms $C^\infty(I, \mathfrak{g} \setminus \{0\}) \to C^\infty(I, \mathfrak{g} \setminus \{0\})$, the assertions on the image of $\mathcal{R}$ and on $\mathcal{R}$ follow directly from the assertions on θ_ω. □

3.1.1 (Reductive SRVT) Let $\mathcal{M}$ be a reductive homogeneous space with reductive complement $\mathfrak{m}$ and $\theta_\omega \colon C^\infty(I, T\mathcal{M}) \to C^\infty(I, \mathfrak{g}),\ f \mapsto \omega \circ f$ be constructed with respect to the 1-form ω from 3.3. Then $\Psi_{g_0^{-1}} \circ \theta_\omega(\mathcal{P}_{c_0}) = C^\infty(I, \mathfrak{m} \setminus \{0\})$ (see Appendix 6.1). Now one constructs a version of the SRVT for reductive spaces via

$$\mathcal{R}_\mathfrak{m} \colon \mathcal{P}_{c_0} \to C^\infty(I, \mathfrak{m} \setminus \{0\}), \quad f \mapsto \frac{\Psi_{g_0^{-1}} \circ \theta_\omega(\dot{f})}{\sqrt{\|\Psi_{g_0^{-1}} \circ \theta_\omega(\dot{f})\|}}.$$

We call this map *reductive SRVT*, to distinguish it from the usual SRVT. Contrary to the SRVT, the reductive SRVT will go into the reductive complement, but it will not be a section of ρ_{c_0}. Instead it is a section of $\rho_{c_0} \circ \Psi_{g_0}$. Finally, we note that by construction (cf. Lemma 28) the image of the reductive SRVT is $C^\infty(I, \mathfrak{m} \setminus \{0\})$.

Arguing as in Theorem 11, we also obtain a first order Sobolev metric by pullback with the reductive SRVT. In general this Riemannian metric will not coincide with the pullback metric obtained from the SRVT. The advantage of the reductive SRVT is that we have full control over its image, which happens to be an open subset (of a subspace of $C^\infty(I, \mathfrak{g})$). Since $C^\infty(I, \mathfrak{g})$ with respect to the L^2 inner product is a flat space (in the sense of Riemannian geometry), it follows that at least locally the geodesic distance on the image of the SRVT is given by the distance

$$d_{\mathcal{P}_{c_0}, \mathfrak{m}}(f, g) := d_{L^2}(\mathcal{R}_\mathfrak{m}(f), \mathcal{R}_\mathfrak{m}(g)).$$

However, we argue as in [9, Theorem 3.16] to obtain the following result.

Proposition 20 *If* $\dim\ \mathfrak{h} + 2 < \dim\ \mathfrak{g}$, *then the geodesic distance of* $(\mathcal{R}(\mathcal{P}_{c_0}), \langle \cdot, \cdot \rangle_{L^2})$ *coincides with the* L^2 *distance. In this case the geodesic distance on* $\mathcal{P}_{c_0}$ *induced by the pullback metric* (5) *(with respect to the reductive SRVT) is given by* $d_{\mathcal{P}_{c_0}, \mathfrak{m}}(f, g) = \sqrt{\int_I \left\| \mathcal{R}_\mathfrak{m}(f)(t) - \mathcal{R}_\mathfrak{m}(g)(t) \right\|^2 \mathrm{d}t}$.

Note that the modification by the reductive SRVT is highly non-linear, e.g. in the Lie group case, Example 14, we obtain:

Example 21 Let G be a Lie group, $c \in^\infty (I, G)$ and $\delta^l(c) = c^{-1}\dot{c}$. Then

$$\Psi(\delta^r(c)) = -\operatorname{Ad}(\operatorname{Evol}(\delta^r(c))^{-1}).\delta^r(c) = -\operatorname{Ad}(c^{-1}).\dot{c}c^{-1} = -\delta^l(c).$$

Recall from [15, 38.4] that $\operatorname{Evol}(-\delta^l(c))(t) = (c(t))^{-1}$. In the Lie group case, the reductive SRVT modifies the formulae to compute distances and interpolations between the pointwise inverses of curves instead of the curves themselves. In particular, this shows that the reductive SRVT will not be a section of ρ_{c_0}.

In particular, we have to prove a version of Lemma 13 for the reductive SRVT.

Lemma 22 *For a reductive space, $d_{\mathcal{P}_{c_0},\mathfrak{m}}$ is reparametrisation invariant.*

Proof For $\mathcal{R}_{\mathfrak{m}}$ we use $\Psi_{g_0^{-1}} \circ \theta_\omega$ instead of $\alpha = \theta_\omega$. Consider $f \in \mathcal{P}_{c_0}$ and $\varphi \in \mathrm{Diff}^+(I)$ to compute as in Lemma 13: $\Psi_{g_0}^{-1}(\theta_\omega(f \circ \varphi)) = \Psi_{g_0}^{-1}(\dot{\varphi} \cdot \theta_\omega(f) \circ \varphi)$. Now

$$\mathrm{Evol}(q) \circ \varphi = \mathrm{Evol}(\dot{\varphi} \cdot q \circ \varphi) \underbrace{\mathrm{Evol}(q)(\varphi(0))}_{=e \text{ since } \varphi(0)=0} = \mathrm{Evol}(\dot{\varphi} \cdot q \circ \varphi)$$

follows from [15, p. 411]. Linearity of the adjoint action yields $\Psi_{g_0}^{-1}(\theta_\omega(f \circ \varphi)) = (\Psi_{g_0}^{-1}(\theta_\omega(f)) \circ \varphi) \cdot \dot{\varphi}$. Inserting this in (9) yields reparametrisation invariance. □

4 The SRVT on Matrix Lie Groups

In order to illustrate our definition of the SRVT in different instances of homogeneous manifolds, we consider in what follows two examples of quotients of finite dimensional matrix Lie groups (for $n \geq 3$ and $p < n$):

1. $\mathrm{SO}(n)/(\mathrm{SO}(n-p) \times \mathrm{SO}(p))$ (see 4.3).
2. $\mathrm{SO}(n)/\mathrm{SO}(n-p)$ (see 4.2).

Note that in both cases the quotients are reductive homogeneous spaces. To prepare our investigation, we will now collect some information on relevant tangent spaces for the matrix Lie groups. These examples are relevant in applications [1].

4.1 Tangent Space of G/H and Tangent Map of $G \to G/H$

For G and H finite dimensional (matrix) Lie groups, we here describe the tangent space of G/H at a prescribed point c_0 and the tangent mapping of the canonical projection $\pi : G \to G/H$. We have seen that any curve $c(t)$ on G/H, $c(0) = c_0$, can be expressed non-uniquely by means of a curve on the Lie group $c(t) = \pi(g(t))$. For matrix Lie groups, the elements of G/H are equivalence classes of matrices. Let the elements of G, $g \in G$, be $n \times n$ matrices, then the group multiplication coincides with matrix multiplication. We identify elements of $H \subset G$ with matrices

$$h = \begin{bmatrix} I & 0 \\ 0 & \Gamma \end{bmatrix}, \quad \Gamma \text{ a } (n-p) \times (n-p) \text{ matrix and } I \text{ the } p \times p \text{ identity.} \tag{11}$$

We obtain $T_{g_0}\pi : T_{g_0}G \to T_{\pi(g_0)}G/H$, $v \mapsto w$, by differentiating $c(t) = \pi(g(t))$. Assuming $g(0) = g_0$, $\pi(g_0) = c_0$, $\dot{g}(0) = v \in T_{g_0}G$, we have

$$w := T_{\pi(g_0)}(v) = \frac{d}{dt}\Big|_{t=0} \pi(g(t)) = \left\{ \frac{d}{dt}\Big|_{t=0} \tilde{g}(t) \mid \tilde{g}(t) = g(t)\,h(t), \quad h(t) \in H \right\}.$$

Assuming $\dot{g}(t) = A(t)g(t)$, where $A(t) \in \mathfrak{g}$, $v = A_0 g_0 = g_0 \operatorname{Ad}_{g_0^{-1}}(A_0)$, and assuming also that $\frac{d}{dt}h(t) = B(t)h(t)$, $B(t) \in \mathfrak{h}$, $\quad B(0) = B_0$, in analogy to (32), we get

$$\frac{d}{dt}\tilde{g}(t) = \big(A(t) + \operatorname{Ad}_{g(t)}(B(t))\big)\; g(t)h(t) = g(t)\left(\operatorname{Ad}_{g(t)^{-1}}(A(t)) + B(t)\right)\; h(t), \tag{12}$$

so we obtain

$$\begin{aligned} w := T_{\pi(g_0)}(v) &= \left\{ \tilde{w} = \frac{d}{dt}\Big|_{t=0} \tilde{g}(t) \mid \tilde{w} = (A_0 + \operatorname{Ad}_{g_0}(B_0))\, g_0\, h, \quad h \in H, B_0 \in \mathfrak{h} \right\} \\ &= \left\{ \tilde{w} = \frac{d}{dt}\Big|_{t=0} \tilde{g}(t) \mid \tilde{w} = g_0\,(\operatorname{Ad}_{g_0^{-1}}(A_0) + B_0)\, h, \quad h \in H, B_0 \in \mathfrak{h} \right\}, \end{aligned}$$

which gives a description of the tangent vector $w \in T_{c_0}G/H$ as well as the characterisation of $T\pi$ for matrix Lie groups. Suppose that we fix a complementary subspace $\mathfrak{m}$ of $\mathfrak{h}$, $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, then there is a unique isotropy element $B_0 \in \mathfrak{h}$ such that $\operatorname{Ad}_{g_0^{-1}}(A_0) + B_0 \in \mathfrak{m}$.

Repeating this procedure for each value of t along a curve $c(t)$, we can assume $c(t) = \pi(g(t))$ and $w(t) \in T_{\pi(g(t))}G/H$, $w(t) = (A(t) + \operatorname{Ad}_{g(t)}(B(t)))c(t)$ with $A(t) \in \mathfrak{g}$, $B(t) \in \mathfrak{h}$, such that $\operatorname{Ad}_{g(t)^{-1}}(A(t)) + B(t) \in \mathfrak{m}$, then we can define

$$\alpha : T_{\pi(g(t))}G/H \to \operatorname{Ad}_{g(t)}(\mathfrak{m}), \qquad \alpha(w(t)) = A(t) + \operatorname{Ad}_{g(t)}(B(t)).$$

This map corresponds to the map θ_ω of 3.4 with ω as described in 3.3. If $\mathfrak{m}$ is reductive, this map is well defined (independently of the choice of representative $g(t)$ of $c(t) = \pi(g(t))$). We refer to Table 1 for different, possible choices of $\mathfrak{m}$ and their implications. In the following examples $\mathfrak{m}$ is reductive and H is compact.

4.2 *SRVT on the Stiefel Manifold:* $\mathbf{SO(n)/SO(n-p)}$

In this section we consider the case when $G = \mathrm{SO}(n)$ and $H = \mathrm{SO}(n-p) \subset \mathrm{SO}(n)$, where the elements of $\mathrm{SO}(n-p)$ are of the type (11) with Γ a $(n-p) \times (n-p)$ orthogonal matrix with determinant equal to 1. We consider the canonical left action

Table 1 Riemannian metrics and decompositions of the Lie algebra

H / $\mathfrak{h}$	Metric on G	Special decompositions of $\mathfrak{g}$	G-action on $\mathcal{M}$
Compact	G-left invariant, H-biinvariant	$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{h}^{\perp}$, the orthogonal complement $\mathfrak{h}^{\perp}$ is $\mathrm{Ad}(H)$-invariant	By isometries
Compact	G-right invariant, H-biinvariant	As above	Only H acts by isometries
Admits reductive complement in $\mathfrak{g}$[a]	G-right invariant	$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, $\mathfrak{m}$ is $\mathrm{Ad}(H)$-invariant $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{h}^{\perp}$, where in general $\mathfrak{m} \neq \mathfrak{h}^{\perp}$	Not by isometries
	G-right invariant	$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{h}^{\perp}$ but $\mathfrak{h}^{\perp}$ is not $\mathrm{Ad}(H)$ invariant	Not by isometries

[a] $\mathfrak{h} = \mathbf{L}(H)$ admits a *reductive complement* $\mathfrak{m}$, if $\mathfrak{m}$ is an $\mathrm{Ad}(H)$-invariant subspace and $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ as vector spaces, cf. 3.1. Then $\mathcal{M} = G/H$ is a reductive homogeneous space

of $\mathrm{SO}(n)$ on the quotient $\mathrm{SO}(n)/\mathrm{SO}(n-p)$. This homogeneous manifold can be identified with the Stiefel manifold, $\mathcal{M} = \mathbb{V}_p(\mathbb{R}^n)$, i.e. the set of p-orthonormal frames in $\mathbb{R}^n$ (real matrices $n \times p$ with orthonormal columns). We will in the following denote by $[U, U^{\perp}]$ the elements of $\mathrm{SO}(n)$ where we have collected in U the first p orthonormal columns and in $U^{\perp}$ the last $n-p$. Multiplication from the right by an arbitrary element in the isotropy subgroup $\mathrm{SO}(n-p)$ gives $[U, U^{\perp}\Gamma]$, leaving the first p columns unchanged and orthonormal to the last $n-p$, for all choices of Γ. Here U alone represents the whole coset of $[U, U^{\perp}]$. When thought of as a map from $\mathrm{SO}(n)$ to $\mathrm{SO}(n)/\mathrm{SO}(n-p)$, the projection $\pi\colon \mathrm{SO}(n) \to \mathrm{SO}(n)/\mathrm{SO}(n-p)$ is

$$\pi([U, U^{\perp}]) = \{\tilde{g} \in \mathrm{SO}(n) \mid \tilde{g} = [U, U^{\perp}\Gamma], \quad \forall\, \Gamma \in \mathrm{SO}(n-p)\}.$$

Otherwise, when thought of as a map from $\pi : \mathrm{SO}(n) \to \mathbb{V}_p(\mathbb{R}^n)$, the canonical projection conveniently becomes $\pi([U, U^{\perp}]) = [U, U^{\perp}]\, I_p = U$, where I_p is the $n \times p$ matrix whose columns are the first p columns of the $n \times n$ identity matrix. The equivalence class of the group identity element $\pi(e)$ is identified with the $n \times p$ matrix I_p. Similarly the tangent mapping of the projection π,

$$T\pi : T\mathrm{SO}(n) \to T\mathrm{SO}(n)/\mathrm{SO}(n-p), \qquad v \in T_g\mathrm{SO}(n) \mapsto w \in T_{\pi(g)}\mathrm{SO}(n)/\mathrm{SO}(n-p),$$

with $g = [U, U^{\perp}]$, $v = A\,[U, U^{\perp}] \in T_{[U,U^{\perp}]}\mathrm{SO}(n)$ and $A \in \mathfrak{so}(n)$, can be realised as

$$T_{[U,U^{\perp}]}\pi(A\,[U, U^{\perp}]) = \left\{ \tilde{w} \in T_{[U,U^{\perp}\Gamma]}\,\mathrm{SO}(n) \;\middle|\; \begin{array}{l} \tilde{w} = [U, U^{\perp}](\mathrm{Ad}_{[U,U^{\perp}]^T}(A) + B)\,\Gamma, \\ \quad \forall\, \Gamma \in \mathrm{SO}(n-p),\ B \in \mathfrak{so}(n-p) \end{array} \right\}, \tag{13}$$

while $T\pi : T\mathrm{SO}(n) \to T\mathbb{V}_p(\mathbb{R}^n)$ by multiplication from the right by I_p, and

$$T_{[U,U^\perp]}\pi(A\,[U, U^\perp]) = A\,[U, U^\perp]\,I_p = AU \in T_U\mathbb{V}_p(\mathbb{R}^n). \tag{14}$$

Alternatively, a characterisation of tangent vectors can be obtained by differentiation of curves on $\mathbb{V}_p(\mathbb{R}^n)$. We have then that

$$T_Q\mathbb{V}_p(\mathbb{R}^n) = \{V \;\; n \times p \;\; \text{matrix} \mid Q^{\mathrm{T}}V \;\; p \times p \;\; \text{skew-symmetric}\}.$$

Proposition 23 ([10]) *Any tangent vector V at $Q \in \mathbb{V}_p(\mathbb{R}^n)$ can be written as*

$$V = (FQ^{\mathrm{T}} - QF^{\mathrm{T}})\,Q, \tag{15}$$

$$F := V - Q\,\frac{Q^{\mathrm{T}}V}{2} \in T_Q\mathcal{M}. \tag{16}$$

And notice that replacing F with $F := V - Q(\frac{Q^{\mathrm{T}}V}{2} + S)$, where S is an arbitrary $p \times p$ symmetric matrix, does not affect (15).

We proceed by using the representation (15) of $T_Q\mathbb{V}_p(\mathbb{R}^n)$ and the framework described in Definition 4 for defining an SRVT on the Stiefel manifold. Consider

$$f_Q : T_Q\mathcal{M} \to T_Q\mathcal{M}, \qquad f_Q(V) = V - Q\frac{Q^{\mathrm{T}}V}{2}, \tag{17}$$

$$a_Q : T_Q\mathcal{M} \to \mathfrak{m}_Q \subset \mathfrak{g}, \qquad a_Q(V) = f_Q(V)Q^{\mathrm{T}} - Qf_Q(V)^{\mathrm{T}}. \tag{18}$$

The SRVT of a curve $Y(t)$ on the Stiefel manifold is a curve on $\mathfrak{so}(n)$ defined by

$$\mathcal{R}(Y) := \frac{a_Y(\dot{Y})}{\sqrt{\|a_Y(\dot{Y})\|}} = \frac{f_Y(\dot{Y})Y^{\mathrm{T}} - Yf_Y(\dot{Y})^{\mathrm{T}}}{\sqrt{\|f_Y(\dot{Y})Y^{\mathrm{T}} - Yf_Y(\dot{Y})^{\mathrm{T}}\|}}. \tag{19}$$

As the Stiefel manifold is a reductive homogeneous space, we can define a reductive SRVT in this case. Denoting with $[Q, Q^\perp]$ a representative in $SO(n)$ of the equivalence class identified by Q on $\mathbb{V}_p(\mathbb{R}^n)$, we observe that

$$V = \mathrm{Ad}_{[QQ^\perp]}(G)I_p \quad \text{with} \quad G := [QQ^\perp]^T F I_p^T - I_p F^T [QQ^\perp].$$

Assuming the right invariant metric on $SO(n)$ is the negative Killing form, then we observe that G belongs to the orthogonal complement of the subalgebra $\mathfrak{so}(n-p)$ in $\mathfrak{so}(n)$ with respect to this inner product. As stated in Table 1, this orthogonal complement is the reductive complement, i.e. $\mathfrak{m} = \mathfrak{so}(n-p)^\perp$, and $\mathrm{Ad}_{\mathrm{SO}(n-p)}(\mathfrak{so}(n-p)^\perp) \subset \mathfrak{so}(n-p)^\perp$. The elements of such an orthogonal

complement $\mathfrak{so}(n-p)^{\perp}$ are matrices $W \in \mathfrak{so}(n)$ of the form

$$W = \begin{bmatrix} \Omega & \Sigma^T \\ -\Sigma & 0 \end{bmatrix}, \tag{20}$$

with $\Omega \in \mathfrak{so}(p)$ and Σ an arbitrary $(n-p) \times p$ matrix. Consider the maps

$$\tilde{f}_Q : T_Q\mathcal{M} \to T_{I_p}\mathcal{M}, \qquad \tilde{f}_Q(V) = [Q\,Q^{\perp}]^T V - I_p \frac{Q^{\mathrm{T}}V}{2}, \tag{21}$$

$$\tilde{a}_Q : T_Q\mathcal{M} \to \mathfrak{m} \subset \mathfrak{g}, \qquad \tilde{a}_Q(V) = \tilde{f}_Q(V) I_p^{\mathrm{T}} - I_p \tilde{f}_Q(V)^{\mathrm{T}}, \tag{22}$$

and we observe that $\tilde{a}_Q(V) \in \mathfrak{m}$. Then the reductive SRVT is

$$\mathcal{R}_{\mathfrak{m}}(Y) := \frac{\tilde{a}_Y(\dot{Y})}{\sqrt{\|\tilde{a}_Y(\dot{Y})\|}} = \frac{\tilde{f}_Y(\dot{Y}) I_p^{\mathrm{T}} - I_p \tilde{f}_Y(\dot{Y})^{\mathrm{T}}}{\sqrt{\|\tilde{f}_Y(\dot{Y}) I_p^{\mathrm{T}} - I_p \tilde{f}_Y(\dot{Y})^{\mathrm{T}}\|}}. \tag{23}$$

4.3 SRVT on the Grassmann Manifold: $\mathbf{SO}(n)/(\mathbf{SO}(n-p) \times \mathbf{SO}(p))$

In this section we consider the case when $G = \mathrm{SO}(n)$ and $H = \mathrm{SO}(n-p) \times \mathrm{SO}(p) \subset \mathrm{SO}(n)$ where the elements of $\mathrm{SO}(n-p) \times \mathrm{SO}(p)$ are of the type

$$h = \begin{bmatrix} \Lambda & 0 \\ 0 & \Gamma \end{bmatrix}, \tag{24}$$

with Λ a $p \times p$ matrix and Γ an $(n-p) \times (n-p)$ matrix, both orthogonal with determinant equal to 1. We consider the canonical left action of $\mathrm{SO}(n)$ on the quotient $\mathrm{SO}(n)/(\mathrm{SO}(n-p) \times \mathrm{SO}(p))$. This homogeneous manifold can be identified with a quotient of the Stiefel manifold $\mathbb{V}_p(\mathbb{R}^n)/\mathrm{SO}(p)$ with equivalence classes $[Q] = \{\tilde{Q} \in \mathbb{V}_p(\mathbb{R}^n) \mid \tilde{Q} = Q\,\Lambda,\ \Lambda \in \mathfrak{so}(p)\}$. We denote such a manifold here with $\mathbf{G}_{p,n}(\mathbb{R})$.[8] The reductive subspace is $\mathfrak{m} = (\mathfrak{so}(p) \times \mathfrak{so}(n-p))^{\perp}$ with elements as in (20) but with $\Omega = 0$. Imposing a choice of isotropy $B \in \mathfrak{so}(p) \times \mathfrak{so}(n-p)$ such that $(\mathrm{Ad}_{[Q,Q^{\perp}]^T}(A) + B) \in \mathfrak{m}$ leads to the following characterisation of tangent vectors.

Proposition 24 *Any tangent vector V at $Q \in \mathbf{G}_{p,n}(\mathbb{R})$ is an $n \times p$ matrix such that $Q^T V = 0$, and V can be expressed in the form (15) with $F = V$.*

[8] An alternative representation of $\mathbf{G}_{p,n}$ is given by considering symmetric matrices P, $n \times n$, with $\mathrm{rank}(P) = p$ and $P^2 = P$, [14].

The proof follows from (12) assuming $g(t) = [Q(t)Q(t)^{\perp}] \in \mathrm{SO}(n)$, and $h(t)$ of the form (24), imposing the stated choice of isotropy, and projecting the resulting curves on $\mathbb{V}_p(\mathbb{R}^n)$ by post-multiplication by I_p.

We proceed by using (15) but with $F = V$. Define $a_Q : T_Q\mathcal{M} \to \mathfrak{g}$ as in (18) with $f_Q : T_Q\mathcal{M} \to T_Q\mathcal{M}$, the identity map $f_Q(V) = V$. Suppose that $Y(t)$ is a curve on the Grassmann manifold, then the SRVT of Y is a curve on $\mathfrak{so}(n)$ and takes the form (19) which here becomes

$$\mathcal{R}(Y) := \frac{\dot{Y}Y^{\mathrm{T}} - Y\dot{Y}^{\mathrm{T}}}{\sqrt{\|\dot{Y}Y^{\mathrm{T}} - Y\dot{Y}^{\mathrm{T}}\|}}. \tag{25}$$

The reductive SRVT is defined by (23) with

$$\tilde{f}_Q(V) = [Q, Q^{\perp}]^T V = \begin{bmatrix} O \\ (Q^{\perp})^T V \end{bmatrix}$$

and $\tilde{a}_Q$ as in (22), which implies $\tilde{a}_Q(V) \in \mathfrak{m}$.

5 Numerical Experiments

To demonstrate an application of the SRVT introduced in this paper, we present a simple example of interpolation between two curves on the unit 2-sphere. In the following we describe some implementation details for this example.

5.1 (Preliminaries) We will use Rodrigues' formula for the Lie group exponential,

$$\exp(\hat{x}) = I + \frac{\sin(\alpha)}{\alpha}\hat{x} + \frac{1-\cos(\alpha)}{\alpha^2}\hat{x}^2, \quad \alpha = \|x\|_2, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \hat{x} = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$$

where $x \mapsto \hat{x}$ defines an isomorphism between vectors in $\mathbb{R}^3$ and 3×3 skew-symmetric matrices in $\mathfrak{so}(3)$.

5.2 (Interpolated curves) Given a continuous curve $c(t), t \in [t_0, t_N]$ on the Stiefel manifold $\mathrm{SO}(3)/\mathrm{SO}(2)$, which is diffeomorphic to S^2, we replace $c(t)$ with the curve $\bar{c}(t)$ interpolating between $N+1$ values $\bar{c}_i = c(t_i)$, with $t_0 < t_1 < \ldots < t_N$, as follows:

$$\bar{c}(t) := \sum_{i=0}^{N-1} \chi_{[t_i, t_{i+1})}(t) \exp\left(\frac{t - t_i}{t_{i+1} - t_i}\left(v_i \bar{c}_i^{\mathrm{T}} - \bar{c}_i v_i^{\mathrm{T}}\right)\right)\bar{c}_i, \tag{26}$$

where χ is the characteristic function, exp is the Lie group exponential, and v_i are approximations to $\frac{d}{dt}c(t)\Big|_{t=t_i}$ found by solving the equations

$$\bar{c}_{i+1} = \exp\left(v_i \bar{c}_i^{\mathrm{T}} - \bar{c}_i v_i^{\mathrm{T}}\right)\bar{c}_i \tag{27}$$

$$\text{constrained by} \quad v_i^{\mathrm{T}}\bar{c}_i = 0. \tag{28}$$

The v_i, $i = 1, \ldots, N$, can be found explicitly, by a simple calculation. We observe that if $\kappa = \bar{c}_i \times v_i$, then $\hat{\kappa} = v_i \bar{c}_i^{\mathrm{T}} - \bar{c}_i v_i^{\mathrm{T}}$. By (28), we have that $\|\bar{c}_i \times v_i\|_2 = \|\bar{c}_i\|_2 \|v_i\|_2 = \|v_i\|_2$, where the last equality follows because we assume the sphere to have radius 1, and so $\|\bar{c}_i\|_2 = \bar{c}_i^{\mathrm{T}}\bar{c}_i = 1$. Using Rodrigues' formula, from (27) we obtain

$$\bar{c}_{i+1} = \frac{\sin\left(\|v_i\|_2\right)}{\|v_i\|_2} v_i + \cos\left(\|v_i\|_2\right)\bar{c}_i.$$

Thus $\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1} = 1 - \cos\left(\|v_i\|_2\right)$ and so $\|v_i\|_2 = \arccos\left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)$ leading to

$$v_i = \left(\bar{c}_{i+1} - \bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\bar{c}_i\right) \frac{\arccos\left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)}{\sqrt{1 - \left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)^2}}.$$

Inserting this into (26) gives

$$\bar{c}(t) = \sum_{i=0}^{N-1} \chi_{[t_i, t_{i+1})}(t) \exp\left(\frac{t - t_i}{t_{i+1} - t_i} \frac{\arccos\left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)}{\sqrt{1 - \left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)^2}} \left(\bar{c}_{i+1}\bar{c}_i^{\mathrm{T}} - \bar{c}_i\bar{c}_{i+1}^{\mathrm{T}}\right) \right) \bar{c}_i. \tag{29}$$

5.3 (The SRVT and its inverse) By Definition 4 and formulae (17), (18) and (19), the SRVT of the curve (29) is a piecewise constant function $\bar{q}(t)$ in $\mathfrak{so}(3)$, taking values $\bar{q}_i = \bar{q}(t_i)$, $i = 0, \ldots, N-1$, where $\bar{q}_i = \mathcal{R}(\bar{c})\big|_{t=t_i}$ is given by

$$\bar{q}_i = \frac{v_i \bar{c}_i^{\mathrm{T}} - \bar{c}_i v_i^{\mathrm{T}}}{\|v_i \bar{c}_i^{\mathrm{T}} - \bar{c}_i v_i^{\mathrm{T}}\|^{\frac{1}{2}}} = \frac{\arccos^{\frac{1}{2}}\left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)}{\left(1 - \left(\bar{c}_i^{\mathrm{T}}\bar{c}_{i+1}\right)^2\right)^{\frac{1}{4}} \|\bar{c}_{i+1}\bar{c}_i^{\mathrm{T}} - \bar{c}_i\bar{c}_{i+1}^{\mathrm{T}}\|^{\frac{1}{2}}} \left(\bar{c}_{i+1}\bar{c}_i^{\mathrm{T}} - \bar{c}_i\bar{c}_{i+1}^{\mathrm{T}}\right). \tag{30}$$

Here the norm $\|\cdot\|$ is induced by the negative (scaled) Killing form, which for skew-symmetric matrices corresponds to the Frobenius inner product, $\|A\| = \sqrt{\mathrm{tr}(AA^{\mathrm{T}})}$.

The inverse SRVT is then given by (29), with

$$\bar{c}_{i+1} = \exp\left(\|\bar{q}_i\|\bar{q}_i\right)\bar{c}_i, \quad i = 1, \ldots, N-1, \quad \bar{c}_0 = c(t_0).$$

5.4 (The reductive SRVT) Since $\mathrm{Evol}(a_{\bar{c}_i}(v_i)) = \exp(a_{\bar{c}_i}(v_i))$, the reductive SRVT (3.1.1) becomes then

$$\mathcal{R}_{\mathfrak{m}}(\bar{c})\big|_{t=t_i} = \mathcal{R}(\bar{c})\big|_{t=t_i} = \frac{\arccos^{\frac{1}{2}}\left(\tilde{c}_i^{\mathrm{T}}\tilde{c}_{i+1}\right)}{\left(1-\left(\tilde{c}_i^{\mathrm{T}}\tilde{c}_{i+1}\right)^2\right)^{\frac{1}{4}}\|\tilde{c}_{i+1}\tilde{c}_i^{\mathrm{T}} - \tilde{c}_i\tilde{c}_{i+1}^{\mathrm{T}}\|^{\frac{1}{2}}}\left(\tilde{c}_{i+1}\tilde{c}_i^{\mathrm{T}} - \tilde{c}_i\tilde{c}_{i+1}^{\mathrm{T}}\right), \tag{31}$$

with

$$\tilde{c}_i = [U, U^{\perp}]_i^{\mathrm{T}}\bar{c}, \quad i = 0, \ldots, N, \qquad [U, U^{\perp}]_{i+1} = \exp(a_{\bar{c}_i}(v_i))[U, U^{\perp}]_i \quad i = 0, \ldots, N-1,$$

where $[U, U^{\perp}]_0$ can be found e.g. by QR-factorization of $c(t_0)$.

5.5 (Curve blending on the 2-sphere) We wish to compute the geodesic in the shape space of curves on the sphere between the two curves $\bar{c}^1(t)$ and $\bar{c}^2(t)$. Following [9], we use a dynamic programming algorithm to solve the optimization problem (7) (see [7, 28] for a detailed description on the use of dynamic programming for shapes):

Algorithm 1 REPARAMETRISATION[7, Section 3.2]

Given $\bar{q}^1(t), \bar{q}^2(t), N, \{t_i\}_{i=0}^N$
for $i, j \in \{0, \ldots, N\}$ **do**
  $c_{\min} = \infty$
  for $k \in \{0, \ldots, i-1\}, l \in \{0, \ldots, j-1\}$ **do**
    $c_{\mathrm{loc}} = \int_{t_i}^{t_k} |\bar{q}^1(t) - \bar{q}^2(t_l + \frac{t_j - t_l}{t_i - t_k}t)|^2 \mathrm{d}t$
    if $\Psi^m(k,l) = \Psi \circ \cdots \circ \Psi(k,l) = (0,0)$ for some $m \geq 0$ **then**
      $z = 0$
    else
      $z = \infty$
    $c = c_{\mathrm{loc}} + A_{k,l} + z$
    if $c < c_{\min}$ **then**
      $c_{\min} = c$
      $\Psi(i,j) = (k,l)$
  $A_{i,j} = c_{\min}$
Create two vectors of indices (p, q) by setting $(p_0, q_0) = (N, N)$ and $(p_{m+1}, q_{m+1}) = \Psi(p_m, q_m)$ until $(p_{m+1}, q_{m+1}) = (0, 0)$
Flip (p, q) so it starts at $(0, 0)$ and ends at (N, N)
for $i \in \{0, \ldots, N\}$ **do**
  $s_i = t_{q_j} + (t_{q_{j+1}} - t_{q_j})\frac{t_i - t_{p_j}}{t_{p_{j+1}} - t_{p_j}}$ for j s.t. $p_j \leq i < p_{j+1}$
Return $s = \{s_i\}_{i=0}^N$

With this approach, we reparametrise optimally the curve $\bar{c}^2(t)$ while minimizing its distance to $\bar{c}^1(t)$. This distance is measured by taking the L^2 norm of $\bar{q}^1(t) - \bar{q}^2(t)$ in the Lie algebra. In the discrete case, this reparametrisation yields an optimal set of grid points $\{s_i\}_{i=0}^N$, where $s_0 = t_0 < s_1 < \ldots < s_N = t_N$, from which we find $\bar{c}_i'^2 = \bar{c}^2(s_i)$ by (29). See [9] for further details.

We interpolate between $\bar{c}^1(t)$ and $\bar{c}'^2(t)$ by performing a linear convex combination of their SRV transforms $\bar{q}^1(t)$ and $\bar{q}'^2(t)$, and then by taking the inverse SRVT of the result. We obtain

$$\bar{c}_{\text{int}}(\bar{c}_1, \bar{c}_2', \theta) = \mathcal{R}^{-1}\left((1-\theta)\,\mathcal{R}(\bar{c}_1) + \theta\,\mathcal{R}(\bar{c}_2')\right), \qquad \theta \in [0,1].$$

Examples are reported in Figs. 2, 3 and 4, where interpolation between two curves is performed with and without reparametrisation. We show curves resulting from using both (30) and (31), and compare these to the results obtained when employing the SRVT introduced in [9] on curves in SO(3) which are then traced out by a vector in $\mathbb{R}^3$ to match the curves in S^2 studied here.

5.6 (Conclusions) We have proposed generalisations of the SRVT approach to curves and shapes evolving on homogenous manifolds using Lie group actions. Different Lie group actions lead to different Riemannian metrics in the infinite dimensional manifolds of curves and shapes opening up for a variety of possibilities which can all be implemented in the same generalised SRVT framework. We have presented only a few preliminary examples here, and further tests and analysis will be the subject of future work. A number of open questions related to the properties of the pullback metrics through the SRVT, to the performance of the algorithms when using different Lie group actions, to the comparison of the SRVT and the reductive SRVT and to the implementation of the approach in examples of non reductive homogeneous manifolds will be addressed in future research.

6 Appendix

6.1 (Auxiliary results for Sect. 3)

Lemma 25 *For the homogeneous space $\mathcal{M} = G/H$ with projection $\pi\colon G \to G/H$ the derivation map $D_{\mathcal{M}}\colon C^\infty(I, G/H) \to C^\infty(I, T(G/H)), c \mapsto \dot{c}$ is smooth.*

Proof The map $D_G\colon C^\infty(I,G) \to C^\infty(I, TG)$, $\gamma \mapsto \dot{\gamma}$ is a smooth group homomorphism by [13, Lemma 2.1]. As $\pi\colon G \to G/H$ is a smooth submersion, $\theta_\pi : C^\infty(I,G) \to C^\infty(I, G/H), c \mapsto \pi \circ c$ is a smooth submersion [2, Lemma 2.4]. Write $\theta_{T\pi} \circ D_G = D \circ \theta_\pi$, whence by [12, Lemma 1.8] $D_{\mathcal{M}}$ is smooth. □

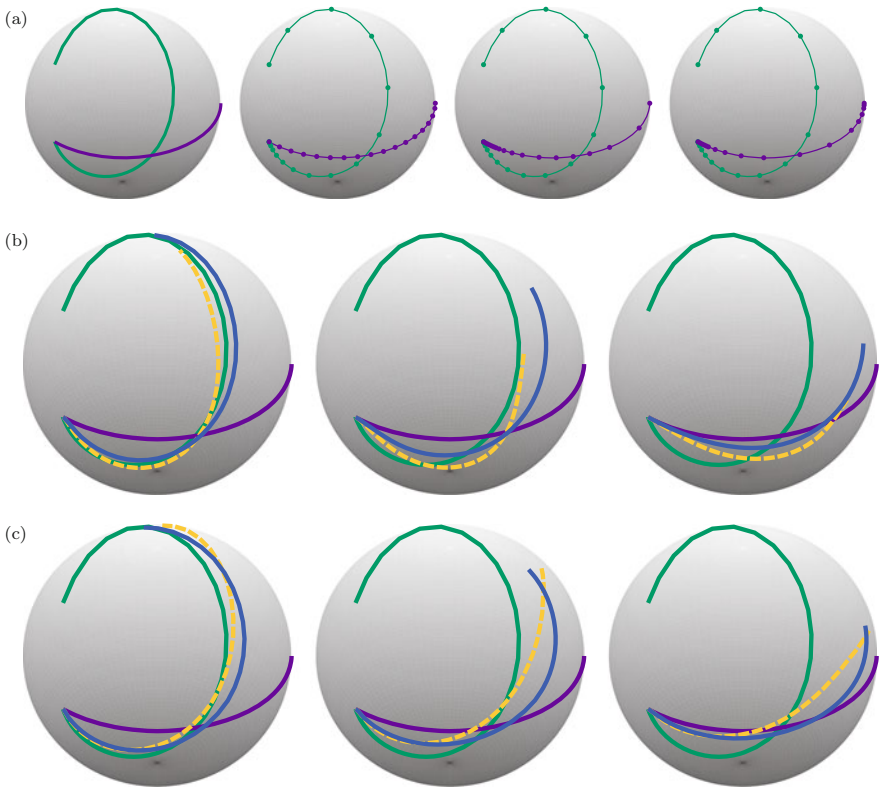


Fig. 2 Interpolation between two curves on S^2, with and without reparametrisation, obtained by the reductive SRVT (31). The results obtained by using the SRVT (30) are not identical to these, but in this case very similar, and therefore omitted. The results are compared to the corresponding SRVT interpolation between curves on SO(3), which are then mapped to S^2 by multiplying with the vector $(1, 0, 0)^T$. The curves are $c^1(t) = R_x(\pi t^3)R_y(\pi t^3)R_y(\pi t^3/2) \cdot (1, 0, 0)^T$ and $c^2(t) = R_z(3\pi t/4)R_x(\pi t) \cdot (1, 0, 0)^T$ for $t \in [0, 1]$, where $R_x(t)$, $R_y(t)$ and $R_z(t)$ are the rotation matrices in SO(3) corresponding to rotation of an angle t around the x-, y- and z-axis, respectively. (**a**) From left to right: Two curves on the sphere, their original parametrisations, the reparametrisation minimizing the distance in SO(3) and the reparametrisation minimizing the distance in S^2, using the reductive SRVT (31). (**b**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, after reparametrisation, on SO(3) (yellow, dashed line) and S^2 (blue, solid line). (**c**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, before reparametrisation, on SO(3) (yellow, dashed line) and S^2 (blue, solid line)

Lemma 26 *With* $\theta := \theta_\omega \circ D$ *The identity* (10) $\mathrm{id}_{C^\infty_{eH}(I,\mathcal{M})} = \pi \circ \mathrm{Evol} \circ \theta$ *holds.*

Proof Let $c\colon I \to \mathcal{M}$ be smooth with $c(0) = eH$ and choose $g\colon I \to G$ smooth with $g(0) = e$ and $\pi \circ g = c$. Set $\gamma(t) := \mathrm{Evol}(\theta(c))(t)$. It suffices to prove that $\gamma(t)^{-1}g(t) \in H$ for all $t \in I$. Then $\pi \circ \gamma = \pi \circ g = c$ and the assertion follows.

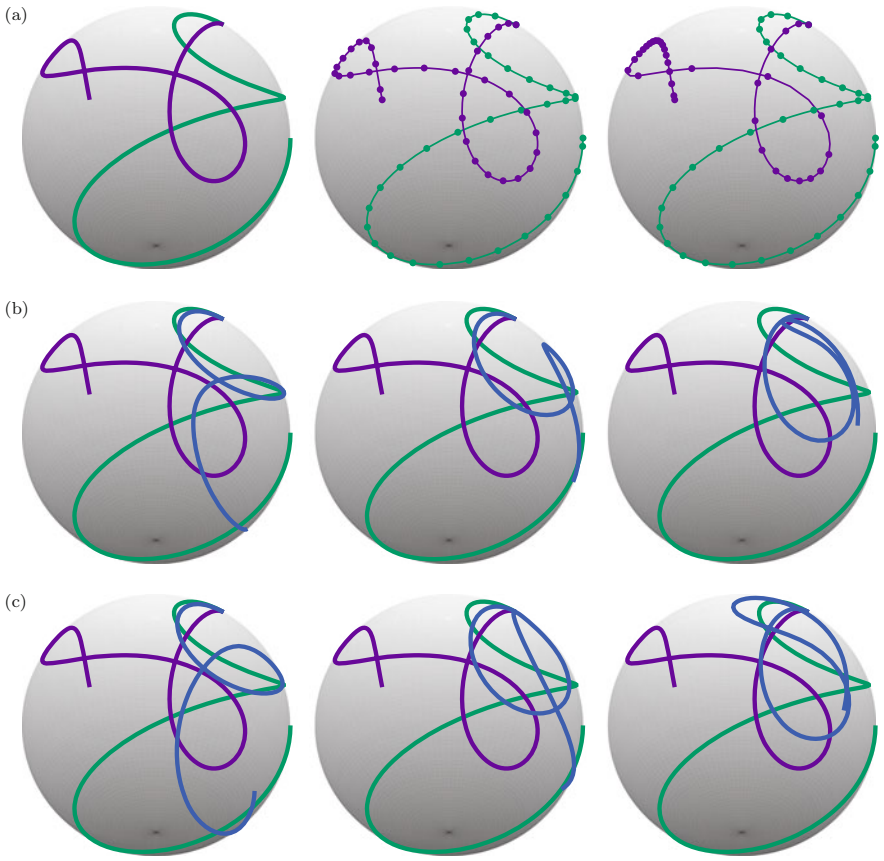


Fig. 3 Interpolation between two curves on S^2, with and without reparametrisation, found by the reductive SRVT (31). The curves are $c^1(t) = R_x(2\pi t)R_y(2\pi t)R_z(\pi t) \cdot (0,1,1)^{\mathrm{T}}/\sqrt{2}$ and $c^2(t) = R_z(2\pi t)R_x(2\pi t)R_y(\pi t/2) \cdot (0,1,1)^{\mathrm{T}}/\sqrt{2}$ for $t \in [0,1]$, where $R_x(t)$, $R_y(t)$ and $R_z(t)$ are the rotation matrices in SO(3) corresponding to rotation of an angle t around the x-, y- and z-axis, respectively. (**a**) From left to right: Two curves on the sphere, their original parametrisations and the reparametrisation minimizing the distance in S^2, using the reductive SRVT (31). (**b**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, before reparametrisation. (**c**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, after reparametrisation

As $\gamma(0)^{-1}g(0) = e \in H$, we only have to prove that $\frac{\mathrm{d}}{\mathrm{d}t}\pi(\gamma(t)^{-1}g(t))$ vanishes everywhere to obtain $\gamma(t)^{-1}g(t) \in H$. Before we compute the derivative of $\pi(\gamma(t)^{-1}g(t))$, let us first collect some facts concerning the logarithmic derivatives $\delta^r(f) = \dot{f}.f^{-1}$ and $\delta^l(f) = f^{-1}.\dot{f}$. By definition $\delta^r(\gamma) = \delta^r(\mathrm{Evol}(\theta(c))) = \theta(c)$.

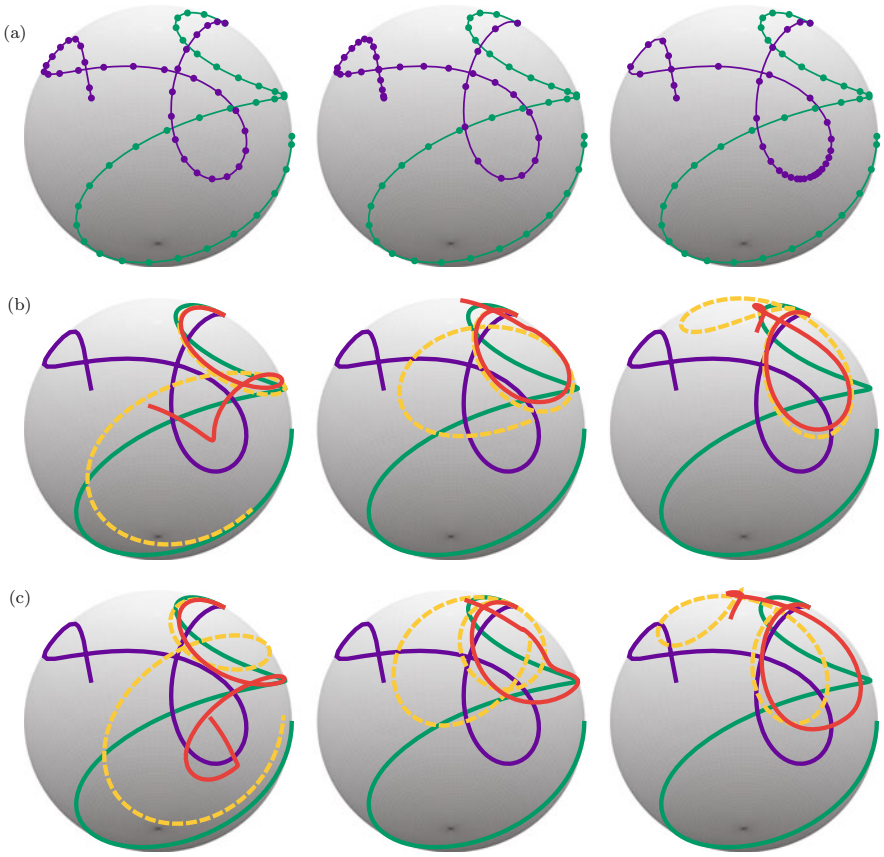


Fig. 4 Interpolation between the same curves as in Fig. 3, with and without reparametrisation, obtained here with the SRVT (30), compared to the corresponding interpolation between curves on SO(3) mapped to S^2 by multiplication with the vector $(0, 1, 1)^T/\sqrt{2}$. (**a**) From left to right: The original parametrisations of the curves to be interpolated, the reparametrisation minimizing the distance in SO(3) and the reparametrisation minimizing the distance in S^2, using the SRVT (30). (**b**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, before reparametrisation, on SO(3) (yellow, dashed line) and S^2 (red, solid line). (**c**) The interpolated curves at times $t = \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$, from left to right, after reparametrisation, on SO(3) (yellow, dashed line) and S^2 (red, solid line)

Further, [15, Lemma 38.1] yields for smooth $f, h\colon I \to G$:

$$\delta^r(f \cdot h) = \delta^r(f) + \mathrm{Ad}(f).\delta^r(h) \quad \text{and} \quad \delta^r(f^{-1}) = -\delta^l(f), \quad \text{whence} \tag{32}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\gamma(t)^{-1}g(t)) &= (\gamma(t)^{-1}g(t)) \cdot \delta^l(\gamma^{-1}g)(t) \overset{(32)}{=} -(\gamma(t)^{-1}g(t)) \cdot \delta^r(g^{-1}\gamma)(t) \\ &\overset{(32)}{=} (\gamma(t)^{-1}g(t)) \cdot (\delta^l(g)(t) - \mathrm{Ad}(g(t)^{-1}).\theta(c)(t))\end{aligned}$$

Recall that by definition, $\theta(c)(t) = \omega(\dot{c}(t)) = \mathrm{Ad}(g(t)).\omega_e(TA^{g(t)^{-1}}(\dot{c}(t)))$ (here $\pi \circ g = c$ is used). Inserting this into the above equation we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma(t)^{-1}g(t)) = (\gamma(t)^{-1}g(t)) \cdot (\delta^l(g)(t) - \omega_e(TA^{g(t)^{-1}} \circ \dot{c}(t))). \tag{33}$$

Observe that $T_e\pi(\delta^l(g)(t)) = TA^{g(t)^{-1}}T\pi\,\dot{g}(t) = TA^{g(t)^{-1}}\dot{c}(t)$ since $\pi \circ g = c$. As ω_e is a section of $T_e\pi$, $T_e\pi(\delta^l(g)(t) - \omega_e(TA^{g(t)^{-1}} \circ \dot{c}(t)) = 0 \in T_{eH}\mathcal{M}$. Summing up

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\pi(\gamma(t)^{-1}g(t)) &\overset{(33)}{=} T\pi((\gamma(t)^{-1}g(t)) \cdot (\delta^l(g)(t) - \omega_e(TA^{g(t)^{-1}} \circ \dot{c}(t))) \\ &\overset{(8)}{=} TA^{\gamma(t)^{-1}g(t)}T_e\pi(\delta^l(g)(t) - \omega_e(TA^{g(t)^{-1}} \circ \dot{c}(t))) = 0.\end{aligned}$$

□

6.2 (A chart for the image of the SRVT) Let G be a Lie group with Lie algebra $\mathfrak{g}$. Using the adjoint action of G on $\mathfrak{g}$ and the evolution $\mathrm{Evol}\colon C^\infty(I,\mathfrak{g}) \to C^\infty(I,G)$, we define the map

$$\Psi\colon C^\infty(I,\mathfrak{g}) \to C^\infty(I,\mathfrak{g}), \quad q \mapsto -\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q,$$

where the dot denotes pointwise application of the linear map $\mathrm{Ad}(\mathrm{Evol}(q)^{-1})$. Observe that Ψ (co)restricts to a mapping $C^\infty(I,\mathfrak{g}\setminus\{0\}) \to C^\infty(I,\mathfrak{g}\setminus\{0\})$.

Lemma 27 *The map $\Psi\colon C^\infty(I,\mathfrak{g}) \to C^\infty(I,\mathfrak{g})$ is a smooth involution.*

Proof To establish smoothness of Ψ, consider the commutative diagram

$$\begin{array}{ccc} C^\infty(I,\mathfrak{g}) & \xrightarrow{\quad\Psi\quad} & C^\infty(I,\mathfrak{g}) \\ {\scriptstyle (\mathrm{Evol},\mathrm{id}_{C^\infty(I,\mathfrak{g})})}\downarrow & & \| \\ C^\infty(I,G)\times C^\infty(I,\mathfrak{g}) & \xrightarrow{(f,g)\mapsto \mathrm{Ad}(f).g} & C^\infty(I,\mathfrak{g}) \end{array}.$$

As $\mathrm{Ad}\colon G\times\mathfrak{g}\to\mathfrak{g}$ is smooth, so is $(f,g)\mapsto \mathrm{Ad}(f).g$ (cf. [13, Proof of Proposition 6.2]) and Ψ is smooth as a composition of smooth maps. Compute for $q \in C^\infty(I,\mathfrak{g})$

$$\begin{aligned}\Psi(\Psi(q)) &= -\mathrm{Ad}(\mathrm{Evol}(\Psi(q))^{-1}).\Psi(q) = -\mathrm{Ad}(\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q)^{-1}).(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q) \\ &= \mathrm{Ad}((\mathrm{Evol}(q)\,\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q))^{-1}).q.\end{aligned}$$

To see that $\Psi(\Psi(q)) = q$, we prove that

$$\gamma_q := \mathrm{Evol}(q)\,\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q)$$

is a constant path. Recall that $\mathrm{Evol}(q)$ and

$$\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q)$$

are smooth paths starting at the identity in G. Hence it suffices to prove $\delta^r(\gamma_q) = 0$. To this end, apply the product formula (32) and $\delta^r(\mathrm{Evol}(q)) = q$:

$$\begin{aligned}\delta^r(\gamma_q) &= \delta^r(\mathrm{Evol}(q)) + \mathrm{Ad}(\mathrm{Evol}(q)).\delta^r(\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q))\\ &= q + \mathrm{Ad}(\mathrm{Evol}(q)).(-\mathrm{Ad}(\mathrm{Evol}(q)^{-1}).q) = q - q = 0.\end{aligned}$$

□

To account for the initial point $c_0 \in \mathcal{M}$, fix $g_0 \in \pi^{-1}(c_0)$ and define

$$\Psi_{g_0}\colon C^\infty(I,\mathfrak{g}) \to C^\infty(I,\mathfrak{g}), \quad \Psi_{g_0}(q) := \mathrm{Ad}(g_0).\Psi(q) = -\mathrm{Ad}(g_0\,\mathrm{Evol}(q)^{-1}).q.$$

For k in the center of G, $\Psi_k = \Psi$ holds, but in general Ψ_{g_0} will not be an involution.

Lemma 28 *For each $g_0 \in G$, the map Ψ_{g_0} is a diffeomorphism with inverse $\Psi_{g_0^{-1}}$.*

Proof From the definition of Ψ_{g_0} and Lemma 27, it is clear that Ψ_{g_0} is a smooth diffeomorphism. We use that $\mathrm{Ad}\colon G \to \mathrm{GL}(\mathfrak{g})$ is a group morphism and compute

$$\begin{aligned}\Psi_{g_0^{-1}}(\Psi_{g_0}(q)) &= \mathrm{Ad}(g_0^{-1}).\Psi(\Psi_{g_0}(q)) = \mathrm{Ad}(g_0).\Psi(\mathrm{Ad}(g_0).\Psi(q))\\ &= \mathrm{Ad}(g_0^{-1}).\Big(-\mathrm{Ad}(\mathrm{Evol}(\mathrm{Ad}(g_0).\Psi(q))^{-1}).\,\mathrm{Ad}(g_0).\Psi(q)\Big)\\ &= -\mathrm{Ad}(g_0^{-1}g_0\,\mathrm{Evol}(\Psi(q))^{-1}g_0^{-1}g_0).\Psi(q) = \Psi(\Psi(q)) = q.\end{aligned}$$

Here we used that $\mathrm{Evol}(\mathrm{Ad}(g).f) = g\,\mathrm{Evol}(f)g^{-1}$, for $g \in G$. □

Lemma 29 *Fix $c_0 \in \mathcal{M}$ and choose $g_0 \in G$ with $\pi(g_0) = c_0$. Assume that $\mathcal{M}$ is reductive with $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, then*

$$\Psi_{g_0}(C^\infty(I,\mathfrak{m}\setminus\{0\})) = \{f \in C^\infty(I,\mathfrak{g}) \mid f = \theta_\omega(\dot c) \quad \textit{for some } c \in \mathrm{Imm}_{c_0}(I,\mathcal{M})\}.$$

With θ_ω as in 3.4 the formula $\theta_\omega \circ D(\rho_{c_0} \circ \Psi_{g_0}(q)) = \Psi_{g_0}(q)$ holds.

Proof Consider $c \in \mathrm{Imm}_{c_0}(I,\mathcal{M})$ and recall from Proposition 17 the identity $\Lambda_{c_0}(\mathrm{Evol}(\theta_\omega(\dot c))) = \pi(\mathrm{Evol}(\theta_\omega(\dot c))g_0) = c$. Choose $\hat c = \mathrm{Evol}(\theta_\omega(\dot c))g_0$ as a smooth lift of c to G and compute as follows:

$$\begin{aligned}\Psi_{g_0^{-1}}(\theta_\omega(\dot c)) &= \mathrm{Ad}(g_0^{-1}).\Big(-\mathrm{Ad}(\mathrm{Evol}(\theta_\omega(\dot c))^{-1}).(\theta_\omega(\dot c))\Big)\\ &= \mathrm{Ad}(g_0^{-1}).\Big(-\mathrm{Ad}(\mathrm{Evol}(\theta_\omega(\dot c))^{-1}).\,\mathrm{Ad}(\hat c).\omega_e(T\Lambda^{\hat c^{-1}}(\dot c))\Big)\end{aligned}$$

$$= \mathrm{Ad}(g_0^{-1}).\left(-\mathrm{Ad}(\mathrm{Evol}(\theta_\omega(\dot{c}))^{-1}).\,\mathrm{Ad}(\mathrm{Evol}(\theta_\omega(\dot{c}))g_0).\omega_e(T\Lambda^{\hat{c}^{-1}}(\dot{c}))\right)$$

$$= -\omega_e(T\Lambda^{\hat{c}^{-1}}(\dot{c})) \in \mathfrak{m}\setminus\{0\}.$$

Conversely, let us show that $\Psi_{g_0}(C^\infty(I,\mathfrak{m}\setminus\{0\}))$ is contained in the image of $\theta_\omega \circ D|_{\mathrm{Imm}_{c_0}(I,\mathcal{M})}$. To this end, consider $q = \Psi_{g_0}(v)$ for $v \in C^\infty(I,\mathfrak{m}\setminus\{0\})$. We compute

$$\begin{aligned}\rho_{c_0}(q) &= \Lambda_{c_0}(\mathrm{Evol}(\mathrm{Ad}(g_0).\Psi(v))) = \pi(\mathrm{Evol}(\mathrm{Ad}(g_0).\Psi(v)g_0))\\ &= \pi(g_0\,\mathrm{Evol}(\Psi(v))) = \pi(g_0\,\mathrm{Evol}(-\mathrm{Ad}(\mathrm{Evol}(v)^{-1}).v)) = \Lambda^{g_0}(\pi(\mathrm{Evol}(\Psi(v)))).\end{aligned} \tag{34}$$

Since Λ^{g_0} is a diffeomorphism, $\rho_{c_0}(q)\colon I \to \mathcal{M}$ is an immersion if and only if the curve $\pi(\mathrm{Evol}(\Psi(v)))$ has a non-vanishing derivative everywhere. Recall from the proof of Lemma 27 that $\mathrm{Evol}(v)\,\mathrm{Evol}(\Psi(v)) = e$, whence we compute the derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\pi(\mathrm{Evol}(\Psi(v))(t)) &= T\pi\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Evol}(\Psi(v))(t)\right) = T\pi(\Psi(v)\,\mathrm{Evol}(\Psi(v)))(t)\\ &= T\pi(-\mathrm{Ad}(\mathrm{Evol}(v)^{-1}).v\,\mathrm{Evol}(\Psi(v)))(t)\\ &= -T\pi\circ(L_{\mathrm{Evol}(v)^{-1}(t)})_*\circ(R_{\mathrm{Evol}(v)(t)\,\mathrm{Evol}(\Psi(v))(t)})_*(v(t))\\ &= -T\Lambda^{\mathrm{Evol}(v)^{-1}(t)}\circ T_e\pi(v(t)).\end{aligned} \tag{35}$$

In passing to the last line, we used that π commutes with the left action. Since $T\Lambda^g$ is an isomorphism, $\frac{\mathrm{d}}{\mathrm{d}t}\pi(\mathrm{Evol}(\Psi(v))(t))$ vanishes if and only if $v(t) \in \ker T_e\pi = \mathfrak{h}$. However, $v(t) \in \mathfrak{m}\setminus\{0\}$, whence $\rho_{c_0}(\Psi_{g_0}(v)) \in \mathrm{Imm}_{c_0}(I,\mathcal{M})$ and we can apply $\theta_\omega \circ D$ to $\rho_{c_0}(q)$. A combination of (34) and (35) yields $\frac{\mathrm{d}}{\mathrm{d}t}\rho_{c_0}(\Psi_{g_0}(v))(t) = -T\Lambda^{g_0(\mathrm{Evol}(v))^{-1}(t)}\circ T_e\pi(v(t))$. With $\pi(g_0\,\mathrm{Evol}(v)^{-1}) = \rho_{c_0}(\Psi_{g_0}(v)(t))$, this yields

$$\theta_\omega\left(\frac{\mathrm{d}}{\mathrm{d}t}\rho_{c_0}(v(t))\right) = \theta_\omega\left(\frac{\mathrm{d}}{\mathrm{d}t}\rho_{c_0}(\Psi_{g_0}(v))(t)\right) = \theta_\omega(-T\Lambda^{g_0(\mathrm{Evol}(v))^{-1}(t)}\circ T_e\pi(v(t)))$$

$$= \mathrm{Ad}(g_0(\mathrm{Evol}(v))^{-1}).\omega_e(-T\Lambda^{(g_0(\mathrm{Evol}(v))^{-1}(t))^{-1}}T\Lambda^{g_0(\mathrm{Evol}(v))^{-1}(t)}\circ T_e\pi(v(t)))$$

$$= -\mathrm{Ad}(g_0(\mathrm{Evol}(v))^{-1}).\omega_e(T_e\pi(v(t))) = -\mathrm{Ad}(g_0(\mathrm{Evol}(v))^{-1}).v(t) = \Psi_{g_0}(v)(t).$$

Note that as $\omega_e = (T_e\pi|_{\mathfrak{m}})^{-1}$, we have $\omega_e(T_e\pi(v(t))) = v(t)$. □

Acknowledgements This work was supported by the Norwegian Research Council, and by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie, grant agreement No. 691070.

References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2007)
2. Amiri, H., Schmeding, A.: A Differentiable Monoid of Smooth Maps on Lie Groupoids (2017). arXiv:1706.04816v1
3. Bastiani, A.: Applications différentiables et variétés différentiables de dimension infinie. J. Anal. Math. **13**, 1–114 (1964)
4. Bauer, M., Bruveris, M., Harms, P., Michor, P.W.: Vanishing geodesic distance for the Riemannian metric with geodesic equation the KdV-equation. Ann. Global Anal. Geom. **41**(4), 461–472 (2012)
5. Bauer, M., Bruveris, M., Michor, P.W.: Overview of the geometries of shape spaces and diffeomorphism groups. J. Math. Imaging Vis. **50**(1), 1–38 (2014)
6. Bauer, M., Bruveris, M., Marsland, S., Michor, P.W.: Constructing reparameterization invariant metrics on spaces of plane curves. Differ. Geom. Appl. **34**, 139–165 (2014)
7. Bauer, M., Eslitzbichler, M., Grasmair, M.: Landmark-guided elastic shape analysis of human character motions. Inverse Prob. Imaging **11**(4), 601–621 (2015). https://doi.org/10.3934/ipi.2017028
8. Bruveris, M.: Optimal reparametrizations in the square root velocity framework. SIAM J. Math. Anal. **48**(6), 4335–4354 (2016)
9. Celledoni, E., Eslitzbichler, M., Schmeding, A.: Shape analysis on Lie groups with applications in computer animation. J. Geom. Mech. **8**(3), 273–304 (2016)
10. Celledoni, E., Owren, B.: On the implementation of Lie group methods on the Stiefel manifold. Numer. Algorithm. **32**(2–4), 163–183 (2003)
11. Gallot, S., Hulin, D., Lafontaine, J.: Riemannian Geometry. Universitext, 3rd edn. Springer, Berlin (2004)
12. Glöckner, H.: Fundamentals of Submersions and Immersions Between Infinite-Dimensional Manifolds (2015). arXiv:1502.05795v3 [math]
13. Glöckner, H.: Regularity Properties of Infinite-Dimensional Lie Groups, and Semiregularity (2015). arXiv:1208.0715v3
14. Huper, K., Leite, F.: On the geometry of rolling and interpolation curves on S^n, SO_n, *and Grassmann manifolds.* J. Dyn. Control. Syst. **13**, 467–502 (2007)
15. Kriegl, A., Michor, P.W.: The convenient setting of global analysis. In: Mathematical Surveys and Monographs, vol. 53. American Mathematical Society, Providence
16. Kobayashi, S., Nomizu, K. Foundations of Differential Geometry, vol. II. Interscience Tracts in Pure and Applied Mathematics, no. 15, vol. II. Interscience Publishers John Wiley, New York/London/Sydney (1969)
17. Knapp, A.W.: Lie groups beyond an introduction. In: Progress in Mathematics, vol. 140, 2nd edn. Birkhäuser, Boston (2002)
18. Le Brigant, A.: Computing distances and geodesics between manifold-valued curves in the SRV framework. J. Geom. Mech. **9**(2) (2017). https://doi.org/10.3934/jgm.2017005
19. Michor, P.W.: Manifolds of Differentiable Mappings. In: Shiva Mathematics Series, vol. 3. Shiva Publishing Ltd., Nantwich (1980)
20. Munthe-Kaas, H., Verdier, O.: Integrators on homogeneous spaces: isotropy choice and connections. Found. Comput. Math. **16**(4), 899–939 (2016)
21. Michor, P.W., Mumford, D.: Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. Doc. Math. **10**, 217–245 (2005)
22. Michor, P.W., Mumford, D.: Riemannian geometries on spaces of plane curves. J. Eur. Math. Soc. (JEMS) **8**(1), 1–48 (2006)
23. Ortega, J.-P., Ratiu, T.S.: Momentum maps and Hamiltonian reduction. In: Progress in Mathematics, vol. 222. Birkhäuser Boston, Inc., Boston (2004)

24. Sharpe, R.W.: Differential geometry. In: Graduate Texts in Mathematics, vol. 166. Springer, New York (1997). Cartan's generalization of Klein's Erlangen program, With a foreword by S. S. Chern
25. Su, Z., Klassen, E., Bauer, M.: Comparing Curves in Homogeneous Spaces (2017). 1712.04586v1
26. Su, Z., Klassen, E., Bauer, M.: The square root velocity framework for curves in a homogeneous space. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 680–689. IEEE (2017)
27. Srivastava, A., Klassen, E., Joshi, S., Jermyn, I.: Shape analysis of elastic curves in euclidean spaces. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1415–1428 (2011)
28. Sebastian, T.B., Klein, P.N., Kimia, B.B.: On aligning curves. IEEE Trans. Pattern Anal. Mach. Intell. **25**(1), 116–125 (2003)
29. Su, J., Kurtek, S., Klassen, E., Srivastava, A.: Statistical analysis of trajectories on Riemmannian manifolds: bird migration, hurricane tracking and video surveillance. Ann. Appl. Stat. **8**(2), 530–552 (2014)

Universality in Numerical Computation with Random Data: Case Studies, Analytical Results and Some Speculations

Percy Deift and Thomas Trogdon

Abstract We discuss various universality aspects of numerical computations using standard algorithms. These aspects include empirical observations and rigorous results. We also make various speculations about computation in a broader sense.

1 Introduction

There are two natural "integrabilities" associated with matrices M. The first concerns random matrix theory where key statistics, such as the distribution of the largest eigenvalue of M, or the gap probability, i.e., the probability that the spectrum of M contains a gap of a given length, are described in an appropriate scaling limit as $N = \dim M \to \infty$, by the solution of completely integrable Hamiltonian systems, viz., Painlevé equations (see e.g. [8]). The second concerns the numerical computation of the eigenvalues of a matrix. Many standard eigenvalue algorithms work in the following way. Let Σ_N denote the set of real $N \times N$ symmetric matrices and let $M \in \Sigma_N$ be a given matrix whose eigenvalues one wants to compute. Associated with each algorithm $\mathscr{A}$, there is, in the discrete case,

The work in this paper was supported in part by DMS Grant #13000965 (P.D.) and DMS Grant #1303018 (T.T.).

P. Deift (✉)
Courant Institute, New York University, New York, NY, USA
e-mail: deift@cims.nyu.edu

T. Trogdon
University of California, Irvine, Irvine, CA, USA
e-mail: ttrogdon@math.uci.edu

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_8

a map $\varphi = \varphi_{\mathscr{A}} : \Sigma_N \to \Sigma_N$, with the properties

- (isospectral) $\operatorname{spec}\big(\varphi_{\mathscr{A}}(H)\big) = \operatorname{spec}(H), \qquad H \in \Sigma_N,$
- (convergence) the iterates $X_{k+1} = \varphi_{\mathscr{A}}(X_k), \quad k \geq 0, \qquad X_0 = M$, converge to a diagonal matrix $X_\infty, \quad X_k \to X_\infty, \quad \text{as} \quad k \to \infty,$

and in the continuum case, there is a flow $t \to X(t) \in \Sigma_N$ with the properties

- (isospectral) $\operatorname{spec}\big(X(t)\big) = \operatorname{spec}\big(X(0)\big),$
- (convergence) the flow $X(t)$, $t \geq 0$, $X(0) = M$, converges to a diagonal matrix X_∞, $X(t) \to X_\infty$ as $t \to \infty$.

In both case, necessarily the (diagonal) entries of X_∞ are the eigenvalues of the given matrix M. Now the fact of the matter is that, in most cases of interest, the flow $t \to X(t)$ is Hamiltonian and completely integrable in the sense of Liouville, and in the discrete case we have a "stroboscope theorem", i.e. there exists a completely integrable Hamiltonian flow $t \to \tilde{X}(t)$ which coincides with the above iterates X_k at integer times, $\tilde{X}(k) = X_k, \quad k \geq 0$ (see, in particular, [2, 4, 13]). The QR algorithm on full $N \times N$ matrices is a prime example of such a discrete algorithm, while the Toda algorithm is an example of the continuous case.

Question: What happens if one tries to "marry" these two integrabilities? In particular, what happens when one computes the eigenvalues of a random matrix? In response to this question, the authors in [11] initiated a statistical study of the performance of various standard algorithms to compute the eigenvalues of random matrices M from Σ_N.

Given $\epsilon > 0$, it follows, in the discrete case, that for some m the off-diagonal entries of X_m are[1] $O(\epsilon)$ and hence the diagonal entries of X_m give the eigenvalues of $X_0 = M$ to $O(\epsilon)$. The situation is similar for continuous algorithms $t \to X(t)$. Rather than running the algorithm until all the off-diagonal entries are $O(\epsilon)$, it is customary to run the algorithm with **deflations** as follows. For an $N \times N$ matrix Y in block from

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

with Y_{11} of size $k \times k$ and Y_{22} of size $(N-k) \times (N-k)$ for some $k \in \{1, 2, \ldots, N-1\}$, the process of projecting $Y \to \operatorname{diag}(Y_{11}, Y_{22})$ is called deflation. For a given $\epsilon > 0$, algorithm $\mathscr{A}$ and matrix $M \in \Sigma_N$, define the ***k*-deflation** time $T^{(k)}(M) = T^{(k)}_{\epsilon,\mathscr{A}}(M), \quad 1 \leq k \leq N-1$, to be the smallest value of

[1] For our purposes, a quantity X is $O(\epsilon)$ if $|X| \leq C\epsilon$ for a (possibly) N-dependent constant C if ϵ is sufficiently small.

m such that X_m, the mth iterate of algorithm $\mathscr{A}$ with $X_0 = M$, has block form

$$X_m = \begin{bmatrix} X_{11}^{(k)} & X_{12}^{(k)} \\ X_{21}^{(k)} & X_{22}^{(k)} \end{bmatrix}$$

with $X_{11}^{(k)}$ of size $k \times k$ and $X_{22}^{(k)}$ of size $(N-k) \times (N-k)$ and $\|X_{12}^{(k)}\| = \|X_{21}^{(k)}\| \leq \epsilon$. The deflation time $T(M)$ is then defined as

$$T(M) = T_{\epsilon,\mathscr{A}}(M) = \min_{1 \leq k \leq N-1} \quad T_{\epsilon,\mathscr{A}}^{(k)}(M).$$

If $\hat{k} \in \{1, \ldots, N-1\}$ is such that $T(M) = T_{\epsilon,\mathscr{A}}^{(\hat{k})}(M)$, it follows that the eigenvalues of $M = X_0$ are given by the eigenvalues of the block-diagonal matrix $\operatorname{diag}\left(X_{11}^{(\hat{k})}, X_{22}^{(\hat{k})}\right)$ to $O(\epsilon)$. After running the algorithm to time $T_{\epsilon,\mathscr{A}}(M)$, the algorithm restarts by applying the basic algorithm $\mathscr{A}$ separately to the smaller matrices $X_{11}^{(\hat{k})}$ and $X_{22}^{(\hat{k})}$ until the next deflation time, and so on. There are again similar considerations for continuous algorithms.

As the algorithm proceeds, the number of matrices after each deflation doubles. This is counterbalanced by the fact that the matrices are smaller and smaller in size, and the calculations are clearly parallelizable. Allowing for parallel computation, the number of deflations to compute all the eigenvalues of a given matrix M to an accuracy ϵ, will vary from $O(\log N)$ to $O(N)$.

In [11] the authors considered the deflation time $T = T_{\epsilon,\mathscr{A}} = T_{\epsilon,\mathscr{A},\mathscr{E}}$ for $N \times N$ matrices chosen from an ensemble $\mathscr{E}$. For a given $\epsilon > 0$, algorithm $\mathscr{A}$ and ensemble $\mathscr{E}$, the authors computed $T(M)$ for 5,000–10,000 samples of matrices M chosen from $\mathscr{E}$, and recorded the **normalized deflation time**

$$\tilde{T}(M) \equiv \frac{T(M) - \langle T \rangle}{\sigma} \tag{1}$$

where $\langle T \rangle$ and $\sigma^2 = \left\langle \left(T - \langle T \rangle\right)^2 \right\rangle$ are the sample average and sample variance of $T(M)$, respectively. What the authors found, surprisingly, was that for the given algorithm $\mathscr{A}$, and ϵ and N in a suitable scaling range with $N \to \infty$, the **histogram of $\tilde{T}$ was universal, independent of the ensemble** $\mathscr{E}$. In other words, the fluctuations in the deflation time $\tilde{T}$, suitably scaled, were universal, independent of $\mathscr{E}$. Figure 1 displays some of the numerical results from [11]. Figure 1a displays data for the QR algorithm, which is discrete, and Fig. 1b displays data for the Toda algorithm, which is continuous. Note that the histograms in Fig. 1a, b are very different: Universality is observed with respect to the ensembles $\mathscr{E}$—not with respect to the algorithms $\mathscr{A}$. The reason these particular histograms are different can be explained by the observation that the deflation time for the Toda algorithm is largely controlled by the largest gap in the spectrum of the matrix which typically occurs at the edge for our matrices. On the other hand, the QR algorithm biases towards

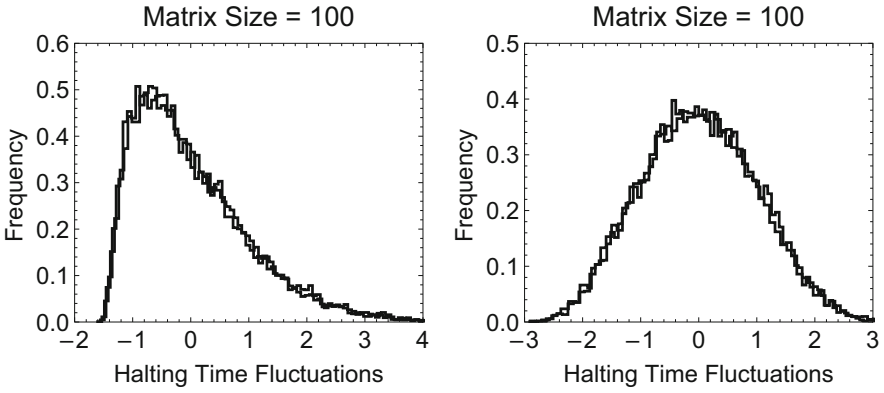


Fig. 1 Universality for $\tilde{T}$ when (**a**) $\mathscr{A}$ is the QR eigenvalue algorithm and when (**b**) $\mathscr{A}$ is the Toda algorithm. Panel (**a**) displays the overlay of two histograms for $\tilde{T}$ in the case of QR, one for each of the two ensembles $\mathscr{E} =$ BE, consisting of iid mean-zero Bernoulli random variables and $\mathscr{E} =$ GOE, consisting of iid mean-zero normal random variables. Here $\epsilon = 10^{-10}$ and $N = 100$. Panel (**b**) displays the overlay of two histograms for $\tilde{T}$ in the case of the Toda algorithm, and again $\mathscr{E} =$ BE or GOE. And here $\epsilon = 10^{-8}$ and $N = 100$

finding small eigenvalues first so that the statistics of the eigenvalues in the bulk of the spectrum control the deflation times.

Subsequently in [3] the authors raised the question of whether the universality results in [11] were limited to eigenvalue algorithms for real symmetric matrices, or whether they were present more generally in numerical computation. And indeed the authors in [3] found similar universality results for a wide variety of numerical algorithms, including

(a) other algorithms such as the QR algorithm with shifts,[2] the Jacobi eigenvalue algorithm, and also algorithms applied to complex Hermitian ensembles
(b) the conjugate gradient and GMRES algorithms to solve linear $N \times N$ systems $Hx = b$ with H and b random
(c) an iterative algorithm to solve the Dirichlet problem $\Delta u = 0$ in a random star-shaped region $\Omega \subset \mathbb{R}^2$ with random boundary data f on $\partial\Omega$
(d) a genetic algorithm to compute the equilibrium measure for orthogonal polynomials on the line.

In [3] the authors also discuss similar universality results obtained by Bakhtin and Correll [1] in a series of experiments with live participants recording

(e) decision making times for a specified task.

[2]The QR algorithm with shifts is the accelerated version of the QR algorithm that is used in practice.

Whereas (a) and (b) concern finite dimensional problems, (c) shows that universality is also present in problems that are genuinely infinite dimensional. And whereas (a), (b) and (c) concern, in effect, deterministic dynamical systems acting on random initial data, problem (d) shows that universality is also present in genuinely stochastic algorithms.

The demonstration of universality in problems (a)–(d) raises the following issue: Given a view commonly discussed by neuroscientists and psychologists that the human brain acts as a big computer with hardware and software (see, e.g., [7, 14] and the references therein) , one should be able to find evidence of universality in some neural computations. It is this issue that led the authors in [3] to the work of Bakhtin and Correll. In [1] each of the participants is shown a large number k of diagrams and then asked to make a decision about a particular geometric feature of each diagram. What is then recorded is the time it takes for the participant to reach his'r decision. Thus each participant produces k decision times t which are then centered and scaled as in (1) to obtain a normalized decision time

$$\tilde{t} = \frac{t - \langle t \rangle}{\sigma}. \tag{2}$$

The distribution of $\tilde{t}$ is then recorded in a histogram. Each of the participants produces such a histogram, and what is remarkable is that the histograms are, with a few exceptions, (essentially) the same. Furthermore, in [1], Bakhtin and Correll developed a Curie-Weiss-type statistical mechanical model for the decision process, and obtained a distribution function f_{BC} which agrees remarkably well with the (common) histogram obtained by the participants. We note that the model of Bakhtin and Correll involves a particular parameter, the spin flip intensity c_i. In [1] the authors made one particular choice for c_i. However, as shown in [3], if one makes various other choices for c_i, then one still obtains the same distribution f_{BC}. In other words, the Bakhtin-Correll model itself has an intrinsic universality. In an independent development Sagun, Trogdon and LeCun [12] considered, amongst other things, search times on Google$^{\text{TM}}$ for a large number of words in English and in Turkish. They then centered and scaled these times as in (1), (2) to obtain two histograms for normalized search times, one for English words and one for Turkish words. To their great surprise, both histograms were the same and, moreover, extremely well described by f_{BC}. So we are left to ponder the following puzzlement: Whatever the neural stochastics of the participants in the study in [1], and whatever the stochastics in the Curie-Weiss model, and whatever the mechanism in Google$^{\text{TM}}$'s search engine, a commonality is present in all three cases expressed through the single distribution function f_{BC}.

2 A Limit Theorem

All of the above results are numerical. In order to establish universality as a bona fide phenomenon in numerical analysis, and not just an artifact suggested, however strongly, by certain computations as above, P. Deift and T. Trogdon in [5] considered the Toda eigenvalue algorithm mentioned above. In place of the deflation time $T(M) = \min_{1\le k\le N-1} \quad T^{(k)}_{\epsilon,\mathscr{A}}(M)$, $\mathscr{A} =$ Toda algorithm, Deift and Trogdon used the 1-deflation time $T^{(1)}(M) = T^{(1)}_{\epsilon,\mathscr{A}}(M)$ as the stopping time for the algorithm. In other words, given $\epsilon > 0$ and an ensemble $\mathscr{E}$, they ran the Toda algorithm $t \to X(t)$ with $X(0) = M \in \mathscr{E}$, until a time t where

$$t = T^{(1)}(M) = \inf\left\{ s \ge 0 : \sum_{j=2}^{N} \left(X_{1j}(s)\right)^2 \le \epsilon^2 \right\}.$$

It follows by perturbation theory that $\left| X_{11}\left(T^{(1)}(M)\right) - \lambda_{j^*}(M) \right| \le \epsilon$ for some eigenvalue $\lambda_{j^*}(M)$ of M. But the Toda algorithm is known to be ordering, i.e. $X(t) \to X_\infty = \operatorname{diag}\left(\lambda_1(M), \lambda_2(M), \ldots \lambda_N(M)\right)$, where the eigenvalues of M are ordered, $\lambda_1(M) \ge \lambda_2(M) \ge \cdots \ge \lambda_N(M)$. It follows then that (for ϵ sufficiently small and $T^{(0)}_{\epsilon,\mathscr{A}}$ correspondingly large) $j^* = 1$ so that the Toda algorithm with stopping time $T^{(1)} = T^{(1)}_{\epsilon,\mathscr{A}}$ computes the largest eigenvalue of M to accuracy ϵ with high probability.

The main result in [5] is the following. For invariant and generalized Wigner random matrix ensembles[3] there is an ensemble dependent constant $c_{\mathscr{E}}$ such that the following limit exists (see [10] and [15])

$$F^{\text{gap}}_{\beta}(t) = \lim_{N\to\infty} \text{Prob}\left(\frac{1}{c^{2/3}_{\mathscr{E}}\, 2^{-2/3}\, N^{2/3}\, (\lambda_1 - \lambda_2)} \le t \right), \qquad t \ge 0. \tag{3}$$

Here $\beta = 1$ for the real symmetric case, $\beta = 2$ for the complex Hermitian case. Thus $F^{\text{gap}}_{\beta}(t)$ is the distribution function for the (inverse of the) gap $\lambda_1 - \lambda_2$ between the largest two eigenvalues of M, on the appropriate scale as $N \to \infty$.

Theorem 1 (Universality for $T^{(1)}$) *Let $0 < \sigma < 1$ be fixed and let (ϵ, N) be in the scaling region*

$$\frac{\log \epsilon^{-1}}{\log N} \ge \frac{5}{3} + \frac{\sigma}{2}. \tag{4}$$

[3] See Appendix A in [5] for a precise description of the matrix ensembles considered in Theorem 1.

Then if M is distributed according to any real ($\beta = 1$) or complex ($\beta = 2$) invariant or Wigner ensemble, we have

$$\lim_{N\to\infty} \text{Prob}\left(\frac{T^{(1)}}{c_{\mathscr{E}}^{2/3}\ 2^{-2/3}\ N^{2/3}\left(\log \epsilon^{-1} - \frac{2}{3}\ \log N\right)} \le t \right) = F_{\beta}^{gap}. \tag{5}$$

Here $c_{\mathscr{E}}$ is the same constant as in (3).

This result establishes universality rigorously for a numerical algorithm of interest, viz., the Toda algorithm with stopping time $T^{(1)}$ to compute the largest eigenvalue of a random matrix. We see, in particular, that $T^{(1)}$ behaves statistically as the inverse of the top gap $\lambda_1 - \lambda_2$, on the appropriate scale as $N \to \infty$. Similar results have now been obtained for the QR algorithm and related algorithms acting on ensembles of strictly positive definite matrices (see [6]).

Remark 1 We point out that Theorem 1 could, in principle, give a robust statistical estimate of the expected run time in the same way that the classical Central Limit Theorem is used to give confidence levels for estimates in elementary statistics. In particular the "3-sigma" confidence level derived from the bell curve, would be replaced by a (possibly different) confidence level derived from F_{β}^{gap}.

Remark 2 Theorem 1 depends on the matrices being distributed according to an unstructured Wigner or invariant ensemble. If the matrices had structured form $M = D + W$ where D is given and deterministic, and W is random, then we would again expect universality for the runtime fluctuations with respect to the choice of ensemble for W. A priori, the histogram would be different from the histogram for the unstructured case, but one would still have universality with respect to W.

However, it turns out that in some cases of interest, the effect of W overwhelms the deterministic structure, and the histogram is the same as in the unstructured case. We recall that at the very beginning of the introduction of random matrix theory into theoretical physics, Wigner postulated, with remarkable success, that the resonances of neutron scattering off a U^{238} nucleus were described by the eigenvalues of a random matrix. In other words, though we might view the uranium nucleus as a system with structure and randomness, the structure is wiped out by the randomness. In the experiments in [3], the authors found a similar phenomenon. Indeed, it turns out that the halting time for the GMRES algorithm gives the same histogram for the fluctuations for unstructured systems $Mx = b$ ($M = I + X$, where X is iid) as it does when it comes from a discretization of the Dirichlet problem on a random star-shaped domain. In terms of the double layer potential method, the Dirichlet problem in a random domain has the form "structure + random" and so we again have a situation where a random, structured system is modeled by a completely random one.

The proof of Theorem 1 depends critically on the integrability of the Toda flow $t \to X(t), \quad X(0) = M$. The evolution of $X(t)$ is governed by the Lax-pair equation

$$\frac{dX}{dt} = \big[X,\ B(X)\big] = X\,B(X) - B(X)\,X$$

where $B(X) = X_- - X_-^T$ and X_- is the strictly lower triangular part of X. Using results of J. Moser [9] one finds that

$$E(t) \equiv \sum_{k=2}^{N} \big|X_{1k}(t)\big|^2 = \sum_{j=1}^{N} \big(\lambda_j - X_{11}(t)\big)^2 \big|u_{1j}(t)\big|^2 \tag{6}$$

$$X_{11}(t) = \sum_{j=1}^{N} \lambda_j \big|u_{1j}(t)\big|^2 \tag{7}$$

$$u_{1j}(t) = \frac{u_{1j}(0)\, e^{\lambda_j t}}{\left(\sum_{k=1}^{N} \big|u_{1k}(0)\big|^2\, e^{2\lambda_k t}\right)^{\frac{1}{2}}}, \qquad 1 \le j \le N, \tag{8}$$

where $u_{1j}(t)$ is the first component of the normalized eigenvector $u_j(t)$ for $X(t)$ corresponding to the eigenvalue $\lambda_j(t) = \lambda_j(0)$ of $X(t)$, $\big(X(t) - \lambda_j(t)\big)\,u_j(t) = 0$. (Note that $t \to X(t)$ is isospectral, so $\mathrm{spec}(X(t)) = \mathrm{spec}(X(0)) = \mathrm{spec}(M)$.) The stopping time $T^{(1)}$ is obtained by solving the equation

$$E(t) = \epsilon^2 \tag{9}$$

for t. Substituting (7) and (8) into (6) we obtain an formula for $E(t)$ involving only the eigenvalues and (the moduli of) the first components of the normalized eigenvectors for $X(0) = M$. It is this explicit formula that the Toda algorithm brings as a gift to the marriage announced earlier of eigenvalue algorithms and random matrices. What random matrix theory brings to the marriage is an impressive collection of very detailed estimates on the statistics of the λ_j's and the $u_{1j}(0)$'s obtained in recent years by a veritable army of researchers including P. Bourgade, L. Erdős, A. Knowles, J. A. Ramírez, B. Rider, B. Virág, T. Tao, V. Vu, J. Yin and H. T. Yau, amongst many others (see [5] and the references therein for more details).

Theorem 1 is a first step towards proving universality for the Toda algorithm with full deflation stopping time $T = T_{\epsilon,\mathscr{A}}$. The analysis of $T_{\epsilon,\mathscr{A}}$ involves very detailed information about the joint statistics of the eigenvalues λ_j and all the components u_{ij} of the normalized eigenvectors of $X(0) = M$, as $N \to \infty$. Such information is not yet known and the analysis of $T_{\epsilon,\mathscr{A}}$ is currently out of reach.

3 Speculations

How should one view the various **two-component** universality results described in this paper? "Two-components" refers to the fact for a random system of size S, say, and halting time T, once the average $\langle T \rangle$ and variance $\sigma^2 = \left\langle \left(T - \langle T \rangle\right)^2 \right\rangle$ are known, the normalized time $\tau = \left(T - \langle T \rangle\right)/\sigma$ is, in the large S limit, universal, independent of the ensemble, i.e. as $S \to \infty$, $T \sim \langle T \rangle + \sigma\, \chi$, where χ is universal. The best known two-component universality theorem is certainly the classical Central Limit Theorem, already mentioned in Remark 1 above: Suppose $Y_1, Y_2, \ldots$ are independent, identically distributed variables with mean μ and variance σ^2. Set $W_n \equiv \sum_{i=1}^{n} Y_i$. Then as $n \to \infty$, $\left(W_n - \langle W_n \rangle\right)/\sigma_n$ converges in distribution to a standard normal $N(0, 1)$, where $\langle W_n \rangle = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = n\mu$ and $\sigma_n^2 = \mathbb{E}\left(\left(W_n - \langle W_n \rangle\right)^2\right) = n\sigma^2$. In words: As $n \to \infty$, the only specific information about the initial distribution of the Y_i's that remains, is the mean μ and the variance σ.

Now, for a moment, set aside histograms for halting times, and imagine you are walking on the boardwalk in some seaside town. Along the way you pass many palm trees. But what do you mean by a "palm tree"? Some are taller, some are shorter, some are bushier, some are less bushy. Nevertheless you recognize them all as "palm trees": Somehow you adjust for the height and you adjust for the bushiness (two components!), and then draw on some internal data base to determine, with high certainty, that the object one is looking at is a "palm tree". The database itself catalogs/summarizes your learning experience with palm trees over many years. It is tempting to speculate that the data base has the form of a histogram. We have in our brains one histogram for palm trees, and another for olive trees, and so on. Then just as we may use a t-test, for example, to test the statistical properties of some sample, so too one speculates that there is a mechanism in one's mind that tests against the "palm tree histogram" and evaluates the likelihood that the object at hand is a palm tree. So in this way of thinking, there is no ideal Platonic object that is a "palm tree": Rather, a palm tree is a histogram.

One may speculate further in the following way. Just imagine if we perceived every palm tree as a distinct species, and then every olive tree as a distinct species, and so on. Working with such a plethora of data, would require access to an enormous bandwidth. From this point of view, the histogram provides a form of "stochastic data reduction", and the fortunate fact is that we have evolved to the point that we have just enough bandwidth to accommodate and evaluate the information "zipped" into the histogram. On the other hand, fortunately, the information in the histogram is sufficiently detailed that we can make meaningful distinctions, and one may speculate that it is precisely this balance between data reduction and bandwidth that is the key to our ability to function successfully in the macroscopic world.

We note finally that there are many similarities between the above speculations and machine learning. In both processes there is a learning phase followed by a

recognition phase. Also, in both cases, there is a balance between data reduction and bandwidth. In the case of the palm trees, etc., however we make the additional assertion/speculation that the stored data is in the form of a histogram, similar in origin to the universal histograms observed in numerical computations.

Now, returning to histograms for halting times, do not mean to suggest that there is a direct correspondence between the histogram which we postulate to be associated with an object and a histogram for halting times. Rather, our point of view is that these histograms are two different manifestations of a deeper form of universality acheived in both cases by a process of stochastic data reduction.

We may summarize the above discussion and speculations in the following way. In a common view, the brain is a computer, with software and hardware, which makes calculations and runs algorithms which reduce data on an appropriate scale—the macroscopic scale on which we live—to a manageable and useful form, viz., a histogram, which is universal[4] for all palm trees, or all olive trees, etc. With this in mind, it is tempting to suggest that **whenever we run an algorithm with random data on a "computer", two-component universal features will emerge on some appropriate scale.** This "computer" could be the electronic machine on our desk, or it could be the device in our mind that runs algorithms to classify random visual objects or to make timed decisions about geometric shapes, or it could be in any of the myriad of ways in which computations are made. Perhaps this is how one should view the various universality results described in this paper.

Acknowledgements One of the authors (P.D.) would like to thank the organizers for the invitation to participate in the Abel Symposium 2016 "Computation and Combinatorics in Dynamics, Stochastics and Control". During the symposium he gave a talk on a condensed version of the paper below.

References

1. Bakhtin, Y., Correll, J.: A neural computation model for decision-making times. J. Math. Psychol. **56**, 333–340 (2012)
2. Deift, P., Li, L.C., Nanda, T., Tomei, C.: The Toda flow on a generic orbit is integrable. Commun. Pure Appl. Math. **39**(2), 183–232 (1986)
3. Deift, P., Menon, G., Olver, S., Trogdon, T.: Universality in numerical computations with random data. Proc. Natl. Acad. Sci. USA **111**(42):14973–14978 (2014)
4. Deift, P., Nanda, T., Tomei, C.: Ordinary differential equations and the symmetric eigenvalue problem. SIAM J. Num. Anal. **20**, 1–22 (1983)

[4]A priori the histogram for a palm tree in one person's mind may be very different from that in another person's mind. Yet the results of Bakhtin and Correll in [1], where the participants produce the same decision time distributions, indicate that this is not so. And indeed, if there was a way to show that the histograms individuals form to catalog a palm tree, say, were all the same, this would have the following implication: The palm tree has an objective existence, and not a subjective one, which varies from person to person.

5. Deift, P., Trogdon, T.: Universality for the Toda algorithm to compute the largest eigenvalue of a random matrix. Commun. Pure Appl. Math. arXiv:1604.07384
6. Deift, P., Trogdon, T.: Universality for eigenvalue algorithms on sample covariance matrices. arXiv:1701.01896
7. Markus, G.: Face it, your brain is a computer. New York Times, 27 June 2015
8. Mehta, M.L.: Random matrices, 3rd edn. Elsevier, Amsterdam, 2004
9. Moser, J.: Three integrable Hamiltonian systems connected with isospectral deformations. Adv. Math. **16**(2), 197–220 (1975)
10. Perret, A., Schehr. G.: Near-extreme eigenvalues and the first gap of Hermitian random matrices. J. Stat. Phys. **156**(5), 843–876 (2014)
11. Pfrang, C.W., Deift, P., Menon, G.: How long does it take to compute the eigenvalues of a random symmetric matrix? Random Matrix Theory Interact. Part. Syst. Integr. Syst. MSRI Publ. **65**, 411–442 (2014)
12. Sagun, L., Trogdon, T., LeCun, Y.: Universal halting times in optimization and machine learning. arXiv:1511.06444
13. Symes, W.W.: The QR algorithm and scattering for the finite nonperiodic Toda lattice. Phys. D Nonlinear Phenom. **4**(2), 275–280 (1982)
14. University of Colorado at Boulder: Human brain region functions like a digital computer. Science Daily, 6 Oct 2006
15. Witte, N.S., Bornemann, F., Forrester, P.J.: Joint distribution of the first and second eigenvalues at the soft edge of unitary ensembles. Nonlinearity **26**(6), 1799–1822 (2013)

BSDEs with Default Jump

Roxana Dumitrescu, Miryana Grigorova, Marie-Claire Quenez, and Agnès Sulem

Abstract We study (nonlinear) Backward Stochastic Differential Equations (BSDEs) driven by a Brownian motion and a martingale attached to a default jump with intensity process $\lambda = (\lambda_t)$. The driver of the BSDEs can be of a generalized form involving a singular *optional* finite variation process. In particular, we provide a comparison theorem and a strict comparison theorem. In the special case of a generalized λ-*linear* driver, we show an explicit representation of the solution, involving conditional expectation and an adjoint exponential semimartingale; for this representation, we distinguish the case where the singular component of the driver is predictable and the case where it is only optional. We apply our results to the problem of (nonlinear) pricing of European contingent claims in an imperfect market with default. We also study the case of claims generating intermediate cashflows, in particular at the default time, which are modeled by a singular *optional* process. We give an illustrating example when the seller of the European

M. Grigorova acknowledges financial support from the German Science Foundation through SFB1283.

R. Dumitrescu
Department of Mathematics, King's College London, London, UK
e-mail: roxana.dumitrescu@kcl.ac.uk

M. Grigorova
Centre for Mathematical Economics, University Bielefeld, Bielefeld, Germany

M.-C. Quenez
Laboratoire de probabilités, statistiques et modélisations (CNRS/Sorbonne Université/Université Paris Diderot), Université Paris 7, Paris, France
e-mail: quenez@lpsm.paris

A. Sulem (✉)
INRIA Paris, Paris Cedex 12, France

Université Paris-Est, Champs-sur-Marne, France
e-mail: agnes.sulem@inria.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_9

option is a large investor whose portfolio strategy can influence the probability of default.

1 Introduction

The aim of the present paper is to study BSDEs driven by a Brownian motion and a compensated default jump process with intensity process $\lambda = (\lambda_t)$. The applications we have in mind are the pricing and hedging of contingent claims in an imperfect financial market with default. The theory of BSDEs driven by a Brownian motion and a Poisson random measure has been developed extensively by several authors (cf., e.g., Barles, Buckdahn and Pardoux [2], Royer [22], Quenez and Sulem [21], Delong [10]). Several of the arguments used in the present paper are similar to those used in the previous literature. Nevertheless, it should be noted that BSDEs with a default jump do not correspond to a particular case of BSDEs with Poisson random measure. The treatment of BSDEs with a default jump requires some specific arguments and we present here a complete analysis of these BSDEs, which is particularly useful in default risk modeling in finance. To our knowledge, there are few works on nonlinear BSDEs with default jump. The papers [6] and [1] are concerned only with the existence and the uniqueness of the solution, which are established under different assumptions. In this paper, we first provide some a priori estimates, from which the existence and uniqueness result directly follows. Moreover, we allow the driver of the BSDEs to have a singular component, in the sense that the driver is allowed to be of the generalized form $g(t, y, z, k)dt + dD_t$, where D is an optional (not necessarily predictable) right-continuous left-limited (RCLL) process with finite variation. We stress that the case of a singular *optional* process D has not been considered in the literature on BSDEs, even when the filtration is associated with a Brownian motion and a Poisson random measure. Moreover, these BSDEs are useful to study the nonlinear pricing problem in imperfect markets with default. Indeed, in this type of markets, the contingent claims often generate intermediate cashflows – in particular at the default time – which can be modeled via an optional singular process D (see e.g. [3, 5, 7, 8]). We introduce the definition of a λ-*linear* driver, where λ refers to the intensity of the jump process, which generalizes the notion of a linear driver given in the literature on BSDEs to the case of BSDEs with default jump and *generalized driver*. When g is λ-*linear*, we provide an explicit solution of the BSDE associated with the *generalized* λ-*linear driver* $g(t, y, z, k)dt + dD_t$ in terms of a conditional expectation and an adjoint exponential semimartingale. We note that this representation formula depends on whether the singular process D is predictable or just optional. Under some suitable assumptions on g, we establish a comparison theorem, as well as a strict comparison theorem. We emphasize that these comparison results are shown for optional (not necessarily) predictable singular processes, which requires some specific arguments. We then give an application in mathematical finance. We consider a financial market with a defaultable risky asset and we study the problems

of pricing and hedging of a European option paying a payoff ξ at maturity T and intermediate dividends (or cashflows) modeled by a singular process D. The option possibly generates a cashflow at the default time, which implies that the dividend process D is not necessarily predictable. We study the case of a market with imperfections which are expressed via the nonlinearity of the wealth dynamics. Our framework includes the case of different borrowing and lending treasury rates (see e.g. [17] and [8]) and "repo rates",[1] which is usual for contracts with intermediate dividends subjected to default (see [7]). We show that the price of the option is given by $X^g_{\cdot,T}(\xi, D)$, where $X^g_{\cdot,T}(\xi, D)$ is the solution of the nonlinear BSDE with default jump (solved under the primitive probability measure P) with *generalized driver* $g(t, y, z, k)dt + dD_t$, terminal time T and terminal condition ξ. This leads to a non linear pricing system $(\xi, D) \mapsto X^g_{\cdot,T}(\xi, D)$, for which we establish some properties. We emphasize that the monotonicity property (resp. no arbitrage property) requires some specific assumptions on the driver g, which are due to the presence of the default. Furthermore, for each driver g and each (fixed) singular process D, we define the (g, D)-conditional expectation by $\mathscr{E}^{g,D}_{t,T}(\xi) := X^g_{t,T}(\xi, D)$, for $\xi \in L^2(\mathscr{G}_T)$. In the case where $D = 0$, it reduces to the g-conditional expectation $\mathscr{E}^g$ (in the case of default). We also introduce the notion of $\mathscr{E}^{g,D}$-martingale, which is a useful tool in the study of nonlinear pricing problems: more specifically, those of American options and game options with intermediate dividends (cf. [13, 14]).

The paper is organized as follows: in Sect. 2, we present the properties of BSDEs with default jump and *generalized driver*. More precisely, in Sect. 2.1, we present the mathematical setup. In Sect. 2.2, we state some a priori estimates, from which we derive the existence and the uniqueness of the solution. In Sect. 2.3, we show the representation property of the solution of the BSDE associated with the *generalized driver* $g(t, y, z, k)dt + dD_t$ in the particular case when g is λ-linear. We distinguish the two cases: the case when the singular process D is predictable and the case when it is just optional. In Sect. 2.4, we establish the comparison theorem and the strict comparison theorem. Section 3 is devoted to the application to the nonlinear pricing of European options with dividends in an imperfect market with default. The properties of the nonlinear pricing system as well as those of the (g, D)-conditional expectation are also studied in this section. As an illustrative example of market imperfections, we consider the case when the seller of the option is a large investor whose hedging strategy (in particular the cost of this strategy) has impact on the default probability.

[1] Which can be seen as securities lending or borrowing rates in a "repo market" (cf. [7]).

2 BSDEs with Default Jump

2.1 Probability Setup

Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a complete probability space equipped with two stochastic processes: a unidimensional standard Brownian motion W and a jump process N defined by $N_t = \mathbf{1}_{\vartheta \leq t}$ for any $t \in [0, T]$, where ϑ is a random variable which models a default time. We assume that this default can appear after any fixed time, that is $P(\vartheta \geq t) > 0$ for any $t \geq 0$. We denote by $\mathbb{G} = \{\mathcal{G}_t, t \geq 0\}$ the *augmented filtration* generated by W and N (in the sense of [9, IV-48]). In the following, $\mathcal{P}$ denotes the $\mathbb{G}$-predictable σ-algebra on $\Omega \times [0, T]$. We suppose that W is a $\mathbb{G}$-Brownian motion.

Let (Λ_t) be the $\mathbb{G}$-predictable compensator of the non decreasing process (N_t). Note that $(\Lambda_{t\wedge\vartheta})$ is then the $\mathbb{G}$-predictable compensator of $(N_{t\wedge\vartheta}) = (N_t)$. By uniqueness of the $\mathbb{G}$-predictable compensator, $\Lambda_{t\wedge\vartheta} = \Lambda_t$, $t \geq 0$ a.s. We assume that Λ is absolutely continuous w.r.t. Lebesgue's measure, so that there exists a nonnegative $\mathbb{G}$-predictable process (λ_t), called the intensity process, such that $\Lambda_t = \int_0^t \lambda_s ds$, $t \geq 0$. Since $\Lambda_{t\wedge\vartheta} = \Lambda_t$, the process λ vanishes after ϑ. We denote by M the $\mathbb{G}$-compensated martingale given by

$$M_t = N_t - \int_0^t \lambda_s ds\,. \tag{1}$$

Let $T > 0$ be the finite horizon. We introduce the following sets:

- $\mathcal{S}^2_T$ (also denoted by $\mathcal{S}^2$) is the set of $\mathbb{G}$-adapted right-continuous left-limited (RCLL) processes φ such that $\mathbb{E}[\sup_{0\leq t\leq T} |\varphi_t|^2] < +\infty$.
- $\mathcal{A}^2_T$ (also denoted by $\mathcal{A}^2$) is the set of real-valued finite variational RCLL $\mathbb{G}$-adapted (thus *optional*) processes A with square integrable total variation process and such that $A_0 = 0$.
- $\mathcal{A}^2_{p,T}$ (also denoted by $\mathcal{A}^2_p$) is the set of *predictable* processes belonging to $\mathcal{A}^2$.
- $\mathbb{H}^2_T$ (also denoted by $\mathbb{H}^2$) is the set of $\mathbb{G}$-predictable processes with $\|Z\|^2 := \mathbb{E}\Big[\int_0^T |Z_t|^2 dt\Big] < \infty$.
- $\mathbb{H}^2_{\lambda,T} := L^2(\Omega \times [0, T], \mathcal{P}, \lambda_t\, dP \otimes dt)$ (also denoted by $\mathbb{H}^2_\lambda$), equipped with scalar product $\langle U, V\rangle_\lambda := \mathbb{E}\Big[\int_0^T U_t V_t \lambda_t dt\Big]$, for all U, V in $\mathbb{H}^2_\lambda$. For all $U \in \mathbb{H}^2_\lambda$, we set $\|U\|^2_\lambda := \mathbb{E}\Big[\int_0^T |U_t|^2 \lambda_t dt\Big]$

For each $U \in \mathbb{H}^2_\lambda$, we have $\|U\|^2_\lambda = \mathbb{E}\left[\int_0^{T\wedge\vartheta} |U_t|^2\lambda_t dt\right]$ because the $\mathbb{G}$-intensity λ vanishes after ϑ. Note that, without loss of generality, we may assume that U vanishes after ϑ.[2]

Moreover, $\mathscr{T}$ is the set of stopping times τ such that $\tau \in [0, T]$ a.s. and for each S in $\mathscr{T}$, $\mathscr{T}_S$ is the set of stopping times τ such that $S \leq \tau \leq T$ a.s.

We recall the martingale representation theorem in this framework (see [18]):

Lemma 1 (Martingale representation) *Let $m = (m_t)_{0\leq t\leq T}$ be a $\mathbb{G}$-local martingale. There exists a unique pair of $\mathbb{G}$-predictable processes (z_t, l_t)[3] such that*

$$m_t = m_0 + \int_0^t z_s dW_s + \int_0^t l_s dM_s\,, \quad \forall\, t \in [0, T] \quad a.s. \tag{2}$$

If m is a square integrable martingale, then $z \in \mathbb{H}^2$ and $l \in \mathbb{H}^2_\lambda$.

We now introduce the following definitions.

Definition 1 (Driver, λ-*admissible* driver)

- A function g is said to be a *driver* if $g : \Omega \times [0, T] \times \mathbf{R}^3 \to \mathbf{R}$; $(\omega, t, y, z, k) \mapsto g(\omega, t, y, z, k)$ is $\mathscr{P} \otimes \mathscr{B}(\mathbf{R}^3)-$ measurable, and such that $g(., 0, 0, 0) \in \mathbb{H}^2$.
- A driver g is called a λ-*admissible driver* if moreover there exists a constant $C \geq 0$ such that for $dP \otimes dt$-almost every (ω, t) , for all (y_1, z_1, k_1), (y_2, z_2, k_2),

$$|g(\omega, t, y_1, z_1, k_1) - g(\omega, t, y_2, z_2, k_2)| \leq C(|y_1 - y_2| + |z_1 - z_2| + \sqrt{\lambda_t(\omega)}|k_1 - k_2|). \tag{3}$$

  A non negative constant C such that (3) holds is called a λ-*constant* associated with driver g.

Note that condition (3) implies that for each (y, z, k), we have $g(t, y, z, k) = g(t, y, z, 0), t > \vartheta\ dP \otimes dt$- a.e. Indeed, on the set $\{t > \vartheta\}$, g does not depend on k, since $\lambda_t = 0$.

Remark 1 Note that a *driver* g supposed to be Lipschitz with respect to (y, z, k) is not generally λ-admissible. Moreover, a *driver* g supposed to be λ-admissible is not generally Lipschitz with respect to (y, z, k) since the process (λ_t) is not necessarily bounded.

Definition 2 Let g be a λ-*admissible* driver, let $\xi \in L^2(\mathscr{G}_T)$.

- A process (Y, Z, K) in $\mathscr{S}^2 \times \mathbb{H}^2 \times \mathbb{H}^2_\lambda$ is said to be a solution of the BSDE with default jump associated with terminal time T, driver g and terminal condition ξ

[2] Indeed, each U in $\mathbb{H}^2_\lambda (= L^2(\Omega \times [0, T], \mathscr{P}, \lambda_t dP \otimes dt))$ can be identified with $U\mathbf{1}_{t\leq\vartheta}$, since $U\mathbf{1}_{t\leq\vartheta}$ is a $\mathbb{G}$-predictable process satisfying $U_t\mathbf{1}_{t\leq\vartheta} = U_t\ \ \lambda_t dP \otimes dt$-a.s.

[3] Such that the stochastic integrals in (2) are well defined.

if it satisfies:

$$-dY_t = g(t, Y_t, Z_t, K_t)dt - Z_t dW_t - K_t dM_t; \quad Y_T = \xi. \tag{4}$$

- Let $D \in \mathscr{A}^2$. A process (Y, Z, K) in $\mathscr{S}^2 \times \mathbb{H}^2 \times \mathbb{H}^2_\lambda$ is said to be a solution of the BSDE with default jump associated with terminal time T, *generalized* λ-*admissible driver* $g(t, y, z, k)dt + dD_t$ and terminal condition ξ if it satisfies:

$$-dY_t = g(t, Y_t, Z_t, K_t)dt + dD_t - Z_t dW_t - K_t dM_t; \quad Y_T = \xi. \tag{5}$$

Remark 2 Let $D = (D_t)_{0 \le t \le T}$ be a finite variational RCLL adapted process such that $D_0 = 0$, and with integrable total variation. We recall that D admits at most a countable number of jumps. We also recall that the process D has the following (unique) canonical decomposition: $D = A - A'$, where A and A' are integrable non decreasing RCLL adapted processes with $A_0 = A'_0 = 0$, and such that dA_t and dA'_t are mutually singular (cf. Proposition A.7 in [11]). If D is predictable, then A and A' are predictable.

Moreover, by a property given in [14], for each $D \in \mathscr{A}^2$, there exist a unique (predictable) process D' belonging to $\mathscr{A}^2_p$ and a unique (predictable) process η belonging to $I\!H^2_\lambda$ such that for all $t \in [0, T]$,

$$D_t = D'_t + \int_0^t \eta_s dN_s \quad \text{a.s.}$$

If D is non decreasing, then D' is non decreasing and $\eta_\vartheta \ge 0$ a.s on $\{\vartheta \le T\}$.

Remark 3 By Remark 2 and Eq. (5), the process Y admits at most a countable number of jumps. It follows that $Y_t = Y_{t^-}$, $0 \le t \le T$ $dP \otimes dt$-a.e. Moreover, we have $g(t, Y_t, Z_t, K_t) = g(t, Y_{t^-}, Z_t, K_t)$, $0 \le t \le T$ $dP \otimes dt$-a.e.

2.2 Properties of BSDEs with Default Jump

We first show some a priori estimates for BSDEs with a default jump, from which we derive the existence and uniqueness of the solution. For $\beta > 0$, $\phi \in I\!H^2$, and $l \in I\!H^2_\lambda$, we introduce the norms $\|\phi\|^2_\beta := \mathbb{E}[\int_0^T e^{\beta s}\phi_s^2 ds]$, and $\|k\|^2_{\lambda,\beta} := \mathbb{E}[\int_0^T e^{\beta s}k_s^2\lambda_s\, ds]$.

2.2.1 A Priori Estimates for BSDEs with Default Jump

Proposition 1 *Let ξ^1, $\xi^2 \in L^2(\mathscr{G}_T)$. Let g^1 and g^2 be two* λ-admissible *drivers. Let C be a λ-constant associated with g^1. Let D be an (optional) process belonging to $\mathscr{A}^2$.*
For $i = 1, 2$, let (Y^i, Z^i, K^i) be a solution of the BSDE associated with terminal time T, generalized driver *$g^i(t, y, z, k)dt + dD_t$ and terminal condition ξ^i. Let $\bar{\xi} := \xi^1 - \xi^2$. For s in $[0, T]$, denote $\bar{Y}_s := Y_s^1 - Y_s^2$, $\bar{Z}_s := Z_s^1 - Z_s^2$ and $\bar{K}_s := K_s^1 - K_s^2$.*
Let $\eta, \beta > 0$ be such that $\beta \geq \frac{3}{\eta} + 2C$ and $\eta \leq \frac{1}{C^2}$. For each $t \in [0, T]$, we have

$$e^{\beta t}(\bar{Y}_t)^2 \leq \mathbb{E}[e^{\beta T}\bar{\xi}^{\,2} \mid \mathscr{G}_t] + \eta\, \mathbb{E}[\int_t^T e^{\beta s}\bar{g}(s)^2 ds \mid \mathscr{G}_t] \quad a.s., \tag{6}$$

where $\bar{g}(s) := g^1(s, Y_s^2, Z_s^2, K_s^2) - g^2(s, Y_s^2, Z_s^2, K_s^2)$. Moreover,

$$\|\bar{Y}\|_\beta^2 \leq T[e^{\beta T}\mathbb{E}[\bar{\xi}^{\,2}] + \eta\|\bar{g}\|_\beta^2]. \tag{7}$$

If $\eta < \frac{1}{C^2}$, we have

$$\|\bar{Z}\|_\beta^2 + \|\bar{K}\|_{\lambda,\beta}^2 \leq \frac{1}{1 - \eta C^2}[e^{\beta T}\mathbb{E}[\bar{\xi}^{\,2}] + \eta\|\bar{g}\|_\beta^2]. \tag{8}$$

Remark 4 If $C = 0$, then (6) and (7) hold for all $\eta, \beta > 0$ such that $\beta \geq \frac{3}{\eta}$, and (8) holds (with $C = 0$) for all $\eta > 0$.

Proof By Itô's formula applied to the semimartingale $(e^{\beta s}\bar{Y}_s^2)$ between t and T, we get

$$\begin{aligned} e^{\beta t}\bar{Y}_t^2 + \beta\int_t^T e^{\beta s}\bar{Y}_s^2 ds + \int_t^T e^{\beta s}\bar{Z}_s^2 ds + \int_t^T e^{\beta s}\bar{K}_s^2\lambda_s ds \\ = e^{\beta T}\bar{Y}_T^2 + 2\int_t^T e^{\beta s}\bar{Y}_s(g^1(s, Y_s^1, Z_s^1, K_s^1) - g^2(s, Y_s^2, Z_s^2, K_s^2))ds \\ - 2\int_t^T e^{\beta s}\bar{Y}_s\bar{Z}_s dW_s - \int_t^T e^{\beta s}(2\bar{Y}_{s^-}\bar{K}_s + \bar{K}_s^2)dM_s. \end{aligned} \tag{9}$$

Taking the conditional expectation given $\mathscr{G}_t$, we obtain

$$\begin{aligned} e^{\beta t}\bar{Y}_t^2 + \mathbb{E}\left[\beta\int_t^T e^{\beta s}\bar{Y}_s^2 ds + \int_t^T e^{\beta s}(\bar{Z}_s^2 + \bar{K}_s^2\lambda_s)ds \mid \mathscr{G}_t\right] \\ \leq \mathbb{E}\left[e^{\beta T}\bar{Y}_T^2 \mid \mathscr{G}_t\right] + 2\mathbb{E}\left[\int_t^T e^{\beta s}\bar{Y}_s(g^1(s, Y_s^1, Z_s^1, K_s^1) - g^2(s, Y_s^2, Z_s^2, K_s^2))ds \mid \mathscr{G}_t\right]. \end{aligned} \tag{10}$$

Now, $g^1(s, Y_s^1, Z_s^1, K_s^1) - g^2(s, Y_s^2, Z_s^2, K_s^2) = g^1(s, Y_s^1, Z_s^1, K_s^1) - g^1(s, Y_s^2, Z_s^2, K_s^2) + \bar{g}_s$.
Since g^1 satisfies condition (3), we derive that

$$|g^1(s, Y_s^1, Z_s^1, K_s^1) - g^2(s, Y_s^2, Z_s^2, K_s^2)| \leq C|\bar{Y}_s| + C|\bar{Z}_s| + C|\bar{K}_s|\sqrt{\lambda_s} + |\bar{g}_s|.$$

Note that, for all non negative numbers λ, y, z, k, g and $\varepsilon > 0$, we have $2y(Cz + Ck\sqrt{\lambda} + g) \leq \frac{y^2}{\varepsilon^2} + \varepsilon^2(Cz + Ck\sqrt{\lambda} + g)^2 \leq \frac{y^2}{\varepsilon^2} + 3\varepsilon^2(C^2y^2 + C^2k^2\lambda + g^2)$.
Hence,

$$e^{\beta t}\bar{Y}_t^2 + \mathbb{E}\left[\beta\int_t^T e^{\beta s}\bar{Y}_s^2 ds + \int_t^T e^{\beta s}(\bar{Z}_s^2 + \bar{K}_s^2\lambda_s)ds \mid \mathcal{G}_t\right] \leq \mathbb{E}\left[e^{\beta T}\bar{Y}_T^2 \mid \mathcal{G}_t\right] +$$
$$+\mathbb{E}\left[(2C + \frac{1}{\varepsilon^2})\int_t^T e^{\beta s}\bar{Y}_s^2 ds + 3C^2\varepsilon^2\int_t^T e^{\beta s}(\bar{Z}_s^2 + \bar{K}_s^2\lambda_s)ds + 3\varepsilon^2\int_t^T e^{\beta s}\bar{g}_s^2 ds \mid \mathcal{G}_t\right]. \tag{11}$$

Let us make the change of variable $\eta = 3\epsilon^2$. Then, for each $\beta, \eta > 0$ chosen as in the proposition, this inequality leads to (6). By integrating (6), we obtain (7). Using (7) and inequality (11), we derive (8).

Remark 5 By classical results on the norms of semimartingales, one similarly shows that $\|\bar{Y}\|_{\mathscr{S}^2} \leq K\left(\mathbb{E}[\bar{\xi}^2] + \|\bar{g}\|_{\mathbb{H}^2}\right)$, where K is a positive constant only depending on T and C.

2.2.2 Existence and Uniqueness Result for BSDEs with Default Jump

By the representation property of $\mathbb{G}$-martingales (Lemma 1) and the a priori estimates given in Proposition 1, we derive the existence and the uniqueness of the solution associated with a *generalized λ-admissible driver*.

Proposition 2 *Let g be a λ-admissible driver, let $\xi \in L^2(\mathcal{G}_T)$, and let D be an (optional) process belonging to $\mathscr{A}^2$. There exists a unique solution (Y, Z, K) in $\mathscr{S}^2 \times \mathbb{H}^2 \times \mathbb{H}^2_\lambda$ of BSDE (5).*

Remark 6 Suppose that $D = 0$. Suppose also that ξ is $\mathcal{G}_{\vartheta \wedge T}$-measurable and that g is replaced by $g\mathbf{1}_{t \leq \vartheta}$ (which is a λ-admissible driver). Then, the solution (Y, Z, K) of the associated BSDE (4) is equal to the solution of the BSDE with random terminal time $\vartheta \wedge T$, driver g and terminal condition ξ, as considered in [6]. Note also that in the present paper, contrary to papers [6, 12], we do not suppose that the default intensity process λ is bounded (which is interesting since this is the case in some models with default).

Proof Let us first consider the case when $g(t)$ does not depend on (y, z, k). Then the solution Y is given by $Y_t = E[\xi + \int_t^T g(s)ds + D_T - D_t|\mathcal{G}_t]$. The processes Z and K are obtained by applying the representation property of $\mathbb{G}$-martingales to the square integrable martingale $E[\xi + \int_0^T g(s)ds + D_T|\mathcal{G}_t]$. Hence, there thus exists a unique solution of BSDE (5) associated with terminal condition $\xi \in L^2(\mathcal{F}_T)$ and *generalized driver* $g(t)dt + dD_t$. Let us now turn to the case with a general λ-admissible driver $g(t, y, z, k)$. Denote by $I\!H^2_\beta$ the space $\mathcal{S}^2 \times I\!H^2 \times I\!H^2_\lambda$ equipped with the norm $\|Y, Z, K\|^2_\beta := \|Y\|^2_\beta + \|Z\|^2_\beta + \|K\|^2_{\lambda,\beta}$. We define a mapping Φ from $I\!H^2_\beta$ into itself as follows. Given $(U, V, L) \in I\!H^2_\beta$, let $(Y, Z, K) = \Phi(U, V, L)$ be the solution of the BSDE associated with *generalized driver* $g(s, U_s, V_s, L_s)ds + dD_s$ and terminal condition ξ. Let us prove that the mapping Φ is a contraction from $I\!H^2_\beta$ into $I\!H^2_\beta$. Let (U', V', L') be another element of $I\!H^2_\beta$ and let $(Y', Z', K') := \Phi(U', V', L')$, that is, the solution of the BSDE associated with the *generalized driver* $g(s, U'_s, V'_s, L'_s)ds + dD_s$ and terminal condition ξ.
Set $\bar{U} = U - U'$, $\bar{V} = V - V'$, $\bar{L} = L - L'$, $\bar{Y} = Y - Y'$, $\bar{Z} = Z - Z'$, $\bar{K} = K - K'$. Set $\Delta g_t := g(t, U_t, V_t, L_t) - g(t, U'_t, V'_t, L'_t)$. By Remark 4 applied to the driver processes $g_1(t) := g(t, U_t, V_t, L_t)$[4] and $g_2(t) := g(t, U'_t, V'_t, L'_t)$, we derive that for all $\eta, \beta > 0$ such that $\beta \geq \frac{3}{\eta}$, we have

$$\|\bar{Y}\|^2_\beta + \|\bar{Z}\|^2_\beta + \|\bar{K}\|^2_{\lambda,\beta} \leq \eta(T+1)\|\Delta g\|^2_\beta.$$

Since the driver g is λ-*admissible* with λ-constant C, we get

$$\|\bar{Y}\|^2_\beta + \|\bar{Z}\|^2_\beta + \|\bar{K}\|^2_{\lambda,\beta} \leq \eta(T+1)3C^2(\|\bar{U}\|^2_\beta + \|\bar{V}\|^2_\beta + \|\bar{L}\|^2_{\lambda,\beta}),$$

for all $\eta, \beta > 0$ with $\beta \geq \frac{3}{\eta}$. Choosing $\eta = \frac{1}{(T+1)6C^2}$ and $\beta \geq \frac{3}{\eta} = 18(T+1)C^2$, we derive that $\|(\overline{Y}, \overline{Z}, \overline{K})\|^2_\beta \leq \frac{1}{2}\|(\overline{U}, \overline{V}, \overline{K})\|^2_\beta$. Hence, for $\beta \geq 18(T+1)C^2$, Φ is a contraction from $I\!H^2_\beta$ into $I\!H^2_\beta$ and thus admits a unique fixed point (Y, Z, K) in the Banach space $I\!H^2_\beta$, which is the (unique) solution of BSDE (4).

2.3 λ-Linear *BSDEs with Default Jump*

We introduce the notion of λ-*linear* BSDEs in our framework with default jump.

Definition 3 (λ-*linear* driver) A driver g is called λ-*linear* if it is of the form:

$$g(t, y, z, k) = \delta_t y + \beta_t z + \gamma_t k \lambda_t + \varphi_t, \tag{12}$$

[4]Note that the driver processes $g_1(t)$ admits $C = 0$ as λ-constant.

where $(\varphi_t) \in \mathbb{H}^2$, and (δ_t), (β_t) and (γ_t) are $\mathbf{R}$-valued predictable processes such that (δ_t), (β_t) and $(\gamma_t\sqrt{\lambda_t})$ are bounded. By extension,

$$(\delta_t y + \beta_t z + \gamma_t k\,\lambda_t)dt + dD_t,$$

where $D \in \mathscr{A}^2$, is called a *generalized λ-linear driver*.

Remark 7 Note that g given by (12) can be rewritten as

$$g(t, y, z, k) = \varphi_t + \delta_t y + \beta_t z + \nu_t\, k\,\sqrt{\lambda_t}, \tag{13}$$

where $\nu_t := \gamma_t\sqrt{\lambda_t}$ is a bounded predictable process.[5] From this remark, it clearly follows that a λ-*linear* driver is λ-*admissible*.

We will now prove that the solution of a λ-*linear* BSDE (or more generally a *generalized λ-linear driver*) can be written as a conditional expectation via an exponential semimartingale. We first show a preliminary result on exponential semimartingales.

Let (β_s) and (γ_s) be two real-valued $\mathbb{G}$-predictable processes such that the stochastic integrals $\int_0^{\cdot} \beta_s dW_s$ and $\int_0^{\cdot} \gamma_s dM_s$ are well-defined. Let (ζ_s) be the process satisfying the forward SDE:

$$d\zeta_s = \zeta_{s^-}(\beta_s dW_s + \gamma_s dM_s);\quad \zeta_0 = 1. \tag{14}$$

Remark 8 Recall that the process (ζ_s) satisfies the so-called Doléans-Dade formula, that is

$$\zeta_s = \exp\{\int_0^s \beta_r dW_r - \frac{1}{2}\int_0^s \beta_r^2 dr\}\exp\{-\int_0^s \gamma_r\lambda_r dr\}(1+\gamma_\vartheta \mathbf{1}_{\{s\geq\vartheta\}}),\quad s \geq 0 \quad \text{a.s.}$$

Hence, if $\gamma_\vartheta \geq -1$ (resp. > -1) a.s, then $\zeta_s \geq 0$ (resp. > 0) for all $s \geq 0$ a.s.

Remark 9 The inequality $\gamma_\vartheta \geq -1$ a.s. is equivalent to the inequality $\gamma_t \geq -1$, $\lambda_t dt \otimes dP$-a.s. Indeed, we have $\mathbb{E}[\mathbf{1}_{\gamma_\vartheta<-1}] = \mathbb{E}[\int_0^{+\infty} \mathbf{1}_{\gamma_r<-1} dN_r]$ $= \mathbb{E}[\int_0^{+\infty} \mathbf{1}_{\gamma_r<-1}\lambda_r dr]$, because the process $(\int_0^t \lambda_r dr)$ is the $\mathbb{G}$-predictable compensator of the default jump process N.

Proposition 3 *Let $T > 0$. Suppose that the random variable $\int_0^T (\beta_r^2 + \gamma_r^2\lambda_r)\,dr$ is bounded.*
Then, the process $(\zeta_s)_{0\leq s\leq T}$, defined by (14), is a martingale and satisfies $\mathbb{E}[\sup_{0\leq s\leq T}\zeta_s^2] < +\infty$.

[5] Actually the formulation (13) is equivalent to (12).

Proof By definition, the process (ζ_s) is a local martingale. Let $T > 0$. Let us show that $\mathbb{E}[\sup_{0\le s\le T} \zeta_s^2] < +\infty$. By Itô's formula applied to ζ_s^2, we get $d\zeta_s^2 = 2\zeta_{s^-} d\zeta_s + d[\zeta,\zeta]_s$. We have

$$d[\zeta,\zeta]_s = \zeta_{s^-}^2 \beta_s^2 ds + \zeta_{s^-}^2 \gamma_s^2 dN_s.$$

Using (1), we thus derive that

$$d\zeta_s^2 = \zeta_{s^-}^2 [2\beta_s dW_s + (2\gamma_s + \gamma_s^2) dM_s + (\beta_s^2 + \gamma_s^2 \lambda_s) ds].$$

It follows that ζ^2 is an exponential semimartingale which can be written:

$$\zeta_s^2 = \eta_s \exp\{\int_0^s (\beta_r^2 + \gamma_r^2 \lambda_r)\, dr\}, \tag{15}$$

where η is the exponential local martingale satisfying

$$d\eta_s = \eta_{s^-} [2\beta_s dW_s + (2\gamma_s + \gamma_s^2) dM_s],$$

with $\eta_0 = 1$. By equality (15), the local martingale η is non negative. Hence, it is a supermartingale, which yields that $\mathbb{E}[\eta_T] \le 1$. Now, by assumption, $\int_0^T (\beta_r^2 + \gamma_r^2 \lambda_r)\, dr$ is bounded. By (15), it follows that

$$\mathbb{E}[\zeta_T^2] \le \mathbb{E}[\eta_T]\, K \le K,$$

where K is a positive constant. By martingale inequalities, we derive that $\mathbb{E}[\sup_{0\le s\le T} \zeta_s^2] < +\infty$. Hence, the process $(\zeta_s)_{0\le s\le T}$ is a martingale.

Remark 10 Note that, under the assumption from Proposition 3, one can prove by an induction argument (as in the proof of Proposition A.1 in [21]) that for all $p \ge 2$, we have $\mathbb{E}[\sup_{0\le s\le T} \zeta_s^p] < +\infty$.

We now show a representation property of the solution of a *generalized* λ*-linear* BSDE when the finite variational process D is supposed to be predictable.

Theorem 1 (Representation result for *generalized* λ*-linear* BSDEs with D predictable) *Let* (δ_t), (β_t) *and* (γ_t) *be* **R***-valued predictable processes such that* (δ_t), (β_t) *and* $(\gamma_t \sqrt{\lambda_t})$ *are bounded.*
Let $\xi \in L^2(\mathcal{G}_T)$ *and let* D *be a process belonging to* $\mathscr{A}_p^2$, *that is, a finite variational RCLL* predictable *process with* $D_0 = 0$ *and square integrable total variation process.*
Let (Y, Z, K) *be the solution in* $\mathscr{S}^2 \times I\!H^2 \times I\!H_\lambda^2$ *of the BSDE associated with* generalized λ-linear driver $(\delta_t y + \beta_t z + \gamma_t k\, \lambda_t)dt + dD_t$ *and terminal condition* ξ, *that is*

$$-dY_t = (\delta_t Y_t + \beta_t Z_t + \gamma_t K_t \lambda_t)dt + dD_t - Z_t dW_t - K_t dM_t; \quad Y_T = \xi. \tag{16}$$

For each $t \in [0, T]$, let $(\Gamma_{t,s})_{s \geq t}$ (called the adjoint process*) be the unique solution of the following forward SDE*

$$d\Gamma_{t,s} = \Gamma_{t,s^-}\left[\delta_s ds + \beta_s dW_s + \gamma_s dM_s\right]; \quad \Gamma_{t,t} = 1. \tag{17}$$

The process (Y_t) satisfies

$$Y_t \; = \; \mathbb{E}[\,\Gamma_{t,T}\;\xi + \int_t^T \Gamma_{t,s^-}\, dD_s \mid \mathscr{G}_t\,], \quad 0 \leq t \leq T, \quad \text{a.s.} \tag{18}$$

Remark 11 From Remark 8, it follows that the process $(\Gamma_{t,s})_{s \geq t}$, defined by (17), satisfies

$$\Gamma_{t,s} = e^{\int_t^s \delta_r dr} \exp\{\int_t^s \beta_r dW_r - \frac{1}{2}\int_t^s \beta_r^2 dr\} e^{-\int_t^s \gamma_r \lambda_r dr}(1 + \gamma_\vartheta \mathbf{1}_{\{s \geq \vartheta > t\}}) \quad s \geq t \quad \text{a.s.}$$

Hence, if $\gamma_\vartheta \geq -1$ (resp. > -1) a.s., we then have $\Gamma_{t,s} \geq 0$ (resp. > 0) for all $s \geq t$ a.s.

Note also that the process $(e^{\int_t^s \delta_r dr})_{t \leq s \leq T}$ is positive, and bounded since δ is bounded. Using Proposition 3, since β and $\gamma\sqrt{\lambda}$ are bounded, we derive that $\mathbb{E}[\sup_{t \leq s \leq T} \Gamma_{t,s}^2] < +\infty$.

Proof Fix $t \in [0, T]$. Note first that since D is a finite variational RCLL process, here supposed to be *predictable*, and since the process $\Gamma_{t,\cdot}$ admits only one jump at the totally inaccessible stopping time ϑ, we get $[\Gamma_{t,\cdot}, D] = 0$. By applying the Itô product formula to $Y_s\Gamma_{t,s}$, we get

$$\begin{aligned} -d(Y_s\Gamma_{t,s}) &= -Y_{s^-}d\Gamma_{t,s} - \Gamma_{t,s^-}dY_s - d[Y,\Gamma]_s \\ &= -Y_s\Gamma_{t,s^-}\delta_s ds + \Gamma_{t,s^-}\left[\delta_s Y_s + \beta_s Z_s + \gamma_s K_s \lambda_s\right] ds + \Gamma_{t,s^-}dD_s \\ &\quad - \beta_s Z_s \Gamma_{t,s^-} ds - \Gamma_{t,s^-}\gamma_s K_s \lambda_s ds - \Gamma_{t,s^-}(Y_s\beta_s + Z_s) dW_s \\ &\quad - \Gamma_{t,s^-}[K_s(1+\gamma_s) + Y_{s^-}\gamma_s] dM_s. \end{aligned} \tag{19}$$

Setting

$$dm_s = -\Gamma_{t,s^-}(Y_s\beta_s + Z_s)dW_s - \Gamma_{t,s^-}[K_s(1+\gamma_s) + Y_{s^-}\gamma_s]dM_s,$$

we get

$$-\,d(Y_s\Gamma_{t,s}) = \Gamma_{t,s^-}dD_s - dm_s. \tag{20}$$

By integrating between t and T, we obtain

$$Y_t = \xi\, \Gamma_{t,T} + \int_t^T \Gamma_{t,s^-} dD_s - (m_T - m_t) \quad \text{a.s.} \tag{21}$$

By Remark 11, we have $(\Gamma_{t,s})_{t\le s\le T} \in \mathscr{S}^2$. Moreover, $Y \in \mathscr{S}^2$, $Z \in I\!H^2$, $K \in I\!H^2_\lambda$, and β and $\gamma\sqrt{\lambda}$ are bounded. It follows that the local martingale $m = (m_s)_{t\le s\le T}$ is a martingale. Hence, by taking the conditional expectation in equality (21), we get equality (18).

When the finite variational process D is no longer supposed to be predictable (which is often the case in the literature on default risk[6]), the representation formula (18) does not generally hold. We now provide a representation property of the solution in that case, that is, when the finite variational process D is only supposed to be RCLL and adapted, which is new in the literature on BSDEs.

Theorem 2 (Representation result for *generalized* λ*-linear* BSDEs with D optional) *Suppose that the assumptions of Theorem 1 hold, except that D is supposed to belong to $\mathscr{A}^2$ instead of $\mathscr{A}^2_p$. Let $D' \in \mathscr{A}^2_p$ and $\eta \in I\!H^2_\lambda$ be such that for all $t \in [0, T]$,*[7]

$$D_t = D'_t + \int_0^t \eta_s dN_s \quad \text{a.s.} \tag{22}$$

Let (Y, Z, K) be the solution in $\mathscr{S}^2 \times I\!H^2 \times I\!H^2_\lambda$ of the BSDE associated with generalized λ-linear driver $(\delta_t y + \beta_t z + \gamma_t\, k\, \lambda_t)dt + dD_t$ *and terminal condition ξ, that is BSDE* (16).

Then, a.s. for all $t \in [0, T]$,

$$\begin{aligned} Y_t &= \mathbb{E}[\, \Gamma_{t,T}\ \xi + \int_t^T \Gamma_{t,s^-}\, (dD'_s + (1+\gamma_s)\eta_s dN_s) \mid \mathscr{G}_t] \\ &= \mathbb{E}[\, \Gamma_{t,T}\ \xi + \int_t^T \Gamma_{t,s^-}\, dD'_s + \Gamma_{t,\vartheta}\, \eta_\vartheta \mathbf{1}_{\{t<\vartheta\le T\}} \mid \mathscr{G}_t] \end{aligned} \tag{23}$$

where $(\Gamma_{t,s})_{s\in[t,T]}$ satisfies (17).

Proof Since D satisfies (22), we get $d[\Gamma_{t,\cdot}, D]_s = \Gamma_{t,s^-}\gamma_s\eta_s dN_s$. The computations are then similar to those of the proof of Theorem 1, with $\Gamma_{t,s^-}\, dD_s$ replaced by $\Gamma_{t,s^-}\, (dD_s + \gamma_s\eta_s dN_s)$ in Eqs. (19), (20) and (21). We thus derive that $Y_t =$

[6] In the case of a contingent claim or a contract subjected to default, ΔD_ϑ represents the cashflow generated by the claim at the default time ϑ (see Sect. 3). It is sometimes called "rebate" (cf. [3, 16]).

[7] See Remark 2.

$\mathbb{E}[\,\Gamma_{t,T}\,\xi + \int_t^T \Gamma_{t,s^-}(dD_s + \gamma_s\eta_s dN_s) \mid \mathcal{G}_t\,]$ a.s. From this together with (22), the first equality of (23) follows. Now, we have a.s.

$$\begin{aligned}\mathbb{E}[\int_t^T \Gamma_{t,s^-}(1+\gamma_s)\eta_s dN_s \mid \mathcal{G}_t] &= \mathbb{E}[\Gamma_{t,\vartheta^-}(1+\gamma_\vartheta)\eta_\vartheta\,\mathbf{1}_{\{t<\vartheta\le T\}} \mid \mathcal{G}_t] \\ &= \mathbb{E}[\,\Gamma_{t,\vartheta}\,\eta_\vartheta\,\mathbf{1}_{\{t<\vartheta\le T\}} \mid \mathcal{G}_t],\end{aligned}$$

where the second equality is due to the fact that $\Gamma_{t,\vartheta^-}(1+\gamma_\vartheta) = \Gamma_{t,\vartheta}$ a.s. (cf. Remark 11). This yields the second equality of (23).

Remark 12 By adapting the arguments of the above proof, this result can be generalized to the case of a BSDE driven by a Brownian motion and a Poisson random measure,[8] which provides a new result in the theory of BSDEs in this framework.

2.4 Comparison Theorems for BSDEs with Default Jump

We now provide a comparison theorem and a strict comparison theorem for BSDEs with *generalized λ-admissible drivers* associated with finite variational RCLL adapted processes.

Theorem 3 (Comparison theorems) *Let ξ_1 and $\xi_2 \in L^2(\mathcal{G}_T)$. Let g_1 and g_2 be two* λ-admissible *drivers. Let D^1 and D^2 be two (optional) processes in $\mathcal{A}^2$.*
For $i = 1, 2$, let (Y^i, Z^i, K^i) be the solution in $\mathcal{S}^2 \times I\!H^2 \times I\!H^2_\lambda$ of the following BSDE

$$-dY_t^i = g_i(t, Y_t^i, Z_t^i, K_t^i)dt + dD_t^i - Z_t^i dW_t - K_t^i dM_t; \quad Y_T^i = \xi_i.$$

(i) (Comparison theorem). *Assume that there exists a predictable process (γ_t) with*

$$(\gamma_t\sqrt{\lambda_t}) \text{ bounded and } \gamma_t \ge -1, \quad dP\otimes dt - a.e. \tag{24}$$

such that

$$g_1(t, Y_t^2, Z_t^2, K_t^1) - g_1(t, Y_t^2, Z_t^2, K_t^2) \ge \gamma_t(K_t^1 - K_t^2)\lambda_t, \quad t\in[0,T], \quad dP\otimes dt - a.e. \tag{25}$$

[8] Since in this case, the jumps times of the Poisson random measure are totally inaccessible.

Suppose that $\xi_1 \geq \xi_2$ a.s., that the process $\bar{D} := D^1 - D^2$ is non decreasing, and that

$$g_1(t, Y_t^2, Z_t^2, K_t^2) \geq g_2(t, Y_t^2, Z_t^2, K_t^2), \quad t \in [0, T], \quad dP \otimes dt - a.e. \tag{26}$$

We then have $Y_t^1 \geq Y_t^2$ for all $t \in [0, T]$ a.s.

(ii) (Strict Comparison Theorem). *Suppose moreover that $\gamma_\vartheta > -1$ a.s.*

If $Y_{t_0}^1 = Y_{t_0}^2$ a.s. for some $t_0 \in [0, T]$, then $\xi^1 = \xi^2$ a.s., and the inequality (26) is an equality on $[t_0, T]$. Moreover, $\bar{D} = D^1 - D^2$ is constant on $[t_0, T]$ and $Y^1 = Y^2$ on $[t_0, T]$.

Remark 13 We stress that the above comparison theorems hold even in the case when the generalized drivers are associated with non-predictable finite variational processes, which thus may admit a jump at the default time ϑ. This is important for the applications to nonlinear pricing of contingents claims. Indeed, in a market with default, contingent claims often generate a cashflow at the default time (see Sect. 3.3 for details).

As seen in the proof below, the treatment of the case of non-predictable finite variational processes requires some additional arguments, compared to the case of predictable ones.

Proof Setting $\bar{Y}_s = Y_s^1 - Y_s^2$; $\bar{Z}_s = Z_s^1 - Z_s^2$; $\bar{K}_s = K_s^1 - K_s^2$, we have

$$-d\bar{Y}_s = h_s ds + d\bar{D}_s - \bar{Z}_s dW_s - \bar{K}_s dM_s; \quad \bar{Y}_T = \xi_1 - \xi_2,$$

where $h_s := g_1(s, Y_{s^-}^1, Z_s^1, K_s^1) - g_2(s, Y_{s^-}^2, Z_s^2, K_s^2)$.

Set $\delta_s := \dfrac{g_1(s, Y_{s^-}^1, Z_s^1, K_s^1) - g_1(s, Y_{s^-}^2, Z_s^1, K_s^1)}{\bar{Y}_{s^-}}$ if $\bar{Y}_{s^-} \neq 0$, and 0 otherwise.

Set $\beta_s := \dfrac{g_1(s, Y_{s^-}^2, Z_s^1, K_s^1) - g_1(s, Y_{s^-}^2, Z_s^2, K_s^1)}{\bar{Z}_s}$ if $\bar{Z}_s \neq 0$, and 0 otherwise.

By definition, the processes δ and β are *predictable*. Moreover, since g_1 satisfies condition (3), the processes δ and β are bounded. Now, we have

$$h_s = \delta_s \bar{Y}_{s^-} + \beta_s \bar{Z}_s + g_1(s, Y_{s^-}^2, Z_s^2, K_s^1) - g_1(s, Y_{s^-}^2, Z_s^2, K_s^2) + \varphi_s,$$

where $\varphi_s := g_1(s, Y_{s^-}^2, Z_s^2, K_s^2) - g_2(s, Y_{s^-}^2, Z_s^2, K_s^2)$.[9]

Using the assumption (25) and the equality $\bar{Y}_{s^-} = \bar{Y}_s$ $dP \otimes ds$-a.e. (cf. Remark 3), we get

$$h_s \geq \delta_s \bar{Y}_s + \beta_s \bar{Z}_s + \gamma_s \ \bar{K}_s \lambda_s + \varphi_s \quad dP \otimes ds - \text{a.e.} \tag{27}$$

[9]Note that, by Remark 3, we have $\varphi_s = g_1(s, Y_s^2, Z_s^2, K_s^2) - g_2(s, Y_s^2, Z_s^2, K_s^2)$ $dP \otimes ds$-a.e.

Fix $t \in [0,T]$. Let $\Gamma_{t,.}$ be the process defined by (17). Since δ, β and $\gamma\sqrt{\lambda}$ are bounded, it follows from Remark 11 that $\Gamma_{t,.} \in \mathscr{S}^2$. Also, since $\gamma_s \geq -1$, we have $\Gamma_{t,.} \geq 0$ a.s. Let us first consider the simpler case when the processes D^1 and D^2 are predictable. By Itô's formula and similar computations to those of the proof of Theorem 1, we derive that

$$-d(\bar{Y}_s\Gamma_{t,s}) = \Gamma_{t,s}(h_s - \delta_s\bar{Y}_s - \beta_s\bar{Z}_s - \gamma_s\,\bar{K}_s\,\lambda_s)\,ds + \Gamma_{t,s^-}d\bar{D}_s - dm_s,$$

where m is a martingale (because $\Gamma_{t,.} \in \mathscr{S}^2$, $\bar{Y} \in \mathscr{S}^2$, $\bar{Z} \in I\!H^2$, $\bar{K} \in I\!H^2_\lambda$ and β, $\gamma\sqrt{\lambda}$ are bounded). Using inequality (27) together with the non negativity of Γ, we thus get $-d(\bar{Y}_s\Gamma_{t,s}) \geq \Gamma_{t,s}\varphi_s ds + \Gamma_{t,s^-}d\bar{D}_s - dm_s$. By integrating between t and T and by taking the conditional expectation, we obtain

$$\bar{Y}_t \;\geq\; \mathbb{E}\left[\,\Gamma_{t,T}\,(\xi_1 - \xi_2) + \int_t^T \Gamma_{t,s^-}\,(\varphi_s ds + d\bar{D}_s)\;|\;\mathscr{G}_t\right],\quad 0 \leq t \leq T,\quad \text{a.s.} \tag{28}$$

By assumption (26), $\varphi_s \geq 0\; dP \otimes ds$-a.e. Moreover, $\xi_1 - \xi_2 \geq 0$ and $\bar{D}$ is non decreasing, which, together with the non negativity of $\Gamma_{t,.}$, implies that $\bar{Y}_t = Y^1_t - Y^2_t \geq 0$ a.s. Since this inequality holds for all $t \in [0,T]$, the assertion (i) follows. Suppose moreover that $Y^1_{t_0} = Y^2_{t_0}$ a.s. and that $\gamma_\cdot > -1$. Since $\gamma_\vartheta > -1$ a.s. , we have $\Gamma_{t,s} > 0$ a.s. for all $s \geq t$. From this, together with (28) applied with $t = t_0$, we get $\xi_1 = \xi_2$ a.s. and $\varphi_t = 0$, $t \in [t_0,T]\; dP \otimes dt$-a.e. On the other hand, set $\tilde{D}_t := \int_{t_0}^t \Gamma_{t_0,s^-}d\bar{D}_s$, for each $t \in [t_0,T]$. By assumption, $\tilde{D}_T \geq 0$ a.s. By (28), we thus get $\mathbb{E}[\tilde{D}_T \mid \mathscr{G}_{t_0}] = 0$ a.s. Hence $\tilde{D}_T = 0$ a.s. Now, since $\Gamma_{t_0,s} > 0$, for all $s \geq t_0$ a.s. , we can write $\bar{D}_T - \bar{D}_{t_0} = \int_{t_0}^T \Gamma^{-1}_{t_0,s^-}d\tilde{D}_s$. We thus get $\bar{D}_T = \bar{D}_{t_0}$ a.s. The proof of (ii) is thus complete.

Let us now consider the case when the processes D^1 and D^2 are not predictable. By Remark 2, for $i = 1,2$, there exist $D'^i \in \mathscr{A}^2_p$ and $\eta^i \in I\!H^2_\lambda$ such that

$$D^i_t = D'^i_t + \int_0^t \eta^i_s dN_s \quad \text{a.s.}$$

Since $\bar{D} := D^1 - D^2$ is non decreasing, we derive that the process $\bar{D}' := D'^1 - D'^2$ is non decreasing and that $\eta^1_\vartheta \geq \eta^2_\vartheta$ a.s. on $\{\vartheta \leq T\}$. By Itô's formula and similar computations to those of the proof of Theorems 1 and 2, we get

$$\begin{aligned}-d(\bar{Y}_s\Gamma_{t,s}) = {}& \Gamma_{t,s}(h_s - \delta_s\bar{Y}_s - \beta_s\bar{Z}_s - \gamma_s\,\bar{K}_s\,\lambda_s)\,ds \\ & + \Gamma_{t,s^-}[d\bar{D}_s + (\eta^1_s - \eta^2_s)\gamma_s dN_s] - dm_s,\end{aligned}$$

where m is a martingale. Using inequality (27) and the equality $\bar{D}_t = \bar{D}'_t + \int_0^t (\eta_s^1 - \eta_s^2) dN_s$ a.s., we thus derive that

$$\bar{Y}_t \geq \mathbb{E}\,[\,\Gamma_{t,T}\,(\xi_1 - \xi_2) + \int_t^T \Gamma_{t,s^-}\,(\varphi_s ds + d\bar{D}'_s + (\eta_s^1 - \eta_s^2)(1+\gamma_s) dN_s) \mid \mathscr{G}_t], \quad 0 \leq t \leq T, \quad \text{a.s.} \tag{29}$$

Since $\eta_\vartheta^1 \geq \eta_\vartheta^2$ a.s. on $\{\vartheta \leq T\}$ and $\gamma_\vartheta \geq -1$ a.s., we have $(\eta_\vartheta^1 - \eta_\vartheta^2)(1+\gamma_\vartheta) \geq 0$ a.s. on $\{\vartheta \leq T\}$. Hence, using the other assumptions made in (i), we derive that $\bar{Y}_t = Y_t^1 - Y_t^2 \geq 0$ a.s. Since this inequality holds for all $t \in [0, T]$, the assertion (i) follows.
Suppose moreover that $Y_{t_0}^1 = Y_{t_0}^2$ a.s. and that $\gamma_\vartheta > -1$ a.s. By the inequality (29) applied with $t = t_0$, we derive that $\xi^1 = \xi^2$ a.s., $\varphi_s = 0\, dP \otimes ds$-a.e. on $[t_0, T]$, $\eta_\vartheta^1 = \eta_\vartheta^2$ a.s. on $\{t_0 < \vartheta \leq T\}$. Moreover, $\bar{D}' = D'^1 - D'^2$ is constant on the time interval $[t_0, T]$. Hence, $\bar{D}$ is constant on $[t_0, T]$. The proof is thus complete.

Remark 14 By adapting the arguments of the above proof, this result can be generalized to the case of BSDEs driven by a Brownian motion and a Poisson random measure (since the jumps times associated with the Poisson random measure are totally inaccessible). This extends the comparison theorems given in the literature on BSDEs with jumps (see [21, Theorems 4.2 and 4.4]) to the case of generalized drivers of the form $g(t, y, z, k)dt + dD_t$, where D is a finite variational RCLL adapted process (not necessarily predictable).

When the assumptions of the comparison theorem (resp. strict comparison theorem) are violated, the conclusion does not necessarily hold, as shown by the following example.

Example 1 Suppose that the process λ is bounded. Let g be a λ-linear driver (see (12)) of the form

$$g(\omega, t, y, z, k) = \delta_t(\omega) y + \beta_t(\omega)\, z + \gamma\, k\, \lambda_t(\omega), \tag{30}$$

where γ is here a real constant. At terminal time T, the associated adjoint process $\Gamma_{0,\cdot}$ satisfies (see (17) and Remark 11):

$$\Gamma_{0,T} = H_T \exp\{-\int_0^T \gamma \lambda_r dr\}(1 + \gamma\, \mathbf{1}_{\{T \geq \vartheta\}}). \tag{31}$$

where (H_t) satisfies $dH_t = H_t(\delta_t dt + \beta_t dW_t)$ with $H_0 = 1$.

Let Y be the solution of the BSDE associated with driver g and terminal condition

$$\xi := \mathbf{1}_{\{T \geq \vartheta\}}.$$

The representation property of λ-linear BSDEs with default jump (see (18)) gives

$$Y_0 = \mathbb{E}[\Gamma_{0,T}\xi] = \mathbb{E}[\Gamma_{0,T}\mathbf{1}_{\{T\geq\vartheta\}}].$$

Hence, by (31), we get

$$\begin{aligned} Y_0 = \mathbb{E}[\Gamma_{0,T}\mathbf{1}_{\{T\geq\vartheta\}}] &= \mathbb{E}[H_T e^{-\gamma\int_0^T \lambda_s ds}(1+\gamma\,\mathbf{1}_{\{T\geq\vartheta\}})\mathbf{1}_{\{T\geq\vartheta\}}] \\ &= (1+\gamma)\mathbb{E}[H_T e^{-\gamma\int_0^T \lambda_s ds}\mathbf{1}_{\{T\geq\vartheta\}}]. \end{aligned} \tag{32}$$

Equation (32) shows that under the additional assumption $P(T \geq \vartheta) > 0$, when $\gamma < -1$, we have $Y_0 < 0$ although $\xi \geq 0$ a.s.

This example also gives a counter-example for the strict comparison theorem by taking $\gamma = -1$. Indeed, in this case, the relation (32) yields that $Y_0 = 0$. Under the additional assumption $P(T \geq \vartheta) > 0$, we have $P(\xi > 0) > 0$, even though $Y_0 = 0$.

3 Nonlinear Pricing in a Financial Market with Default

3.1 Financial Market with Defaultable Risky Asset

We consider a complete financial market with default as in [4], which consists of one risk-free asset, with price process S^0 satisfying $dS_t^0 = S_t^0 r_t dt$ with $S_0^0 = 1$, and two risky assets with price processes S^1, S^2 evolving according to the equations:

$$\begin{aligned} dS_t^1 &= S_t^1[\mu_t^1 dt + \sigma_t^1 dW_t] \quad \text{with} \quad S_0^1 > 0; \\ dS_t^2 &= S_{t-}^2[\mu_t^2 dt + \sigma_t^2 dW_t - dM_t] \quad \text{with} \quad S_0^2 > 0, \end{aligned} \tag{33}$$

where the process (M_t) is given by (1).

The processes $\sigma^1, \sigma^2, r, \mu^1, \mu^2$ are predictable (that is $\mathscr{P}$-measurable). We set $\sigma = (\sigma^1, \sigma^2)'$, where $'$ denotes transposition.

We suppose that $\sigma^1, \sigma^2 > 0$, and $r, \mu^1, \mu^2, \sigma^1, \sigma^2, (\sigma^1)^{-1}, (\sigma^2)^{-1}$ are bounded. Note that the intensity process (λ_t) is not necessarily bounded, which is useful in market models with default where the intensity process is modeled by the solution (which is not necessarily bounded) of a forward stochastic differential equation.

Remark 15 By Remark 11, we have

$$S_t^2 = e^{\int_0^t \mu_r^2 dr}\exp\{\int_0^t \sigma_r^2 dW_r - \frac{1}{2}\int_0^t (\sigma_r^2)^2 dr\}\, e^{\int_0^t \lambda_r dr}(1-\mathbf{1}_{\{t\geq\vartheta\}}), \quad t \geq 0 \quad \text{a.s.}$$

The second risky asset is thus defaultable with total default: we have $S_t^2 = 0$, $t \geq \vartheta$ a.s.

We consider an investor who, at time 0, invests an initial amount $x \in \mathbf{R}$ in the three assets. For $i = 1, 2$, we denote by φ_t^i the amount invested in the i^{th} risky asset. After time ϑ, the investor does not invest in the defaultable asset since its price is equal to 0. We thus have $\varphi_t^2 = 0$ on $t > \vartheta$. A process $\varphi = (\varphi^1, \varphi^2)'$ belonging to $\mathbb{H}^2 \times \mathbb{H}^2_\lambda$ is called a *risky assets strategy*. Let C be a finite variational optional process belonging to $\mathscr{A}^2$, representing the *cumulative cash amount withdrawn* from the portfolio.

The value at time t of the portfolio (or *wealth*) associated with x, φ and C is denoted by $V_t^{x,\varphi,C}$. The amount invested in the risk-free asset at time t is then given by $V_t^{x,\varphi,C} - (\varphi_t^1 + \varphi_t^2)$.

3.2 Pricing of European Options with Dividends in a Perfect (Linear) Market with Default

In this section, we place ourselves in a perfect (linear) market model with default. In this case, by the self financing condition, the wealth process $V^{x,\varphi,C}$ (simply denoted by V) follows the dynamics:

$$\begin{aligned} dV_t &= (r_t V_t + \varphi_t^1(\mu_t^1 - r_t) + \varphi_t^2(\mu_t^2 - r_t))dt - dC_t + (\varphi_t^1\sigma_t^1 + \varphi_t^2\sigma_t^2)dW_t - \varphi_t^2 dM_t \\ &= \left(r_t V_t + \varphi_t{}'\sigma_t\theta_t^1 - \varphi_t^2\theta_t^2\lambda_t\right)dt - dC_t + \varphi_t'\sigma_t dW_t - \varphi_t^2 dM_t, \end{aligned} \tag{34}$$

where $\varphi_t{}'\sigma_t = \varphi_t^1\sigma_t^1 + \varphi_t^2\sigma_t^2$, and

$$\theta_t^1 := \frac{\mu_t^1 - r_t}{\sigma_t^1}; \quad \theta_t^2 := -\frac{\mu_t^2 - \sigma_t^2\theta_t^1 - r_t}{\lambda_t}\mathbf{1}_{\{\lambda_t \neq 0\}}.$$

Suppose that the processes θ^1 and $\theta^2\sqrt{\lambda}$ are bounded.

Let $T > 0$. Let ξ be a $\mathscr{G}_T$-measurable random variable belonging to L^2, and let D be a finite variational optional process belonging to $\mathscr{A}_T^2$. We consider a European option with maturity T, which generates a *terminal payoff* ξ, and intermediate cashflows called *dividends*, which are not necessarily positive (cf. for example [7, 8]). For each $t \in [0, T]$, D_t represents the cumulative intermediate cashflows paid by the option between time 0 and time t. The process $D = (D_t)$ is called the *cumulative dividend process*. Note that D is not necessarily non decreasing.

The aim is to price this contingent claim. Let us consider an agent who wants to sell the option at time 0. With the amount the seller receives at time 0 from the buyer, he/she wants to be able to construct a portfolio which allows him/her to pay to the buyer the amount ξ at time T, as well as the intermediate dividends.

Now, setting

$$Z_t := \varphi_t{}'\sigma_t \;\; ; \;\; K_t := -\varphi_t^2, \tag{35}$$

by (34), we derive that the process (V, Z, K) satisfies the following dynamics:

$$-dV_t = -(r_t V_t + \theta_t^1 Z_t + \theta_t^2 K_t \lambda_t)dt + dD_t - Z_t dW_t - K_t dM_t\,.$$

We set for each (ω, t, y, z, k),

$$g(\omega, t, y, z, k) := -r_t(\omega)y - \theta_t^1(\omega)z - \theta_t^2(\omega)\, k\ \lambda_t(\omega). \tag{36}$$

Since by assumption, the coefficients $r, \theta^1, \theta^2\sqrt{\lambda}$ are predictable and bounded, it follows that g is a λ-*linear* driver (see Definition 3). By Proposition 2, there exists a unique solution $(X, Z, K) \in \mathscr{S}^2 \times \mathbb{H}^2 \times \mathbb{H}^2_\lambda$ of the BSDE associated with terminal time T, *generalized* λ-*linear driver* $g(t, y, z, k)dt + dD_t$ (with g defined by (36)) and terminal condition ξ.

Let us show that the process (X, Z, K) provides a replicating portfolio. Let φ be the risky-assets strategy such that (35) holds. Note that this defines a change of variables Φ as follows:
$\Phi : \mathbb{H}^2 \times \mathbb{H}^2_\lambda \to \mathbb{H}^2 \times \mathbb{H}^2_\lambda$; $(Z, K) \mapsto \Phi(Z, K) := \varphi$, where $\varphi = (\varphi^1, \varphi^2)$ is given by (35), which is equivalent to

$$\varphi_t^2 = -K_t \ \ ; \ \ \varphi_t^1 = \frac{Z_t - \varphi_t^2 \sigma_t^2}{\sigma_t^1} = \frac{Z_t + \sigma_t^2 K_t}{\sigma_t^1}. \tag{37}$$

The process D corresponds here to the cumulative cash withdrawn by the seller from his/her hedging portfolio. The process X thus coincides with $V^{X_0,\varphi,D}$, the value of the portfolio associated with initial wealth $x = X_0$, risky-assets strategy φ and cumulative cash withdrawal D. We deduce that this portfolio is a replicating portfolio for the seller since, by investing the initial amount X_0 in the reference assets along the strategy φ, the seller can pay the terminal payoff ξ to the buyer at time T, as well as the intermediate dividends (since the cash withdrawals perfectly replicate the dividends of the option). We derive that X_0 is the initial price of the option, called *hedging price*, denoted by $X_{0,T}(\xi, D)$, and that φ is the *hedging risky-assets strategy*. Similarly, for each time $t \in [0, T]$, X_t is the *hedging price* at time t of the option, and is denoted by $X_{t,T}(\xi, D)$.

Suppose that the *cumulative dividend process* D is predictable. Since the driver g given by (36) is λ-*linear*, the representation property of the solution of a *generalized* λ-*linear* BSDE (see Theorem 1) yields

$$X_{t,T}(\xi, D) = \mathbb{E}[e^{-\int_t^T r_s ds}\zeta_{t,T}\xi + \int_t^T e^{-\int_t^s r_u du}\zeta_{t,s^-}dD_s \,|\, \mathscr{G}_t] \quad \text{a.s.}\,, \tag{38}$$

where ζ satisfies

$$d\zeta_{t,s} = \zeta_{t,s^-}[-\theta_s^1 dW_s - \theta_s^2 dM_s]; \quad \zeta_{t,t} = 1.$$

Suppose now that $\theta_t^2 < 1$ $dP \otimes dt$-a.e. By Proposition 3 and Remark 8, the process $\zeta_{0,\cdot}$ is a square integrable positive martingale. Let Q be the probability measure which admits $\zeta_{0,T}$ as density with respect to P on $\mathscr{G}_T$.[10] By the equality (38), we have

$$X_{t,T}(\xi, D) = \mathbb{E}_Q[e^{-\int_t^T r_s ds}\xi + \int_t^T e^{-\int_t^s r_u du} dD_s \,|\, \mathscr{G}_t] \quad \text{a.s.} \tag{39}$$

When the *cumulative dividend process* D is not predictable and thus admits a jump at time ϑ, the representation formulas (38) and (39) for the no-arbitrage price of the contingent claim do not generally hold. In this case, by Remark 2, there exist a (unique) process $D' \in \mathscr{A}_p^2$ and a (unique) process $\eta \in I\!H_\lambda^2$ such that for all $t \in [0, T]$,

$$D_t = D'_t + \int_0^t \eta_s dN_s \quad \text{a.s.} \tag{40}$$

The random variable η_ϑ (sometimes called "rebate" in the literature) represents the cash flow generated by the contingent claim at the default time ϑ (see e.g. [3, 7, 8, 16] for examples of such contingent claims). By Theorem 2, we get

$$X_{t,T}(\xi, D) = \mathbb{E}[e^{-\int_t^T r_s ds}\zeta_{t,T}\xi + \int_t^T e^{-\int_t^s r_u du}\zeta_{t,s^-} dD'_s + e^{-\int_t^\vartheta r_s ds}\zeta_{t,\vartheta}\eta_\vartheta \mathbf{1}_{\{t<\vartheta\le T\}} \,|\, \mathscr{G}_t] \quad \text{a.s.},$$

or equivalently

$$X_{t,T}(\xi, D) = \mathbb{E}_Q[e^{-\int_t^T r_s ds}\xi + \int_t^T e^{-\int_t^s r_u du} dD'_s + e^{-\int_t^\vartheta r_s ds}\eta_\vartheta \mathbf{1}_{\{t<\vartheta\le T\}} \,|\, \mathscr{G}_t] \quad \text{a.s.}$$

We thus recover the risk-neutral pricing formula of [3, 16], which we have established here by working under the primitive probability measure, using BSDE techniques.

We note that the pricing system (for a fixed maturity T): $(\xi, D) \mapsto X_{\cdot,T}(\xi, D)$ is linear.

[10] Note that the discounted price process $(e^{-\int_0^t r_s ds} S_t^1)_{0\le t\le T}$ (resp. $(e^{-\int_0^t r_s ds} S_t^2)_{0\le t\le T}$) is a martingale (resp. local martingale) under Q. Suppose now that $E[e^{q\int_0^T \lambda_r dr}] < +\infty$ for some $q > 2$. Using Remark 15, we show that $e^{-\int_0^T r_s ds} S_T^2 \in L_Q^2$, which, by martingale inequalities, implies that $(e^{-\int_0^t r_s ds} S_t^2)_{0\le t\le T}$ is a martingale under Q. In other terms, Q is a *martingale probability measure*. By classical arguments, Q can be shown to be the unique *martingale probability measure*.

3.3 Nonlinear Pricing of European Options with Dividends in an Imperfect Market with Default

From now on, we assume that there are imperfections in the market which are taken into account via the *nonlinearity* of the dynamics of the wealth. More precisely, we suppose that the *wealth* process $V_t^{x,\varphi,C}$ (or simply V_t) associated with an initial wealth x, a risky-assets strategy $\varphi = (\varphi^1, \varphi^2)$ in $\mathbb{H}^2 \times \mathbb{H}^2_\lambda$ and a cumulative withdrawal process $C \in \mathscr{A}^2$ satisfies the following dynamics:

$$-dV_t = g(t, V_t, {\varphi_t}'\sigma_t, -\varphi_t^2)dt - {\varphi_t}'\sigma_t dW_t + dC_t + \varphi_t^2 dM_t;\ V_0 = x, \tag{41}$$

where g is a nonlinear λ-*admissible* driver (see Definition 1). Equivalently, setting $Z_t = {\varphi_t}'\sigma_t$ and $K_t = -\varphi_t^2$, we have

$$-dV_t = g(t, V_t, Z_t, K_t)dt - Z_t dW_t + dC_t - K_t dM_t;\ V_0 = x. \tag{42}$$

Let us consider a European option with maturity T, terminal payoff $\xi \in L^2(\mathscr{G}_T)$, and dividend process $D \in \mathscr{A}^2$ (with a possible jump at the default time ϑ) in this market model. Let $(X^g_{\cdot,T}(\xi, D), Z^g_{\cdot,T}(\xi, D), K^g_{\cdot,T}(\xi, D))$, simply denoted by (X, Z, K), be the solution of BSDE associated with terminal time T, *generalized driver* $g(t, y, z, k)dt + dD_t$ and terminal condition ξ, that is satisfying

$$-dX_t = g(t, X_t, Z_t, K_t)dt + dD_t - Z_t dW_t - K_t dM_t; \quad X_T = \xi.$$

The process $X = X^g_{\cdot,T}(\xi, D)$ is equal to the wealth process associated with initial value $x = X_0$, strategy $\varphi = \Phi(Z, K)$ (see (37)) and cumulative amount D of cash withdrawals, that is $X = V^{X_0,\varphi,D}$. Its initial value $X_0 = X^g_{0,T}(\xi, D)$ is thus a sensible price (at time 0) of the option for the seller since this amount allows him/her to construct a risky-assets strategy φ, called *hedging* strategy, such that the value of the associated portfolio is equal to ξ at time T, and such that the cash withdrawals perfectly replicate the dividends of the option. We call $X_0 = X^g_{0,T}(\xi, D)$ the *hedging price* at time t of the option. Similarly, for each $t \in [0, T]$, $X_t = X^g_{t,T}(\xi, D)$ is the *hedging price* at time t of the option.

Thus, for each maturity $S \in [0, T]$ and for each pair *payoff-dividend* $(\xi, D) \in L^2(\mathscr{G}_S) \times \mathscr{A}^2_S$, the process $X^g_{\cdot,S}(\xi, D)$ is called the *hedging price* process of the option with maturity S and *payoff-dividend* (ξ, D). This leads to a *pricing* system

$$\mathbf{X}^g : (S, \xi, D) \mapsto X^g_{\cdot,S}(\xi, D), \tag{43}$$

which is generally *nonlinear* with respect to (ξ, D).

We now give some properties of this *nonlinear pricing* system $\mathbf{X}^g$ which generalize those given in [15] to the case with a default jump and dividends.

- **Consistency.** By the flow property for BSDEs, the pricing system $\mathbf{X}^g$ is *consistent*. More precisely, let $S' \in [0,T]$, $\xi \in L^2(\mathscr{G}_{S'})$, $D \in \mathscr{A}^2_{S'}$, and let $S \in [0,S']$. Then, the *hedging price* of the option associated with payoff ξ, cumulative dividend process D and maturity S' coincides with the *hedging price* of the option associated with maturity S, payoff $X^g_{S,S'}(\xi, D)$ and dividend process $(D_t)_{t\leq S}$ (still denoted by D), that is

$$X^g_{\cdot,S'}(\xi, D) = X^g_{\cdot,S}\left(X^g_{S,S'}(\xi, D),\ D\right).$$

- When $g(t,0,0,0) = 0$,[11] then the price of the European option with null payoff and no dividends is equal to 0, that is, for all $S \in [0,T]$, $X^g_{\cdot,S}(0,0) = 0$.

Due the presence of the default, the *nonlinear pricing* system $\mathbf{X}^g$ is not necessarily monotone with respect to the payoff and the dividend. We introduce the following assumption.

Assumption 4 *Assume that there exists a map*

$$\gamma : \Omega \times [0,T] \times \mathbf{R}^4 \to \mathbf{R}\,;\ (\omega, t, y, z, k_1, k_2) \mapsto \gamma_t^{y,z,k_1,k_2}(\omega)$$

$\mathscr{P} \otimes \mathscr{B}(\mathbf{R}^4)$*-measurable, satisfying* $dP \otimes dt$*-a.e., for each* $(y,z,k_1,k_2) \in \mathbf{R}^4$,

$$|\gamma_t^{y,z,k_1,k_2}\sqrt{\lambda_t}| \leq C \quad \textit{and} \quad \gamma_t^{y,z,k_1,k_2} \geq -1, \tag{44}$$

and

$$g(t,y,z,k_1) - g(t,y,z,k_2) \geq \gamma_t^{y,z,k_1,k_2}(k_1 - k_2)\lambda_t \tag{45}$$

(where C *is a positive constant).*

Remark 16 Suppose (λ_t) bounded (as in [12]). Then the first inequality in (44) holds if, for example, γ is bounded.

Recall that λ vanishes after ϑ and $g(t,\cdot)$ does not depend on k on $\{t > \vartheta\}$. Hence, inequality (45) is always satisfied on $\{t > \vartheta\}$. Note that Assumption 4 holds when $g(t,\cdot)$ is non decreasing with respect to k, or when g is $\mathscr{C}^1$ in k with $\partial_k g(t,\cdot) \geq -\lambda_t$.

Before giving some additional properties of the nonlinear pricing system under Assumption 4, we introduce the following partial order relation, defined for each fixed time $S \in [0,T]$, on the set of pairs "payoff-dividends" by: for each $(\xi^1, D^1), (\xi^2, D^2) \in L^2(\mathscr{G}_S) \times \mathscr{A}^2_S$

$$(\xi^1, D^1) \succ (\xi^2, D^2) \quad \text{if} \quad \xi^1 \geq \xi^2 \text{ a.s. and } D^1 - D^2 \text{ is non decreasing.}$$

[11] Note that when the market is perfect, g is given by (36) and thus satisfies $g(t,0,0,0) = 0$.

Loosely speaking, the non decreasing property of $D^1 - D^2$ corresponds to the fact that the dividends paid by the option associated with (ξ^1, D^1) are greater than or equal to those paid by the option associated with (ξ^2, D^2).

Using the comparison theorem for BSDEs with generalized drivers (Theorem 3 (i)), we derive the following properties:

- **Monotonicity.** Under Assumption 4, the nonlinear pricing system $\mathbf{X}^g$ is non decreasing with respect to the payoff and the dividend. More precisely, for all maturity $S \in [0, T]$, for all payoffs $\xi_1, \xi_2 \in L^2(\mathscr{G}_S)$, and cumulative dividend processes $D^1, D^2 \in \mathscr{A}^2_S$, the following property holds:
  If $(\xi^1, D^1) \succ (\xi^2, D^2)$, then we have $X^g_{t,S}(\xi_1, D^1) \geq X^g_{t,S}(\xi_2, D^2), \quad t \in [0, S]$ a.s.[12]
- **Convexity.** Under Assumption 4, if g is convex with respect to (y, z, k), then the nonlinear pricing system $\mathbf{X}^g$ is convex with respect to (ξ, D), that is, for any $\alpha \in [0, 1]$, $S \in [0, T]$, $\xi_1, \xi_2 \in L^2(\mathscr{G}_S)$, $D^1, D^2 \in \mathscr{A}^2_S$, for all $t \in [0, S]$, we have

$$X^g_{t,S}\left(\alpha\xi_1 + (1-\alpha)\xi_2\,,\, \alpha D^1 + (1-\alpha)D^2\right) \leq \alpha\, X^g_{t,S}(\xi_1, D^1) + (1-\alpha)\, X^g_{t,S}(\xi_2, D^2) \quad \text{a.s.}$$

- **Nonnegativity.** Under Assumption 4, when $g(t, 0, 0, 0) \geq 0$, the nonlinear pricing system $\mathbf{X}^g$ is nonnegative, that is, for each $S \in [0, T]$, for all non negative $\xi \in L^2(\mathscr{G}_S)$ and all non decreasing processes $D \in \mathscr{A}^2_S$, we have $X^g_{t,S}(\xi, D) \geq 0$ for all $t \in [0, S]$ a.s.

By the strict comparison theorem (see Theorem 3(ii)), we have the following additional property.

- **No arbitrage.** Under Assumption 4 with $\gamma^{y,z,k_1,k_2}_{\vartheta} > -1$, the nonlinear pricing system $\mathbf{X}^g$ satisfies the *no arbitrage* property: for all maturity $S \in [0, T]$, for all payoffs $\xi^1, \xi^2 \in L^2(\mathscr{G}_S)$, and cumulative dividend processes $D^1, D^2 \in \mathscr{A}^2_S$, for each $t_0 \in [0, S]$, the following holds:
  If $(\xi^1, D^1) \succ (\xi^2, D^2)$ and if the prices of the two options are equal at time t_0, that is, $X^g_{t_0,S}(\xi_1, D^1) = X^g_{t_0,S}(\xi_2, D^2)$ a.s., then, $\xi_1 = \xi_2$ a.s. and $(D^1_t - D^2_t)_{t_0 \leq t \leq S}$ is a.s. constant.[13]

Remark 17 In the perfect market model with default, the driver is given by (36). When $\theta^2_t \leq 1$, then Assumption 4 is satisfied with $\gamma_t^{y,z,k_1,k_2} = -\theta^2_t$, which ensures in particular the **monotonicity** property of the pricing system. Note that when (θ^2_t) is a constant $\theta > 1$ and $P(T \geq \vartheta) > 0$, the pricing system is no longer monotone (see Example 1 with $\delta_t = -r_t$, $\beta_t = -\theta^1_t$ and $\gamma = -\theta$). Moreover, when $\theta^2_t < 1$, then the above **no arbitrage** property holds. This is no longer the case when, for

[12] This property follows from Theorem 3 (i) applied to $g^1 = g^2 = g$ and ξ^1, ξ^2, D^1, D^2. Indeed by Assumption 4, Assumption (25) holds with γ_t replaced by the predictable process $\gamma_t^{Y^2_{t-}, Z^2_t, K^1_t, K^2_t}$.

[13] In other words, the intermediate dividends paid between t_0 and S are equal a.s.

example, $\theta_t^2 = 1$ and $P(T \geq \vartheta) > 0$ (see Example 1 with $\delta_t = -r_t$, $\beta_t = -\theta_t^1$ and $\gamma = -1$).

3.4 The (g, D)-Conditional Expectation $\mathscr{E}^{g,D}$ and $\mathscr{E}^{g,D}$-Martingales

Let g be a λ-admissible driver and let D be an optional singular process belonging to $\mathscr{A}_T^2$.
We define the (g, D)-conditional expectation for each $S \in [0, T]$ and each $\xi \in L^2(\mathscr{G}_S)$ by

$$\mathscr{E}_{t,S}^{g,D}(\xi) := X_{t,S}^{g}(\xi, D), \quad 0 \leq t \leq S.$$

In other terms, $\mathscr{E}_{.,S}^{g,D}(\xi)$ is defined as the first coordinate of the solution of the BSDE associated with terminal time S, *generalized driver* $g(t, y, z, k)dt + dD_t$ and terminal condition ξ.
In the case where $D = 0$, it reduces to the g-conditional expectation $\mathscr{E}^g$ (in the case of default).

Note that $\mathscr{E}_{.,S}^{g,D}(\xi)$ can be defined on the whole interval $[0, T]$ by setting $\mathscr{E}_{t,S}^{g,D}(\xi) := \mathscr{E}_{t,T}^{g^S,D^S}(\xi)$ for $t \geq S$, where $g^S(t, .) := g(t, .)\mathbf{1}_{t \leq S}$ and $D_t^S := D_{t \wedge S}$.
We also define $\mathscr{E}_{.,\tau}^{g,D}(\xi)$ for each stopping time $\tau \in \mathscr{T}_0$ and each $\xi \in L^2(\mathscr{G}_\tau)$ as the solution of the BSDE associated with terminal time T, driver $g^\tau(t, .) := g(t, .)\mathbf{1}_{t \leq \tau}$ and singular process $D_t^\tau := D_{t \wedge \tau}$.

We now give some properties of the (g, D)-conditional expectation which generalize those given in [20] to the case of a default jump and generalized driver.

The (g, D)-conditional expectation $\mathscr{E}^{g,D}$ is **consistent**. More precisely, let τ' be a stopping time in $\mathscr{T}_0$, $\xi \in L^2(\mathscr{G}_{\tau'})$, and let τ be a stopping time smaller or equal to τ'.
We then have $\mathscr{E}_{t,\tau'}^{g,D}(\xi) = \mathscr{E}_{t,\tau}^{g,D}(\mathscr{E}_{\tau,\tau'}^{g,D}(\xi))$ for all $t \in [0, T]$ a.s.

The (g, D)-conditional expectation $\mathscr{E}^{g,D}$ satisfies the following property:
for all $\tau \in \mathscr{T}_0$, $\xi \in L^2(\mathscr{G}_\tau)$, and for all $t \in [0, T]$ and $A \in \mathscr{F}_t$, we have:
$\mathscr{E}_{t,\tau}^{g^A,D^A}(\mathbf{1}_A\xi) = \mathbf{1}_A\mathscr{E}_{t,\tau}^{g,D}(\xi)$ a.s., where $g^A(s, \cdot) = g(s, \cdot)\mathbf{1}_A\mathbf{1}_{]t,T]}(s)$ and $D_s^A := (D_s - D_t)\mathbf{1}_A\mathbf{1}_{s \geq t}$.[14]

Using the comparison theorem for BSDEs with default and generalized drivers (Theorem 3(i)), we derive that, under Assumption 4, the (g, D)-conditional

[14]From this property, we derive the following **Zero-one law**: if $g(\cdot, 0, 0, 0) = 0$, then $\mathscr{E}_{t,\tau}^{g,D^A}(\mathbf{1}_A\xi) = \mathbf{1}_A\mathscr{E}_{t,\tau}^{g,D}(\xi)$ a.s.

expectation $\mathcal{E}^{g,D}$ is **monotone** with respect to ξ. If moreover g is convex with respect to (y, z, k), then $\mathcal{E}^{g,D}$ is convex with respect to ξ.

From the strict comparison theorem (see Theorem 3(ii)), we derive that, under Assumption 4 with $\gamma_\vartheta^{y,z,k_1,k_2} > -1$, $\mathcal{E}^{g,D}$ satisfies the **no arbitrage** property. More precisely, for all $S \in [0, T]$, $\xi^1, \xi^2 \in L^2(\mathcal{G}_S)$, and for all $t_0 \in [0, S]$ and $A \in \mathcal{G}_{t_0}$, we have:
If $\xi^1 \geq \xi^2$ a.s. and $\mathcal{E}^{g,D}_{t_0,S}(\xi_1) = \mathcal{E}^{g,D}_{t_0,S}(\xi_2)$ a.s. on A, then $\xi_1 = \xi_2$ a.s. on A.

The **no arbitrage** property also ensures that when $\gamma_\vartheta^{y,z,k_1,k_2} > -1$, the (g, D)-conditional expectation $\mathcal{E}^{g,D}$ is **strictly monotone**.[15]

We now introduce the definition of an $\mathcal{E}^{g,D}$-martingale which generalizes the classical notion of $\mathcal{E}^g$-martingale.

Definition 4 Let $Y \in \mathcal{S}^2$. The process Y is said to be a $\mathcal{E}^{g,D}$-martingale if $\mathcal{E}^{g,D}_{\sigma,\tau}(Y_\tau) = Y_\sigma$ a.s. on $\sigma \leq \tau$, for all $\sigma, \tau \in \mathcal{T}_0$.

Proposition 4 *For all $S \in [0, T]$, payoff $\xi \in L^2(\mathcal{G}_S)$ and dividend process $D \in \mathcal{A}^2_S$, the associated* hedging price *process $\mathcal{E}^{g,D}_{\cdot,S}(\xi)$ is an $\mathcal{E}^{g,D}$-martingale.*

Moreover, for all $x \in \mathbb{R}$, risky-assets strategy $\varphi \in \mathbb{H}^2 \times \mathbb{H}^2_\lambda$ and cash withdrawal process $D \in \mathcal{A}^2$, the associated wealth process $V^{x,\varphi,D}$ is an $\mathcal{E}^{g,D}$-martingale.

Proof The first assertion follows from the consistency property of $\mathcal{E}^{g,D}$. The second one is obtained by noting that $V^{x,\varphi,D}$ is the solution of the BSDE with *generalized driver* $g(t, \cdot)dt + dD_t$, terminal time T and terminal condition $V_T^{x,\varphi,D}$.

Remark 18 The above result is used in [12, Section 5.4] to study the nonlinear pricing of game options with intermediate dividends in an imperfect financial market with default.

Some examples of market models with default and imperfections or constraints, leading to a nonlinear pricing are given in [7, 8, 13, 14, 19]. We now provide another example.

3.5 *Example: Large Seller Who Affects the Default Probability*

We consider a European option with maturity T, terminal payoff $\xi \in L^2(\mathcal{G}_T)$, and dividend process $D \in \mathcal{A}^2_T$. We suppose that the seller of this option is a *large trader*. More precisely, her hedging strategy (as well as its associated cost) may affect the

[15] In the case without default, it is well-known that, up to a minus sign, the g-conditional expectation $\mathcal{E}^g$ can be seen as a dynamic risk measure (see e.g. [20, 21]). In our framework, we can define a dynamic risk measure ρ^g by setting $\rho^g := -\mathcal{E}^g$ $(= -\mathcal{E}^{g,0})$. This dynamic risk-measure thus satisfies similar properties to those satisfied by $\mathcal{E}^g$.

prices of the risky assets and the default probability. She takes into account these feedback effects in her market model in order to price the option. To the best of our knowledge, the possible impact on the default probability has not been considered in the literature before.

In order to simplify the presentation, we consider the case when the seller's strategy affects only the default intensity. We also suppose in this example that the default intensity is bounded.

We are given a family of probability measures parametrized by V and φ. More precisely, for all $V \in \mathscr{S}^2$, $\varphi \in \mathbb{H}^2 \times \mathbb{H}^2_\lambda$, let $Q^{V,\varphi}$ be the probability measure equivalent to P, which admits $L^{V,\varphi}$ as density with respect to P, where $L^{V,\varphi}$ is the solution of the following SDE:

$$dL_t^{V,\varphi} = L_{t^-}\gamma(t, V_{t^-}, \varphi_t)dM_t; \quad L_0^{V,\varphi} = 1.$$

Here, $\gamma : (\omega, t, y, \varphi_1, \varphi_2) \mapsto \gamma(\omega, t, y, \varphi_1, \varphi_2)$ is a $\mathscr{P} \otimes \mathscr{B}(\mathbf{R}^3)$-measurable function defined on $\Omega \times \mathbf{R}^+ \times \mathbf{R}^3$, bounded, and such that the map $y \mapsto \gamma(\omega, t, y, \varphi_1, \varphi_2)/\varphi_2$ is uniformly Lipschitz. We suppose that $\gamma(t, \cdot) > -1$. Note that by Proposition 3 and Remark 8, the process $L^{V,\varphi}$ is positive and belongs to $\mathscr{S}^2$. By Girsanov's theorem, the process W is a $Q^{V,\varphi}$-Brownian motion and the process $M^{V,\varphi}$ defined as

$$M_t^{V,\varphi} := N_t - \int_0^t \lambda_s(1 + \gamma(s, V_s, \varphi_s))ds = M_t - \int_0^t \lambda_s\gamma(s, V_s, \varphi_s)ds \tag{46}$$

is a $Q^{V,\varphi}$-martingale. Hence, under $Q^{V,\varphi}$, the $\mathbb{G}$-default intensity process is equal to $\lambda_t(1 + \gamma(t, V_t, \varphi_t))$. The process $\gamma(t, V_t, \varphi_t)$ represents the *impact of the seller's strategy on the default intensity* in the case when φ is the seller's risky-assets strategy and V is the value of her portfolio.

The large seller considers the following pricing model. For a fixed pair "wealth/risky-assets strategy" $(V, \varphi) \in \mathscr{S}^2 \times \mathbb{H}^2 \times \mathbb{H}^2_\lambda$, the dynamics of the risky-assets under the probability $Q^{V,\varphi}$ are given by

$$\begin{aligned} dS_t^1 &= S_t^1[\mu_t^1 dt + \sigma_t^1 dW_t]; \\ dS_t^2 &= S_{t^-}^2[\mu_t^2 dt + \sigma_t^2 dW_t - dM_t^{V,\varphi}]. \end{aligned}$$

The value process (V_t) of the portfolio associated with an initial wealth x, a risky-assets strategy φ, and with a cumulative withdrawal process, that the seller chooses to be equal to the dividend process D of the option, must satisfy the following dynamics:

$$dV_t = \left(r_t V_t + \varphi_t'\sigma_t\theta_t^1 - \varphi_t^2\theta_t^2\lambda_t\right)dt - dD_t + \varphi_t'\sigma_t dW_t - \varphi_t^2 dM_t^{V,\varphi}. \tag{47}$$

Note that the dynamics of the wealth (47) can be written

$$dV_t = \left(r_t V_t + \varphi_t' \sigma_t \theta_t^1 - \varphi_t^2 \theta_t^2 \lambda_t + \gamma(t, V_t, \varphi_t)\lambda_t \varphi_t^2\right) dt - dD_t + \varphi_t' \sigma_t dW_t - \varphi_t^2 dM_t. \tag{48}$$

Let us suppose that the large seller has an initial wealth equal to x and follows a risky-assets strategy φ. By the assumptions made on γ, there exists a unique process $V^{x,\varphi}$ satisfying (48) with initial condition $V_0^{x,\varphi} = x$. This model is thus well posed.

Moreover, it can be seen as a particular case of the general model described in Sect. 3.3. Indeed, setting $Z_t = \varphi_t{}' \sigma_t$ and $K_t = -\varphi_t^2$, the dynamics (48) can be written

$$-dV_t = g(t, V_t, Z_t, K_t)dt + dD_t - Z_t dW_t - K_t dM_t, \tag{49}$$

where

$$g(t, y, z, k) = -r_t y - \theta_t^1 z - \theta_t^2 \lambda_t k + \gamma\left(t, y, (\sigma_t^1)^{-1}(z + \sigma_t^2 k), -k\right)\lambda_t k.$$

Assuming that there exists a positive constant C such that inequality (3) holds, g is λ-admissible. We are thus led to the model from Sect. 3.3 associated with this nonlinear driver g. Thus, by choosing this pricing model, the seller prices the option at time t, where $t \in [0, T]$, at the price $X_{t,T}^g(\xi, D)$. In other terms, the seller 's price process[16] will be equal to X, where (X, Z, K) is the solution of the BSDE:

$$-dX_t = g(t, X_t, Z_t, K_t)dt + dD_t - Z_t dW_t - K_t dM_t; \quad X_T = \xi.$$

Moreover, her hedging risky-assets strategy φ will be such that $Z_t = \varphi_t{}' \sigma_t$ and $K_t = -\varphi_t^2$, that is, equal to $\Phi(Z, K)$, where Φ is given by (37).

This model can be easily generalized to the case when the coefficients μ^1, σ^1, μ^2, σ^2 also depend on the hedging cost V (equal to the seller's price X of the option) and on the hedging strategy φ^2.[17]

4 Concluding Remarks

In this paper, we have established properties of BSDEs with default jump and *generalized driver* which involves a finite variational process D. We treat the case

[16] Note that the seller 's price is not necessarily equal to the market price of the option.

[17] The coefficients may also depend on $\varphi = (\varphi^1, \varphi^2)$, but in this case, we have to assume that the map $\Psi : (\omega, t, y, \varphi) \mapsto (z, k)$ with $z = \varphi' \sigma_t(\omega, t, y, \varphi)$ and $k = -\varphi^2$ is one to one with respect to φ, and such that its inverse Ψ_φ^{-1} is $\mathscr{P} \otimes \mathscr{B}(\mathbf{R}^3)$-measurable.

when D is not necessarily predictable and may admit a jump at the default time. This allows us to study nonlinear pricing of European options generating intermediate dividends (with in particular a cashflow at the default time) in complete imperfect markets with default. Due to the default jump, we need an appropriate assumption on the driver g to ensure that the associated nonlinear pricing system $\mathbf{X}^g : (T, \xi, D) \mapsto \mathscr{E}^{g,D}_{\cdot,T}(\xi)$ is monotonous, and a stronger condition to ensure that it satisfies the so-called **no-arbitrage** property. Some complements concerning the nonlinear pricing of European options are given in [13] (cf. Section 4 and Section 5.1). The nonlinear pricing of *American* options (resp. *game* options) in *complete* imperfect markets with default are addressed in [13] (resp. [12]). The case of American options in *incomplete* imperfect financial markets with default is studied in [14].

Appendix

BSDEs with Default Jump in L^p, for $p \geq 2$

For $p \geq 2$, let $\mathscr{S}^p$ be the set of $\mathbb{G}$-adapted RCLL processes φ such that $\mathbb{E}[\sup_{0\leq t\leq T} |\varphi_t|^p] < +\infty$,
$\mathbb{H}^p$ the set of $\mathbb{G}$-predictable processes such that $\|Z\|_p^p := \mathbb{E}\Big[(\int_0^T |Z_t|^2 dt)^{p/2}\Big] < \infty$,
$\mathbb{H}^p_\lambda$ the set of $\mathbb{G}$-predictable processes such that $\|U\|^p_{p,\lambda} := \mathbb{E}\Big[(\int_0^T |U_t|^2 \lambda_t dt)^{p/2}\Big] < \infty$.

Proposition 5 *Let $p \geq 2$ and $T > 0$. Let g be a λ-admissible driver such that $g(t, 0, 0, 0) \in I\!H^p$. Let $\xi \in L^p(\mathscr{G}_T)$. There exists a unique solution (Y, Z, K) in $\mathscr{S}^p \times \mathbb{H}^p \times \mathbb{H}^p_\lambda$ of the BSDE with default (4).*

Remark 19 The above result still holds in the case when there is a $\mathbb{G}$-martingale representation theorem with respect to W and M, even if $\mathbb{G}$ is not generated by W and M.

Proof We now introduce the same arguments as in the proof of Proposition A.2 in [21] together with the arguments used in the proof of Proposition 2.

BSDEs with Default Jump and Change of Probability Measure

Let (β_s) and (γ_s) be two real-valued $\mathbb{G}$-predictable processes such that (β_s) and $(\gamma_s\sqrt{\lambda_s})$ are bounded. Let (ζ_s) be the process satisfying the forward SDE:

$$d\zeta_s = \zeta_{s^-}(\beta_s dW_s + \gamma_s dM_s),$$

with $\zeta_0 = 1$. By Remark 10, we have $\mathbb{E}[\sup_{0\le s\le T} \zeta_s^p] < +\infty$ for all $p \ge 2$. We suppose that $\gamma_\vartheta > -1$ a.s., which, by Remark 8, implies that $\zeta_s > 0$ for all $s \ge 0$ a.s. Let Q be the probability measure equivalent to P which admits ζ_T as density with respect to P on $\mathscr{G}_T$.

By Girsanov's theorem (see [16] Chapter 9.4 Corollary 4.5), the process $W_t^\beta := W_t - \int_0^t \beta_s ds$ is a Q-Brownian motion and the process M^γ defined as

$$M_t^\gamma := M_t - \int_0^t \lambda_s \gamma_s ds = N_t - \int_0^t \lambda_s (1+\gamma_s) ds \tag{50}$$

is a Q-martingale. We now state a representation theorem for $(Q, \mathbb{G})$-local martingales with respect to W^β and M^γ.

Proposition 6 *Let $m = (m_t)_{0\le t\le T}$ be a $(Q, \mathbb{G})$-local martingale. There exists a unique pair of predictable processes (z_t, k_t) such that*

$$m_t = m_0 + \int_0^t z_s dW_s^\beta + \int_0^t k_s dM_s^\gamma \quad 0 \le s \le T \quad \text{a.s.} \tag{51}$$

Proof Since m is a Q-local martingale, the process $\bar{m}_t := \zeta_t m_t$ is a P-local martingale. By the martingale representation theorem (Lemma 1), there exists a unique pair of predictable processes (Z, K) such that

$$\bar{m}_t = \bar{m}_0 + \int_0^t Z_s dW_s + \int_0^t K_s dM_s \quad 0 \le t \le T \quad \text{a.s.}$$

Then, by applying Itô's formula to $m_t = \bar{m}_t (\zeta_t)^{-1}$ and by classical computations, one can derive the existence of (z, k) satisfying (51).

From this result together with Proposition 5 and Remark 19, we derive the following corollary.

Corollary 1 *Let $p \ge 2$ and let $T > 0$. Let g be a λ-admissible driver such that $g(t, 0, 0, 0) \in I\!H_Q^p$. Let $\xi \in L_Q^p(\mathscr{G}_T)$. There exists a unique solution (Y, Z, K) in $\mathscr{S}_Q^p \times \mathbb{H}_Q^p \times \mathbb{H}_{Q,\lambda}^p$ of the BSDE with default:*

$$-dY_t = g(t, Y_t, Z_t, K_t)dt - Z_t W_t^\beta - K_t dM_t^\gamma; \quad Y_T = \xi.$$

Here the spaces $\mathscr{S}_Q^p$, $\mathbb{H}_Q^p$, and $\mathbb{H}_{Q,\lambda}^p$ are defined as $\mathscr{S}^p$, $\mathbb{H}^p$, and $\mathbb{H}_\lambda^p$, by replacing the probability P by Q.

Remark 20 Note that the results given in the Appendix are used in [12] (Section 4.3) to study the nonlinear pricing problem of game options in an imperfect market with default and model uncertainty.

References

1. Ankirchner, S., Blanchet-Scalliet, C., Eyraud-Loisel, A.: Credit risk premia and quadratic BSDEs with a single jump. Int. J. Theor. App. Fin. **13**(7), 1103–1129 (2010)
2. Barles, G., Buckdahn, R., Pardoux, E.: Backward stochastic differential equations and integral-partial differential equations. Stochastics and Stochastics Reports (1995)
3. Bielecki, T., Jeanblanc, M., Rutkowski, M.: Hedging of defaultable claims. In: Bielecki, T.R. et al. (eds.) Paris-Princeton Lectures on Mathematical Finance. Lecture Notes in Mathematics 1847, pp. 1–132. Springer (2004). ISBN:3-540-22266-9. https://doi.org/10.1007/b98353
4. Bielecki, T., Jeanblanc, M., Rutkowski, M.: PDE approach to valuation and hedging of credit derivatives. Quant. Finan. **5**, 257–270 (2005)
5. Bielecki, T., Crepey, S., Jeanblanc, M., Rutkowski, M.: Defaultable game options in a hazard process model. Int. J. Stoch. Anal. **2009**, 1–33 (2009)
6. Blanchet-Scalliet, C., Eyraud-Loisel, A., Royer-Carenzi, M.: Hedging of defaultable contingent claims using BSDE with uncertain time horizon. Le bulletin français d'actuariat **20**(10), 102–139 (2010)
7. Brigo, D., Franciscello, M., Pallavicini, A.: Analysis of nonlinear valuation equations under credit and funding effects. In: Glau, K., Grbac, Z., Scherer, M., Zagst, R. (eds.) Innovations in Derivative Markets, Springer Proceedings in Mathematics and Statistics, vol. 165, pp. 37–52. Springer, Heidelberg (2016)
8. Crepey, S.: Bilateral counterparty risk under funding constraints Part I: CVA. Pricing. Math. Financ. **25**(1), 1–22 (2015)
9. Dellacherie, C., Meyer, P.-A.: Probabilités et Potentiel, Chaps. I–IV. Nouvelle édition. Hermann. **MR**0488194 (1975)
10. Delong, L.: Backward Stochastic Differential Equations with Jumps and Their Actuarial and Financial Applications. EAA Series. Springer, London/New York (2013)
11. Dumitrescu, R., Quenez, M.-C., Sulem, A.: Generalized dynkin games and doubly reflected BSDEs with jumps. Electron. J. Probab. **21**(64), 32 (2016)
12. Dumitrescu, R., Quenez, M.-C., Sulem, A.: Game options in an imperfect market with default. SIAM J. Financ. Math. **8**, 532–559 (2017)
13. Dumitrescu, R., Quenez, M.-C., Sulem, A.: American options in an imperfect complete market with default. ESAIM Proc. Surv. (2018, to appear)
14. Grigorova, M., Quenez, M.-C., Sulem, A.: $\mathscr{E}^g$-pricing of American options in incomplete markets with default. (2018), manuscript
15. El Karoui, N., Quenez, M.C.: Non-linear pricing theory and backward stochastic differential equations. In: Runggaldier, W.J. (ed.) Collection. Lectures Notes in Mathematics 1656, Bressanone, 1996. Springer (1997)
16. Jeanblanc, M., Yor, M., Chesney, M.: Mathematical Methods for Financial Markets. Springer Finance, London (2009)
17. Korn, R.: Contingent claim valuation in a market with different interest rates. Math. Meth. Oper. Res. **42**, 255–274 (1995)
18. Kusuoka, S.: A remark on default risk models. Adv. Math. Econ. **1**, 69–82 (1999)
19. Lim, T., Quenez, M.-C.: Exponential utility maximization in an incomplete market with defaults. Electron. J. Probab. **16**(53), 1434–1464 (2011)
20. Peng, S.: Nonlinear Expectations, Nonlinear Evaluations and Risk Measures. Lecture Notes in Mathematics, pp. 165–253. Springer, Berlin (2004)
21. Quenez, M.-C., Sulem, A.: BSDEs with jumps, optimization and applications to dynamic risk measures. Stoch. Process. Appl. **123**, 3328–3357 (2013)
22. Royer, M.: Backward stochastic differential equations with jumps and related non-linear expectations. Stoch. Process. Appl. **116**, 1358–1376 (2006)

The Faà di Bruno Hopf Algebra for Multivariable Feedback Recursions in the Center Problem for Higher Order Abel Equations

Kurusch Ebrahimi-Fard and W. Steven Gray

Abstract Poincaré's center problem asks for conditions under which a planar polynomial system of ordinary differential equations has a center. It is well understood that the Abel equation naturally describes the problem in a convenient coordinate system. In 1990, Devlin described an algebraic approach for constructing sufficient conditions for a center using a linear recursion for the generating series of the solution to the Abel equation. Subsequent work by the authors linked this recursion to feedback structures in control theory and combinatorial Hopf algebras, but only for the lowest degree case. The present work introduces what turns out to be the nontrivial multivariable generalization of this connection between the center problem, feedback control, and combinatorial Hopf algebras. Once the picture is completed, it is possible to provide generalizations of some known identities involving the Abel generating series. A linear recursion for the antipode of this new Hopf algebra is also developed using coderivations. Finally, the results are used to further explore what is called the composition condition for the center problem.

1 Introduction

The classical center problem first studied by Henri Poincaré [38] considers a system of planar ordinary differential equations

$$\frac{dx}{dt} = X(x, y), \qquad \frac{dy}{dt} = Y(x, y), \tag{1}$$

K. Ebrahimi-Fard
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: kurusch.ebrahimi-fard@ntnu.no

W. Steven Gray (✉)
Old Dominion University, Norfolk, VA, USA
e-mail: sgray@odu.edu

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_10

where X, Y are homogeneous polynomials with a linear part of center type. The equilibrium at the origin is a center if it is contained in an open neighborhood U having no other equilibria, and every trajectory of system (1) in U is closed with the same period ω. The problem is usually studied in its canonical form via a reparametrization that transforms (1) into the Abel equation

$$\dot{z}(t) = v_1(t)z^2(t) + v_2(t)z^3(t), \tag{2}$$

where v_1 and v_2 are continuous real-valued functions [3, 9, 35]. In this setting, the origin $z = 0$ is a center if $z(0) = z(\omega) = r$ for $r > 0$ sufficiently small and $\omega > 0$ fixed. The center problem is to determine the largest class of functions v_1 and v_2 that will render $z = 0$ a center.

An algebraic approach to the center problem was first proposed by Devlin in 1990 [10, 11], which was based on the work of Alwash and Lloyd [3, 35]. In modern parlance, Devlin's method was to first write the solution of the Abel equation (2) in terms of a *Chen–Fliess functional expansion* or *Fliess operator* [18, 19] whose coefficients are parameterized by r. A Fliess operator is simply a weighted sum of iterated integrals of v_1 and v_2 indexed by words in the noncommuting symbols x_1 and x_2, respectively. The concept is widely used, for example, in control theory to describe the input-output map of a system modeled in terms of ordinary differential equations. (For readers not familiar with this subject, the following references provide a good overview [18, 19, 32, 33, 37, 42–46].) Devlin showed that the generating series for his particular Fliess operator with $r = 1$, which is a formal power series c_A over words in the alphabet $X = \{x_1, x_2\}$, can be decomposed as

$$c_A = \sum_{n=1}^{\infty} c_A(n), \tag{3}$$

where the polynomials $c_A(n)$, $n \geq 1$ satisfy the linear recursion

$$c_A(n) = (n-1)c_A(n-1)x_1 + (n-2)c_A(n-2)x_2, \quad n \geq 2$$

with $c_A(1) = 1$ and $c_A(0) = 0$. Here $\deg(x_i) := i$, and each letter x_i encodes the contribution of v_i to the series solution of (2). His derivation used the underlying shuffle algebra induced by products of iterated integrals rather than the fact that the operator coefficients are differentially generated from the vector fields in the Abel equation (2) [18, 32, 37]. Devlin also provided a recursion for the higher-order Abel equation

$$\dot{z}(t) = \sum_{i=1}^{m} v_i(t)z^{i+1}(t), \quad m \geq 2, \tag{4}$$

though the calculations become somewhat intractable. Using such recursions, it was then possible to synthesize various sufficient conditions on the v_i under which the

origin was a center. This included a generalization of the *composition condition* in [3]. The latter states that a sufficient condition for a center is the existence of a differentiable function q such that $q(\omega) = q(0)$ for some $\omega > 0$ and

$$v_i(t) = \bar{v}_i(q(t))\dot{q}(t), \quad i = 1, \ldots, m, \tag{5}$$

where the $\bar{v}_i$ are continuous functions. For a time it was conjectured that this condition was also a necessary condition for a center if certain constraints were imposed on the v_i, for example, if they were polynomial functions of $\cos \omega t$ and $\sin \omega t$. However, a counterexample to this claim was later given by Alwash in [1]. It is still believed, however, to be a necessary condition when the v_i are polynomials. This is now called the *composition conjecture* (see [2, 5–7, 47] and the references in the survey article [22]).

Recently, the authors revisited Devlin's method in a combinatorial Hopf algebra setting in light of the fact that the Abel equation was found to play a central role in determining the radius of convergence of feedback connected Fliess operators as shown in Fig. 1 [41]. This recursive structure is described by the feedback equation

$$y(t) = F_c[v_1(t) + F_d[y(t)]],$$

which by a suitable choice of generating series c and d involving an arbitrary function $v_2(t)$ can be written directly in the form

$$\begin{aligned} \dot{z}(t) &= z^2(t)[v_1(t) + v_2(t)y(t)] \\ &= v_1(t)z^2(t) + v_2(t)z^3(t), \end{aligned}$$

where $y(t) = z(t)$. It was shown in [14] that the decomposition (3) is exactly the sum of the graded components of a Hopf algebra antipode applied to the formal power series $-c_F$, where

$$c_F = \sum_{k=0}^{\infty} k!\, x_1^k$$

is the Ferfera series, that is, the generating series for solution of the equation $\dot{z} = z^2 u$, $z(0) = 1$ [15, 16]. The link is made using the Hopf algebra of output feedback

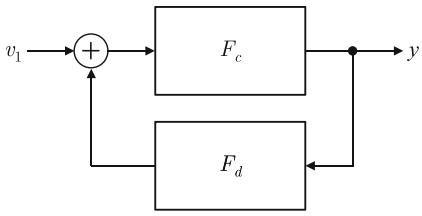


Fig. 1 Feedback connection of Fliess operators F_c and F_d

which encodes the *composition* of iterated integrals rather than their products [12, 23, 29]. As a consequence, another algebraic structure at play in Devlin's approach beyond the shuffle algebra is a Faà di Bruno type Hopf algebra. Now it is a standard theorem that the antipode of every connected graded Hopf algebra can be computed recursively [17, 36]. This fact was exploited, for example, in the authors' application of the output feedback Hopf algebra to compute the feedback product, a device used to compute the generating series for the Fliess operator representation of the interconnection shown in Fig. 1 [12, 27]. But somewhat surprisingly it was also shown in [14] that for this Hopf algebra the antipode could be computed *in general* using a linear recursion of Devlin type. This method has been shown empirically to be more efficient than all existing methods for computing the antipode [4], which is useful in control applications [13, 26, 30, 31]. What was not evident, however, was how all of these ideas could be related for higher order Abel equations, i.e., Eq. (4) when $m > 2$.

The goal of this paper is to present what turns out to be the nontrivial generalization of the connection between the center problem, control theory, and combinatorial Hopf algebras for higher order Abel equations. It requires a new class of matrix-valued Fliess operators with a certain Toeplitz structure in order to provide the proper grading. In addition, a new type of multivariable output feedback Hopf algebra is needed, one which is distinct from that described in [12, 23, 29] and is more closely related to the *output affine feedback* Hopf algebra introduced in [28] for the $m = 2$ case with $v_2 = 1$ (so effectively the single-input–single-output case) to describe *multiplicative* output feedback. Once the picture is completed, it is possible to provide higher order extensions of some known identities for the Abel generating series, c_A. A linear recursion for the antipode of this new Hopf algebra is also developed using coderivations. Finally, a new sufficient condition for a center is given inspired by viewing the Abel equation in terms of a feedback condition. This in turn provides another way of interpreting the composition condition.

2 Linear Recursions for Differentially Generated Series and Their Inverses

The starting point is to show how any formal power series whose coefficients are differentially generated by a set of analytic vector fields can be written in terms of a linear recursion, as can its inverse in a certain compositional sense. This implicitly describes a group that will be utilized in the next section to describe recursions derived from feedback systems.

Consider the set of formal power series $\mathbb{R}\langle\langle X\rangle\rangle$ over the set of words X^* generated by an alphabet of noncommuting symbols $X = \{x_1, \ldots, x_m\}$. Elements of X are called *letters*, and *words* over X consist of finite sequences of letters, $\eta = x_{i_1} \cdots x_{i_k} \in X^*$. The *length* of a word η is denoted $|\eta|$ and is equivalent to the number of letters it contains. When viewed as a graded vector space, where

$\deg(x_i) := i$ and $\deg(\mathrm{e}) := 0$ with e denoting the empty word $\emptyset \in X^*$, any $c \in \mathbb{R}\langle\langle X\rangle\rangle$ can be uniquely decomposed into its homogeneous components $c = \sum_{n\geq 1} c(n)$ with $\deg(c(n)) = n-1$, $n \geq 1$. In particular, if X^k is the set of all words of length k, then $c(1) = \langle c, \mathrm{e}\rangle \mathrm{e}$ and

$$c(n) = \sum_{i=1}^{\min(m,n-1)} \sum_{\eta\in X^{n-1-i}} \langle c, \eta x_i\rangle\, \eta x_i, \quad n \geq 2. \tag{6}$$

A series $c \in \mathbb{R}\langle\langle X\rangle\rangle$ is said to be *differentially generated* if there exists a set of analytic vector fields $\{g_1, g_2, \ldots, g_m\}$ defined on a neighborhood W of $z_0 \in \mathbb{R}^n$ and an analytic function $h : W \to \mathbb{R}$ such that for every word η in X^* the corresponding coefficient of c can be written as

$$\langle c, \eta\rangle = L_{g_{j_1}} \cdots L_{g_{j_k}} h(z_0), \quad \eta = x_{j_k} \cdots x_{j_1},$$

where the Lie derivative of h with respect to g_j is defined as the linear operator

$$L_{g_j} h : W \to \mathbb{R} : z \mapsto L_{g_j} h(z) := \frac{\partial h}{\partial z}(z)\, g_j(z).$$

The tuple $(g_1, g_2, \ldots, g_m, z_0, h)$ will be referred to as a *generator* of c. It follows directly that $c(n) = P_{n-1}(z_0)$, where $P_0(z_0) = h(z_0)\mathrm{e}$ and for $n > 0$, $P_n(z) := \sum_{\eta\in X^n} L_{g_\eta} h(z)\, \eta$, with $L_{g_\eta} := L_{g_{j_1}} \cdots L_{g_{j_k}}$, and (6) can be rewritten as the linear recursion

$$P_n(z_0) = \sum_{i=1}^{\min(m,n)} L_{g_i} P_{n-i}(z_0) x_i, \quad n \geq 1. \tag{7}$$

In this case the grading on $\mathbb{R}\langle\langle X\rangle\rangle$ can be encoded in the sequence $P_n(z_0)$, $n \geq 1$, by assigning degrees to the vector fields, namely, $\deg(g_i) := \deg(x_i) = i$, $i = 1, \ldots, m$.

Example 1 Suppose $m = 1$, $g_1(z) = z^2$, $z_0 = 1$, and $h(z) = z$. Then $c(1) = P_0(1) = h(1)\mathrm{e} = \mathrm{e}$ and

$$c(n) = P_{n-1}(1) = L_{z^2} P_{n-2}(1) x_1 = (n-1) P_{n-2}(1) x_1 = (n-1) c(n-1) x_1, \quad n \geq 2.$$

In which case,

$$\sum_{n=1}^{\infty} c(n) = \sum_{n=1}^{\infty} (n-1)!\, x_1^{n-1} = \sum_{n=0}^{\infty} n!\, x_1^n =: c_F.$$

This is the well studied generating series of Ferfera [15, 16].

Now suppose $d \in \mathbb{R}\langle\langle X\rangle\rangle$ is differentially generated, and consider the corresponding *Chen–Fliess series* or *Fliess operator*

$$F_d[u](t) := \sum_{\eta \in X^*} \langle d, \eta \rangle \, E_\eta[u](t, t_0),$$

where $E_\eta[u]$ is defined inductively for each word $\eta \in X^*$ as an iterated integral over the *controls* $u := (u_1(t), \ldots, u_m(t))$, $u_i : [t_0, t] \to \mathbb{R}$, by $E_\emptyset[u] := 1$ and

$$E_{x_i \bar{\eta}}[u](t, t_0) := \int_{t_0}^{t} u_i(\tau) E_{\bar{\eta}}[u](\tau, t_0) \, d\tau$$

with $x_i \in X$, $\bar{\eta} \in X^*$. If $u \in L_1^m[t_0, t_0 + T]$, that is, u is measurable with finite L_1-norm, $\|u\|_{L_1} := \max\{\|u_i\|_1 : 1 \leq i \leq m\} < R$, then the analyticity of the generator for d is sufficient to guarantee that the Fliess operator $F_d[u](t)$ converges absolutely and uniformly on $[0, T]$ for sufficiently small $R, T > 0$ [24]. Suppose next that $d = (d_1, \ldots, d_{m-1})$ is a family of series $d_i \in \mathbb{R}\langle\langle X\rangle\rangle$, $i = 1, \ldots, m-1$ which are differentially generated by $(g_1, \ldots, g_m, z_0, h_1, \ldots, h_{m-1})$, and define the associated *Toeplitz matrix*

$$d_{\text{Toep}} := \begin{bmatrix} 1 & d_1 & d_2 & \cdots & d_{m-1} \\ 0 & 1 & d_1 & \cdots & d_{m-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & d_1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = I + \sum_{i=1}^{m-1} d_i N^i,$$

where $I \in \mathbb{R}^{m \times m}$ is the identity matrix, and $N \in \mathbb{R}^{m \times m}$ is the nilpotent matrix consisting of zero entries except for a super diagonal of ones. The *Toeplitz affine Fliess operator* is taken to be $y = F_{d_\delta}[u] := F_{d_{\text{Toep}}}[u]u$, which can be written in expanded form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & F_{d_1}[u] & F_{d_2}[u] & \cdots & F_{d_{m-1}}[u] \\ 0 & 1 & F_{d_1}[u] & \cdots & F_{d_{m-2}}[u] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & F_{d_1}[u] \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix}.$$

Note in particular that $0_{\text{Toep}} = I$ so that $F_{0_\delta}[u] = u$. The operator F_{d_δ} is *realized* by the analytic state space system

$$\dot{z} = \sum_{i=1}^{m} g_i(z) u_i, \quad z(0) = z_0 \tag{8a}$$

$$y = H(z)u, \tag{8b}$$

where

$$H = \begin{bmatrix} 1 & h_1 & h_2 & \cdots & h_{m-1} \\ 0 & 1 & h_1 & \cdots & h_{m-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = I + \sum_{i=1}^{m-1} h_i N^i, \tag{9}$$

in the sense that on some neighborhood W of z_0, (8a) has a well defined solution $z(t)$ on $[t_0, t_0+T]$ and $y = F_{d_{\text{Toep}}}[u]u = H(z)u$ on this same interval. Since the Toeplitz matrix H is always invertible and Toeplitz, it follows that the inverse operator $u = F_{d_\delta^{-1}}[y] := F_{d_{\text{Toep}}^{-1}}[y]y$ is another Toeplitz affine Fliess operator realized by the state space system

$$\dot{z} = \sum_{i=1}^{m} g_i(z)[H^{-1}(z)y]_i, \quad z(0) = z_0 \tag{10a}$$

$$u = H^{-1}(z)y. \tag{10b}$$

so that $F_{d_\delta} \circ F_{d_\delta^{-1}} = F_{d_\delta^{-1}} \circ F_{d_\delta} = I$. (Here $[y]_i$ denotes the i component of $y \in \mathbb{R}^m$.) The generating series for the inverse operator, $d^{-1} = (d_1^{-1}, \ldots, d_{m-1}^{-1})$, is differentially generated by $(\tilde{g}_1, \ldots, \tilde{g}_m, z_0, \tilde{h}_1, \ldots, \tilde{h}_{m-1})$, where $\tilde{g}_i := \sum_{j=1}^m g_j H_{ji}^{-1} = g_i + \sum_{j=1}^{i-1} g_{i-j}\tilde{h}_j$ with $\tilde{h}_j := H_{1,1+j}^{-1}$.

Example 2 For the case where $m = 3$, system (10) becomes

$$\dot{z} = g_1 y_1 + (g_2 - g_1 h_1)y_2 + (g_3 - g_2 h_1 + g_1(h_1^2 - h_2))y_3, \quad z(0) = z_0 \tag{11a}$$

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1 & -h_1 & h_1^2 - h_2 \\ 0 & 1 & -h_1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{11b}$$

In which case,

$$F_{d_{\text{Toep}}^{-1}}[y] = \begin{bmatrix} 1 & F_{d_1^{-1}}[y] & F_{d_2^{-1}}[y] \\ 0 & 1 & F_{d_1^{-1}}[y] \\ 0 & 0 & 1 \end{bmatrix},$$

where $d^{-1} = (d_1^{-1}, d_2^{-1})$ is generated by $(\tilde{g}_1, \tilde{g}_2, \tilde{g}_3, z_0, \tilde{h}_1, \tilde{h}_2)$ with $\tilde{g}_1 := g_1$, $\tilde{g}_2 := g_2 - g_1 h_1$, $\tilde{g}_3 := g_3 - g_2 h_1 + g_1(h_1^2 - h_2)$, $\tilde{h}_1 = -h_1$ and $\tilde{h}_2 = h_1^2 - h_2$. If coordinate functions are defined as linear maps on $\mathbb{R}^2\langle\langle X\rangle\rangle$ by

$$a_\eta^i(d) := (d_i, \eta) = L_{g_\eta} h_i(z_0), \quad \eta \in X^*, \quad i = 1, 2,$$

and S is defined as a mapping on $\mathbb{R}\langle X\rangle$ seen as the dual space of $\mathbb{R}\langle\langle X\rangle\rangle$, so that

$$(S(a_\eta^i))(d) := (d_i^{-1}, \eta) = L_{\tilde{g}_\eta} \tilde{h}_i(z_0), \quad \eta \in X^*, \quad i = 1, 2,$$

then the coordinates, i.e., coefficients of the inverse series are described compactly by the following polynomials:

$$S(a_{\mathrm{e}}^1) = -a_{\mathrm{e}}^1 \tag{12a}$$

$$S(a_{\mathrm{e}}^2) = -a_{\mathrm{e}}^2 + a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 \tag{12b}$$

$$S(a_{x_1}^1) = -a_{x_1}^1 \tag{12c}$$

$$S(a_{x_2}^1) = -a_{x_2}^1 + a_{x_1}^1 a_{\mathrm{e}}^1 \tag{12d}$$

$$S(a_{x_1}^2) = -a_{x_1}^2 + 2a_{x_1}^1 a_{\mathrm{e}}^1 \tag{12e}$$

$$S(a_{x_3}^1) = -a_{x_3}^1 + a_{x_2}^1 a_{\mathrm{e}}^1 - a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 + a_{x_1}^1 a_{\mathrm{e}}^2 \tag{12f}$$

$$S(a_{x_2}^2) = -a_{x_2}^2 + 2a_{x_2}^1 a_{\mathrm{e}}^1 - 2a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 + a_{x_1}^2 a_{\mathrm{e}}^1 \tag{12g}$$

$$\begin{aligned} S(a_{x_3}^2) = &-a_{x_3}^2 + 2a_{x_3}^1 a_{\mathrm{e}}^1 - 2a_{x_2}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 + a_{x_2}^2 a_{\mathrm{e}}^1 - a_{x_1}^2 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 + a_{x_1}^2 a_{\mathrm{e}}^2 + \\ &2a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 - 2a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^2 \end{aligned} \tag{12h}$$

$$\vdots$$

It is not obvious in general whether the generators for the inverse series d_i^{-1} will necessarily satisfy a linear recursion of the form (7). This is contingent on whether the new vector fields $\tilde{g}_i$ are consistent with the grading on $\mathbb{R}\langle\langle X\rangle\rangle$, that is, whether $\deg(\tilde{g}_i) = \deg(g_i)$, $i = 1, \ldots, m$. The next theorem gives a sufficient condition under which the upper triangular Toeplitz structure of H in (9) guarantees this property.

Theorem 1 *Given any Toeplitz matrix of the form* (9) *and a set of vector fields* g_i, $i = 1, \ldots, m$ *with* $\deg(g_i) = i$, *it follows that* $\tilde{g}_i := \sum_{j=1}^m g_j H_{ji}^{-1}$ *has the property* $\deg(\tilde{g}_i) = \deg(g_i)$ *provided* $\deg(h_i) := \deg(g_i) = i$, $i = 1, \ldots, m-1$.

Proof First observe that

$$H^{-1} = \left(I + \sum_{i=1}^{m-1} h_i N^i \right)^{-1} = \sum_{n=0}^{m-1} (-1)^n \left(\sum_{i=1}^{m-1} h_i N^i \right)^n,$$

using the fact that $N^n = 0, n \geq m$. Now applying the multinomial theorem gives

$$H^{-1} = I + \sum_{j=1}^{m-1} \left[\sum_{k=1}^{j} (-1)^k k! \sum_{\substack{k_1+k_2+\cdots+k_j=k \\ k_1+2k_2+\cdots+jk_j=j}} \frac{1}{k_1! \cdots k_j!} h_1^{k_1} \cdots h_j^{k_j} \right] N^j. \tag{13}$$

$$=: I + \sum_{j=1}^{m-1} \tilde{h}_j N^j.$$

This means that $\deg(\tilde{h}_j) = \deg(h_1^{k_1} \cdots h_j^{k_j}) = k_1 + 2k_2 + \cdots + jk_j = j$. Therefore, since $\tilde{g}_i = \sum_{j=1}^{m} g_j H_{ji}^{-1} = \sum_{j=0}^{i} g_{i-j} \tilde{h}_j (\tilde{h}_0 := 1)$ and

$$\deg \left(g_{i-j} \tilde{h}_j \right) = \deg \left(g_{i-j} \right) + \deg(\tilde{h}_j) = (i - j) + j = i,$$

it follows that $\deg(\tilde{g}_i) = i, i = 1, \ldots, m$ as required.

Example 3 Reconsider Example 2 in the particular case where $(g_1, g_2, g_3, z_0, h_1, h_2) = (z^2, 0, 0, 1, -z, 0)$ so that $d = (-c_F, 0)$. This is an embedding of Example 1 into the case where $m = 3$. The series $d^{-1} = (d_1^{-1}, d_2^{-1})$ has the generator $(\tilde{g}_1, \tilde{g}_2, \tilde{g}_3, z_0, \tilde{h}_1, \tilde{h}_2) = (z^2, z^3, z^4, 1, z, z^2)$. The system (11) reduces to the Abel system

$$\dot{z} = z^2 y_1 + z^3 y_2 + z^4 y_3, \quad z(0) = 1$$

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1 & z & z^2 \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix},$$

and therefore, $c_{A,3} := d_1^{-1}$. Using (7) with the generator for d_1^{-1}, the Abel generating series $c_{A,3}$ can also be written as $c_{A,3} = \sum_{n \geq 1} c_{A,3}(n)$, where $c_{A,3}(1) = P_0(1) = \mathrm{e}$ and

$$c_{A,3}(n) = P_{n-1}(1) = L_{z^2} P_{n-2}(1) x_1 + L_{z^3} P_{n-3}(1) x_2 + L_{z^4} P_{n-4}(1) x_3, \quad n \geq 2$$

($P_n(1) := 0$ for $n < 0$). A polynomial recursion follows from proving the identity $L_{z^{i+1}} P_{n-i-1}(1) = (n-i)P_{n-i-1}(1)$, $i = 1, 2, 3$, so that

$$c_{A,3}(n) = (n-1)c_{A,3}(n-1)x_1+(n-2)c_{A,3}(n-2)x_2+(n-3)c_{A,3}(n-3)x_3, \quad n \geq 2$$

($c_{A,3}(n) = 0$ for $n < 1$). The first few of these polynomials are:

$$c_{A,3}(1) = 1$$

$$c_{A,3}(2) = x_1$$

$$c_{A,3}(3) = 2x_1x_1 + x_2$$

$$c_{A,3}(4) = 6x_1x_1x_1 + 3x_2x_1 + 2x_1x_2 + x_3$$

$$c_{A,3}(5) = 24x_1x_1x_1x_1 + 12x_2x_1x_1 + 8x_1x_2x_1 + 4x_3x_1 + 6x_1x_1x_2 + 3x_2x_2 + 2x_1x_3.$$

Note that each $c_{A,3}(n)$ consists only of words of degree $n-1$. These polynomials were first identified by Devlin in [10]. The example can be generalized to any $m \geq 2$ so that

$$c_{A,m} = (I - c_F N)_1^{-1}, \tag{14}$$

and the corresponding Abel series $c_{A,m} = \sum_{n\geq 1} c_{A,m}(n)$ can be computed from the recursion

$$c_{A,m}(n) = \sum_{i=1}^{m} (n-i)c_{A,m}(n-i)x_i, \quad n \geq 2,$$

with $c_{A,m}(1) = 1$ and $c_{A,m}(n) = 0$ for $n < 1$.

It is interesting to note that the construction above has some elements in common with the Faà di Bruno Hopf algebra $\mathscr{H}_{FdB} = (\mu, \Delta_{FdB})$ for the group $\mathscr{G}_{diff}$ of diffeomorphisms h on $\mathbb{R}$ satisfying $h(0) = 0$, $\dot{h}(0) = 1$. See [17] for details. First observe that (13) can also be written in terms of the partial exponential Bell polynomials

$$B_{j,k}(t_1, \ldots, t_l) := \sum_{\substack{k_1+k_2+\cdots+k_l=k \\ k_1+2k_2+\cdots+lk_l=j}} \frac{j!}{k_1!\cdots k_l!} \left(\frac{t_1}{1!}\right)^{k_1} \cdots \left(\frac{t_l}{l!}\right)^{k_l},$$

where $l = j - k + 1$, using the Faà di Bruno formula

$$f(h(t)) = \sum_{j=1}^{\infty} \sum_{k=1}^{j} \beta_k B_{j,k}(\alpha_1, \ldots, \alpha_{j-k+1}) \frac{t^j}{j!}$$

with $f(t) := \sum_{n=1}^{\infty} \beta_n t^n/n!$ and $h(t) := \sum_{n=1}^{\infty} \alpha_n t^n/n!$. Specifically, setting

$$f(t) = \frac{1}{1+t} - 1 = \sum_{n=1}^{\infty} (-1)^n n! \frac{t^n}{n!}$$

$$h(t) = \sum_{n=1}^{m-1} n! h_n \frac{t^n}{n!}$$

gives

$$H^{-1} = I + f(h(N)) = I + \sum_{j=1}^{m-1} \sum_{k=1}^{j} (-1)^k \left[\frac{k!}{j!} B_{j,k}(h_1, 2!h_2, \ldots, (j-k+1)!h_{j-k+1}) \right] N^j.$$

The expressions in brackets above, i.e., the partial *ordinary* Bell polynomials, are used in [8] to define a variation (flipped/co-opposite version) of the coproduct $\bar{\Delta}_{FdB}$ on $\mathscr{H}_{FdB}$ (see equations (4.1)–(4.2) in this citation). A faithful representation of the group $\mathscr{G}_{diff}$ is

$$M_h := \left[\frac{k!}{j!} B_{j,k}(h_1, 2!h_2, \ldots, (j-k+1)!h_{j-k+1}) \right]^T$$

$$= \begin{bmatrix} h_1 & h_2 & h_3 & h_4 & h_5 & \cdots \\ 0 & h_1^2 & 2h_1h_2 & h_2^2 + 2h_1h_3 & 2h_2h_3 + 2h_1h_4 & \cdots \\ 0 & 0 & h_1^3 & 3h_1^2h_2 & 3h_1h_2^2 + 3h_1^2h_3 & \cdots \\ 0 & 0 & 0 & h_1^4 & 4h_1^3h_2 & \cdots \\ 0 & 0 & 0 & 0 & h_1^5 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

(cf. [21]) and

$$H^{-1} = I + \sum_{j=1}^{m-1} [\mathbb{1} M_h]_j N^j = I + \sum_{j=1}^{m-1} \tilde{h}_j N^j,$$

where $\mathbb{1} := [-1 \; 1 \; -1 \; \cdots]$. Therefore, defining the coordinate functions $a_i(h) = h_i$, $i \geq 1$, it follows that $\tilde{h}_i = \mu(\bar{\Delta}_{FdB} a_i(\mathbb{1}, h))$. For example,

$$\begin{aligned} \tilde{h}_3 &= \mu(\bar{\Delta}_{FdB}(a_3)(\mathbb{1}, h)) \\ &= \mu((a_1 \otimes a_3 + a_2 \otimes 2a_1a_2 + a_3 \otimes a_1^3)(\mathbb{1}, h)) \\ &= -h_3 + 2h_1h_2 - h_1^3. \end{aligned}$$

Further observe, setting $h_1 = 1$, that the antipode S_{FdB} of $\mathscr{H}_{FdB}$ can be identified from the top row of

$$M_h^{-1} = \begin{bmatrix} 1 & -h_2 & 2h_2^2 - h_3 & -5h_2^3 + 5h_2h_3 - h_4 & 14h_2^4 - 21h_2^2h_3 + 3h_3^2 + 6h_2h_4 - h_5 & \cdots \\ 0 & 1 & -2h_2 & 5h_2^2 - 2h_3 & -14h_2^3 + 12h_2h_3 - 2h_4 & \cdots \\ 0 & 0 & 1 & -3h_2 & 9h_2^2 - 3h_3 & \cdots \\ 0 & 0 & 0 & 1 & -4h_2 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

That is, using a standard expression for S_{FdB} (see [17]), the $(j+1)th$ entry in the top row of M_h^{-1} is given by

$$(S_{FdB}(a_{j+1}))(h) = \sum_{k=1}^{j} (-1)^k B_{j+k,k}(0, 2!h_2, 3!h_3, \ldots, (j+1)!h_{j+1}), \quad j \geq 1.$$

The assertion to be explored in Sect. 4 is that this construction has deeper connections to another kind of Faà di Bruno type Hopf algebra, one that is derived from a group of Fliess operators and used to described their feedback interconnection. In fact, the compositional inverse described above corresponds to the group inverse.

3 Devlin's Polynomials and Feedback Recursions

It was shown in [14] that the Devlin polynomials describing the Abel generating series $c_{A,m}$ when $m = 2$ can be related to a certain feedback structure commonly encountered in control theory. This in turn led to a Hopf algebra interpretation of these polynomials since feedback systems have been characterized in such terms in [12, 23, 26, 27, 29]. In this section, the generalization of the theory is given for any $m \geq 2$. This will again provide a Hopf algebra interpretation of Devlin's polynomials as well as a shuffle formula for the Abel series which is distinct from that derived directly from the Abel equation, namely, the non-linear recursion

$$c_{A,m} = 1 + \sum_{i=1}^{m} x_i c_{A,m}^{\sqcup\!\sqcup\, i+1}, \quad m \geq 2,$$

where $c_{A,m}^{\sqcup\!\sqcup\, i}$ denotes the i-th shuffle power of $c_{A,m}$. Recall that $\mathbb{R}\langle\langle X \rangle\rangle$ consisting of all formal power series over the alphabet X with coefficients in $\mathbb{R}$ forms a unital associative $\mathbb{R}$-algebra under the concatenation product and a unital, commutative and associative $\mathbb{R}$-algebra under the shuffle product, denoted here by the shuffle symbol $\sqcup\!\sqcup$. The latter is the $\mathbb{R}$-bilinear extension of the shuffle product of two words,

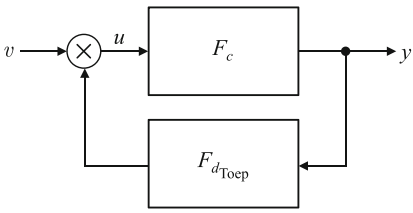


Fig. 2 Multiplicative feedback system

which is defined inductively by

$$(x_i\eta) \sqcup\!\sqcup (x_j\xi) = x_i(\eta \sqcup\!\sqcup (x_j\xi)) + x_j((x_i\eta) \sqcup\!\sqcup \xi)$$

with $\eta \sqcup\!\sqcup \emptyset = \emptyset \sqcup\!\sqcup \eta = \eta$ for all words $\eta, \xi \in X^*$ and letters $x_i, x_j \in X$ [18, 39]. For instance, $x_i \sqcup\!\sqcup x_j = x_ix_j + x_jx_i$ and

$$x_{i_1}x_{i_2} \sqcup\!\sqcup x_{i_3}x_{i_4} = x_{i_1}x_{i_2}x_{i_3}x_{i_4} + x_{i_3}x_{i_4}x_{i_1}x_{i_2} + x_{i_1}x_{i_3}(x_{i_2} \sqcup\!\sqcup x_{i_4}) + x_{i_3}x_{i_1}(x_{i_2} \sqcup\!\sqcup x_{i_4}).$$

Consider the (componentwise) multiplicative feedback interconnection shown in Fig. 2 consisting of a Fliess operator F_c in the forward path, where $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$, and an $m \times m$ matrix-valued Toeplitz Fliess operator $F_{d_{\text{Toep}}}$ in the feedback path. It is useful here to define a *generalized unital series*, δ_i, so that $F_{\delta_i}[y] := y_i$ for all $i = 1, \ldots, m$ and $F_{\delta_i^{\sqcup\!\sqcup j}}[y] := (F_{\delta_i}[y])^j = y_i^j$. With $d = (\delta_1, \delta_1^{\sqcup\!\sqcup 2}, \ldots, \delta_1^{\sqcup\!\sqcup m-1})$, the closed-loop system shown in Fig. 2 is described by

$$y = F_c[u], \quad u = F_{d_{\text{Toep}}}[y]v = v + \sum_{i=1}^{m-1} y_1^i N^i v,$$

and, in particular,

$$u_1 = v_1 + y_1v_2 + y_1^2v_3 + \cdots + y_1^{m-1}v_m. \tag{15}$$

Example 4 Suppose $c_{F,m} = [c_F, 0, \ldots, 0] \in \mathbb{R}^m\langle\langle X\rangle\rangle$, where $c_F = \sum_{k=0}^{\infty} k!\, x_1^k$ as in Example 1. In which case, $y_1 = F_{c_F}[u]$ is realized by the one dimensional state space model

$$\dot{z} = z^2u_1, \quad z(0) = 1, \quad y_1 = z.$$

Applying the feedback (15) gives the following realization for the closed-loop system

$$\dot{z} = \sum_{i=1}^{m} v_i z^{i+1}, \quad z(0) = 1, \quad y_1 = z, \quad i = 1, \ldots, m.$$

Hence, the generating series for the closed-loop system, denoted here by $c_{F,m} @ d_{\text{Toep}}$, has the property that

$$c_{A,m} = [c_{F,m} @ d_{\text{Toep}}]_1. \tag{16}$$

This is a generalization of the result given in [14] for the $m = 2$ case.

In control theory, feedback is often described algebraically in terms of transformation groups. This approach is useful here as it will lead to an explicit way to compute the generating series of any closed-loop system as shown in Fig. 2. Consider the group of Toeplitz affine Fliess operators

$$\mathscr{T} := \left\{ y = F_{d_\delta}[u] = F_{d_{\text{Toep}}}[u]u : d \in \mathbb{R}^{m-1}\langle\langle X \rangle\rangle \right\}$$

under the operator composition

$$(F_{c_\delta} \circ F_{d_\delta})[u] = F_{c_{\text{Toep}}}[F_{d_{\text{Toep}}}[u]u]F_{d_{\text{Toep}}}[u]u,$$

which is associative and has the identity element F_{0_δ}. Strictly speaking, one should limit the definition of the group to those generating series whose corresponding Fliess operators converge. But the algebraic set up presented here carries through in general if one considers the non-convergent case in a formal sense (see [25]). The group inverse has already been described for the case where d is differentially generated, i.e., by Eq. (10). It can be shown by other arguments to exist in general (via contractive maps on ultrametric spaces, see [28]). The group product on $\mathscr{T}$ in turn induces a formal power series product on $\mathbb{R}^{m-1}\langle\langle X \rangle\rangle$ denoted by $c_\delta \circ d_\delta$ satisfying $F_{c_\delta \circ d_\delta} = F_{c_\delta} \circ F_{d_\delta}$. Given that generating series are unique and the bijection between $\mathbb{R}^{m-1}\langle\langle X \rangle\rangle$ and their associated Toeplitz matrices, this means that $\mathbb{R}^{m-1}\langle\langle X \rangle\rangle$ inherits a group structure. A right action of the group $\mathscr{T}$ on the set of all Fliess operators F_c, $c \in \mathbb{R}^m\langle\langle X \rangle\rangle$ is given by

$$(F_c \circ F_{d_\delta})[u] = F_c[F_{d_{\text{Toep}}}[u]u] = F_c\left[u + \sum_{i=1}^{m-1} F_{d_i}[u]N^i u \right].$$

This composition induces a second formal power series product, the *mixed composition product* $c \,\tilde{\circ}\, d_\delta$, satisfying

$$F_c \circ F_{d_\delta} = F_{c \,\tilde{\circ}\, d_\delta}. \tag{17}$$

It can be viewed as a right action of the group $\mathbb{R}^{m-1}\langle\langle X \rangle\rangle$ on the set $\mathbb{R}^m\langle\langle X \rangle\rangle$. This product is left linear, nonassociative, and can be computed explicitly when $c \in \mathbb{R}\langle\langle X \rangle\rangle$ by

$$c \,\tilde{\circ}\, d_\delta = \phi_d(c)(\mathbf{1}) = \sum_{\eta \in X^*} \langle c, \eta \rangle \, \phi_d(\eta)(\mathbf{1}),$$

where $\mathbf{1} := 1\mathrm{e}$, and ϕ_d is the continuous (in the ultrametric sense) algebra homomorphism from $\mathbb{R}\langle\langle X\rangle\rangle$ to $\mathrm{End}(\mathbb{R}\langle\langle X\rangle\rangle)$ uniquely specified by $\phi_d(x_i\eta) = \phi_d(x_i) \circ \phi_d(\eta)$ with

$$\phi_d(x_i)(e) = x_i e + \sum_{j=1}^{m-i} x_{i+j}(d_j \sqcup\!\sqcup e), \quad i = 1, \ldots, m$$

for any $e \in \mathbb{R}\langle\langle X\rangle\rangle$, and where $\phi_d(\emptyset)$ denotes the identity map on $\mathbb{R}\langle\langle X\rangle\rangle$. For any $c \in \mathbb{R}^{i\times j}\langle\langle X\rangle\rangle$ the product is extended componentwise such that

$$[c \mathbin{\tilde{\circ}} d]_{kl} = [c]_{kl} \mathbin{\tilde{\circ}} d \tag{18}$$

for all $k = 1, 2, \ldots, i$ and $l = 1, 2, \ldots, j$. The following pre-Lie product results from the right linearization of the mixed composition product

$$x_i\eta \mathbin{\triangleleft} d := x_i(\eta \mathbin{\triangleleft} d) + \sum_{j=1}^{m-i} x_{i+j}(d_j \sqcup\!\sqcup \eta)$$

with $\emptyset \triangleleft d := 0$. In which case, $c \mathbin{\tilde{\circ}} d_\delta = c + c \triangleleft d + O(d^2)$. In particular, it can be shown directly that

$$(c_\delta \circ d_\delta)_{\mathrm{Toep}} = (c_{\mathrm{Toep}} \mathbin{\tilde{\circ}} d_\delta) \sqcup\!\sqcup d_{\mathrm{Toep}},$$

where the shuffle product on matrix-valued series is defined componentwise. Another useful composition product is the (unmixed) composition product $c \circ d$ induced simply by $F_{c\circ d} = F_c \circ F_d$.

With these various formal power series products defined, it is now possible to give a general formula for the *feedback product* $c@d_{\mathrm{Toep}}$ describing the generating series for the interconnected system in Fig. 2. The following lemma is needed.

Lemma 1 *The set* $\mathscr{G}_{\sqcup\!\sqcup} := \{c \in \mathbb{R}^{m\times m}\langle\langle X\rangle\rangle : \langle c, \mathrm{e}\rangle \in Gl_m(\mathbb{R})\}$ *is a group under the shuffle product with the identity element being the constant series* $\mathbf{I} := I\mathrm{e}$, *and the inverse of any* $c \in \mathscr{G}_{\sqcup\!\sqcup}$ *is*

$$c^{\sqcup\!\sqcup -1} = (\langle c, \mathrm{e}\rangle(\mathbf{I} - c'))^{\sqcup\!\sqcup -1} = (c')^{\sqcup\!\sqcup *}\langle c, \mathrm{e}\rangle^{-1},$$

where c' *is proper (i.e,* $\langle c', \mathrm{e}\rangle = 0$*), and* $(c')^{\sqcup\!\sqcup *} := \sum_{k\geq 0}(c')^{\sqcup\!\sqcup k}$.

Theorem 2 *For any* $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$ *and* $d \in \mathbb{R}^{m-1}\langle\langle X\rangle\rangle$ *it follows that* $c@d_{\mathrm{Toep}} = c \mathbin{\tilde{\circ}} ((d_{\mathrm{Toep}} \circ c)^{\sqcup\!\sqcup -1})^{-1}_\delta$.

Proof The feedback law requires that $u = F_{d_{\text{Toep}}}[y]v = F_{d_{\text{Toep}}}[F_c[u]]v = F_{d_{\text{Toep}} \circ c}[u]v$. From Lemma 1 it follows that

$$v = F_{(d_{\text{Toep}} \circ c) \sqcup\!\sqcup -1}[u]u = F_{((d_{\text{Toep}} \circ c) \sqcup\!\sqcup -1)_\delta}[u].$$

As the latter is now a group element in $\mathscr{T}$, one can write

$$u = F_{((d_{\text{Toep}} \circ c) \sqcup\!\sqcup -1)_\delta^{-1}}[v].$$

Making this substitution for u into $y = F_c[u]$ and writing the result in terms of the group action gives

$$y = F_{c @ d_{\text{Toep}}}[v] = F_c[F_{((d_{\text{Toep}} \circ c) \sqcup\!\sqcup -1)_\delta^{-1}}[v]] = F_{c \,\tilde{\circ}\, ((d_{\text{Toep}} \circ c) \sqcup\!\sqcup -1)_\delta^{-1}}[v].$$

As generating series are known to be unique, the theorem is proved.

Corollary 1 *The feedback product satisfies the fixed point equation* $c @ d_{\text{Toep}} = c \,\tilde{\circ}\, (d_{\text{Toep}} \circ (c @ d_{\text{Toep}}))_\delta$.

Proof Observe that $y = F_c[F_{d_{\text{Toep}}}[y]v]$. So if $y = F_{c @ d_{\text{Toep}}}[v]$ then necessarily

$$y = F_c[F_{d_{\text{Toep}}}[F_{c @ d_{\text{Toep}}}[v]]v] = F_c[F_{d_{\text{Toep}} \circ (c @ d_{\text{Toep}})}[v]v] = F_{c \,\tilde{\circ}\, (d_{\text{Toep}} \circ (c @ d_{\text{Toep}}))_\delta}[v].$$

The uniqueness of generating series then proves the claim.

The tools above are now applied to compute the feedback product $c_{F,m} @ d_{\text{Toep}}$ in (16). This will in turn render identities satisfied by the Abel series. The following lemma is useful.

Lemma 2 *If in* (9) $h_i = h^i$, $i = 1, \ldots, m-1$ *for some* $h \in C^\omega$ *then* $H^{-1} = I - hN$.

Proof Given that $H = I + hN + h^2N^2 + \cdots + h^{m-1}N^{m-1}$, observe

$$\begin{aligned} H^{-1} &= ((I - hN)^{-1} - (hN)^m (I - hN)^{-1})^{-1} \\ &= (I - hN)(I - (hN)^m)^{-1} \\ &= I - hN + O((hN)^m) \\ &= I - hN, \end{aligned}$$

since $N^n = 0$ when $n \geq m$.

Theorem 3 *For any* $m \geq 2$, $c_{A,m} = c_F \,\tilde{\circ}\, (I - c_F N)_\delta^{-1}$.

Proof Starting from the formula in Theorem 2 for the feedback product with $c = c_{F,m} = [c_F, 0, \ldots, 0]$ and $d = (\delta_1, \delta_1^{\sqcup\!\sqcup 2}, \ldots, \delta_1^{\sqcup\!\sqcup m-1})$ and using the definition of

the shuffle inverse in Lemma 1, observe that

$$\begin{aligned} c_{F,m} @ d_{\text{Toep}} &= c_{F,m} \,\tilde{\circ}\, ((d_{\text{Toep}} \circ c_{F,m})^{\sqcup\!\sqcup -1})_{\delta}^{-1} \\ &= c_{F,m} \,\tilde{\circ}\, \left(\left(\left(I + \sum_{i=1}^{m-1} \delta_1^{\sqcup\!\sqcup i} N^i \right) \circ c_{F,m} \right)^{\sqcup\!\sqcup -1} \right)_{\delta}^{-1} \\ &= c_{F,m} \,\tilde{\circ}\, \left(\left(\sum_{i=0}^{m-1} c_F^{\sqcup\!\sqcup i} N^i \right)^{\sqcup\!\sqcup -1} \right)_{\delta}^{-1}. \end{aligned}$$

Now note that if h in Lemma 2 is identified with F_{c_F} then $h^i = F_{c_F}^i = F_{c_F^{\sqcup\!\sqcup i}}$. So the shuffle version of the identity in this lemma is $\left(\sum_{i=0}^{m-1} c_F^{\sqcup\!\sqcup i} N^i \right)^{\sqcup\!\sqcup -1} = I - c_F N$. In which case, $c_{F,m} @ d_{\text{Toep}} = c_{F,m} \,\tilde{\circ}\, (I - c_F N)_{\delta}^{-1}$. Next, in light of (16) and (18), it is clear that $c_{A,m} = [c_{F,m} \,\tilde{\circ}\, (I - c_F N)_{\delta}^{-1}]_1 = c_F \,\tilde{\circ}\, (I - c_F N)_{\delta}^{-1}$ as claimed.

Example 5 Consider evaluating $c_{A,m} = c_F \,\tilde{\circ}\, (I - c_F N)_{\delta}^{-1}$ when $m = 3$. In this case

$$c_F \,\tilde{\circ}\, (I - c_F N)_{\delta}^{-1} = \sum_{k=0}^{\infty} k!\, \phi_d(x_1^k)(\mathbf{1}),$$

where

$$\phi_d(x_1)(e) = x_1 e + x_2(d_1 \sqcup\!\sqcup e) + x_3(d_2 \sqcup\!\sqcup e)$$

with $d_1 = (I - c_F N)_1^{-1}$ and $d_2 = (I - c_F N)_2^{-1}$. Using (12) to compute the inverses gives

$$\begin{aligned} \langle d_1, \mathrm{e} \rangle &= S(a_{\mathrm{e}}^1)(-c_F) = -a_{\mathrm{e}}^1(-c_F) = -\langle -c_F, \mathrm{e} \rangle = 1 \\ \langle d_1, x_1 \rangle &= S(a_{x_1}^1)(-c_F) = -a_{x_1}^1(-c_F) = -\langle -c_F, x_1 \rangle = 1 \\ \langle d_1, x_2 \rangle &= S(a_{x_2}^1)(-c_F) = (-a_{x_2}^1 + a_{x_1}^1 a_{\mathrm{e}}^1)(-c_F) \\ &= -\langle -c_F, x_2 \rangle + \langle -c_F, x_1 \rangle \langle -c_F, \mathrm{e} \rangle = 1 \\ \langle d_1, x_3 \rangle &= S(a_{x_3}^1)(-c_F) = (-a_{x_3}^1 + a_{x_2}^1 a_{\mathrm{e}}^1 - a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 + a_{x_1}^1 a_{\mathrm{e}}^2)(-c_F) \\ &= -\langle -c_F, x_3 \rangle + \langle -c_F, x_2 \rangle \langle -c_F, \mathrm{e} \rangle - \langle -c_F, x_1 \rangle \langle -c_F, \mathrm{e} \rangle^2 + \langle -c_F, x_1 \rangle \langle 0, \mathrm{e} \rangle \\ &= 1. \end{aligned}$$

Therefore, $d_1 = 1 + x_1 + x_2 + x_3 + \cdots$, which from (14) should be $c_{A,3}$. Similarly, $d_2 = 1 + 2x_1 + 2x_2 + 2x_3$, so that

$$c_F \tilde{\circ} (I - c_F N)_\delta^{-1} = 1 + x_1 + x_2 + x_3 + 2x_1x_1 + 2x_1x_2 + 2x_1x_3 + 3x_2x_1 + 3x_2x_2 + 3x_2x_3 + 4x_3x_1 + 4x_3x_2 + 4x_3x_3 + \cdots,$$

which is also equivalent to $c_{A,3}$ as expected.

Theorem 4 *For any* $m \geq 2$

$$c_{A,m} = 1 + c_{A,m} \sqcup \left(\sum_{i=1}^{m} x_i c_{A,m}^{\sqcup i-1} \right).$$

Proof Applying Corollary 1, Theorem 3, and the fact that the mixed composition product distributes to the left over the shuffle product gives

$$\begin{aligned} c_{A,m} &= c_F \tilde{\circ} (d_{\text{Toep}} \circ c_{A,m})_\delta = c_F \tilde{\circ} \left(\sum_{i=0}^{m} c_A^{\sqcup i} N^i \right)_\delta \\ &= \sum_{k=0}^{\infty} x_1^{\sqcup k} \tilde{\circ} \left(\sum_{i=0}^{m} c_A^{\sqcup i} N^i \right)_\delta = \sum_{k=0}^{\infty} \left(x_1 \tilde{\circ} \left(\sum_{i=0}^{m} c_A^{\sqcup i} N^i \right)_\delta \right)^{\sqcup k} \\ &= \sum_{k=0}^{\infty} \left(\sum_{i=1}^{m} x_i c_{A,m}^{\sqcup i-1} \right)^{\sqcup k}. \end{aligned}$$

Hence, the identity in question then follows directly.

Theorem 4 was first observed in functional form for the $m = 2$ case in [34] (see equation (2.3)). In fact, one of the main results of this paper (Theorem 4.1) is actually just a graded version of this result as described next.

Corollary 2 *For any* $m, n \geq 2$

$$c_{A,m}(n) = c_{A,m}(n-1) \sqcup x_1 + \sum_{i=2}^{m} \sum_{k_1 + \cdots + k_i = n-1} c_{A,m}(k_1) \sqcup (x_i (c_{A,m}(k_2) \sqcup \cdots \sqcup c_{A,m}(k_i))).$$

Example 6 When $m = 3$ observe $c_{A,3} = 1 + c_{A,3} \sqcup (x_1 + x_2 c_{A,3} + x_3 c_{A,3}^{\sqcup 2})$. Therefore, if $a := F_{c_{A,3}}[u]$ then

$$a(t) = 1 + a(t) \left[\int_0^t u_1(\tau)\, d\tau + \int_0^t u_2(\tau) a(\tau)\, d\tau + \int_0^t u_3(\tau) a^2(\tau)\, d\tau \right].$$

Defining $a_n = F_{c_{A,3}(n)}[u]$, $n \geq 1$ gives the recursion

$$a_n(t) = a_{n-1}(t) \int_0^t u_1(\tau)\, d\tau + \sum_{k_1+k_2=n-1} a_{k_1}(t) \int_0^t u_2(\tau) a_{k_2}(\tau)\, d\tau + \sum_{k_1+k_2+k_3=n-1} a_{k_1}(t) \int_0^t u_3(\tau) a_{k_2}(\tau) a_{k_3}(\tau)\, d\tau.$$

The $m = 2$ case of this recursion appears in [34] as equation (1.7).

The final theorem will be generalized in Sect. 5 to provide a sufficient condition for a center of the Abel equation.

Theorem 5 *Let $v_1, v_2, \ldots, v_m \in L_1[0, \omega]$ and $m \geq 2$ be fixed. Then the $m+1$ degree Abel equation (4) with $z(0) = 1$ has the solution*

$$z(t) = \frac{1}{1 - E_{x_1}[u](t)},$$

if there exists functions $u_1, u_2, \ldots, u_m \in L_1[0, \omega]$ satisfying

$$\begin{aligned} v_1(t) &= u_1(t) - \frac{u_2(t)}{1 - E_{x_1}[u](t)} \\ v_2(t) &= u_2(t) - \frac{u_3(t)}{1 - E_{x_1}[u](t)} \\ &\vdots \\ v_{m-1}(t) &= u_{m-1}(t) - \frac{u_m(t)}{1 - E_{x_1}[u](t)} \\ v_m(t) &= u_m(t) \end{aligned}$$

with $E_{x_1}[u](t) := \int_0^t u_1(\tau)\, d\tau < 1$ on $[0, \omega]$.

Proof In light of Theorem 3, it is clear that $c_{A,m} = c_F \,\tilde{\circ}\, (I - c_F N)_\delta^{-1}$, and thus, $c_F = c_{A,m} \,\tilde{\circ}\, (I - c_F N)_\delta$. So assume there exists $u \in L_1^m[0, \omega]$ such that

$$v = F_{(I - c_F N)_\delta}[u] = \begin{bmatrix} u_1 - F_{c_F}[u] u_2 \\ u_2 - F_{c_F}[u] u_3 \\ \vdots \\ u_{m-1} - F_{c_F}[u] u_m \\ u_m \end{bmatrix}.$$

Then, observing that $F_{c_F}[u] = 1/(1 - E_{x_1}[u])$, it follows from (17) that

$$z(t) = F_{c_{A,m}}[v] = F_{c_{A,m}}[F_{(I-c_F N)_\delta}[u]] = F_{c_{A,m} \tilde{\circ} (I-c_F N)_\delta}[u] = F_{c_F}[u] = \frac{1}{1 - E_{x_1}[u](t)}.$$

In the next section a Hopf algebra structure is defined on the coordinate functions.

4 Multivariable Hopf Algebra for Toeplitz Multiplicative Output Feedback

All algebraic structures considered in this section are over the field $\mathbb{K}$ of characteristic zero. Let $X = \{x_1, \ldots, x_m\}$ be a finite alphabet with m letters. Each letter has an integer degree $\deg(x_k) := k$. The monoid of words is denoted by X^* and includes the empty word $\mathrm{e} = \emptyset$ for which $\deg(\mathrm{e}) = 0$. The degree of a word $\eta = x_{i_1} \cdots x_{i_n} \in X^*$ of length $|\eta| := n$ is defined by

$$\deg(\eta) := \sum_{k=1}^{m} k|\eta|_k.$$

Here $|\eta|_k$ denotes the number of times the letter $x_k \in X$ appears in the word η.

Consider the polynomial algebra $H^{(\bar{m})}$ generated by the *coordinate functions* a_η^k, where $\eta \in X^*$ and the so called *root index* $k \in [\bar{m}] := \{1, \ldots, \bar{m}\}$, $\bar{m} \leq m$. By defining the degree

$$\|a_\eta^k\| := k + \deg(\eta),$$

$H^{(\bar{m})}$ becomes a graded connected algebra, $H^{(\bar{m})} := \bigoplus_{n \geq 0} H_n^{(\bar{m})}$, and $\|a_\eta^k a_\kappa^l\| = \|a_\eta^k\| + \|a_\kappa^l\|$. The unit in $H^{(\bar{m})}$ is denoted by $\mathbf{1}$, and $\|\mathbf{1}\| = 0$, whereas $\|a_\mathrm{e}^k\| = k$.

The *left-* and *right-shift* maps, $\theta_{x_j} : H^{(\bar{m})} \to H^{(\bar{m})}$ respectively $\tilde{\theta}_{x_j} : H^{(\bar{m})} \to H^{(\bar{m})}$, for $x_j \in X$, are taken to be

$$\theta_{x_j} a_\eta^p := a_{x_j \eta}^p, \qquad \tilde{\theta}_{x_j} a_\eta^p := a_{\eta x_j}^p$$

with $\theta_{x_j} \mathbf{1} = \tilde{\theta}_{x_j} \mathbf{1} = 0$. On products in $H^{(\bar{m})}$ both these maps act as derivations

$$\theta_{x_j} a_\eta^p a_\mu^q := (\theta_{x_j} a_\eta^p) a_\mu^q + a_\eta^p (\theta_{x_j} a_\mu^q),$$

and analogously for $\tilde{\theta}_{x_j}$. For a word $\eta = x_{i_1} \cdots x_{i_n} \in X^*$

$$\theta_\eta := \theta_{x_{i_1}} \cdots \theta_{x_{i_n}}, \qquad \tilde{\theta}_\eta := \tilde{\theta}_{x_{i_n}} \cdots \tilde{\theta}_{x_{i_1}}.$$

Hence, any element $a_\eta^i \in H^{(\bar{m})}$ with $\eta = x_{i_1} \cdots x_{i_n} \in X^*$ can be written

$$a_\eta^i = \theta_\eta a_{\mathrm{e}}^i = \tilde{\theta}_\eta a_{\mathrm{e}}^i.$$

In the following it will be shown how $\tilde{\theta}_\eta$ can be employed to define a coproduct $\Delta : H^{(\bar{m})} \to H^{(\bar{m})} \otimes H^{(\bar{m})}$. First, for the coordinate functions with respect to the empty word, a_{e}^l, $1 \le l \le \bar{m}$, the coproduct is defined to be

$$\Delta a_{\mathrm{e}}^l := a_{\mathrm{e}}^l \otimes \mathbf{1} + \mathbf{1} \otimes a_{\mathrm{e}}^l + \sum_{k=1}^{l-1} a_{\mathrm{e}}^k \otimes a_{\mathrm{e}}^{l-k}. \tag{19}$$

Note that a_{e}^1 is by definition primitive, i.e., $\Delta a_{\mathrm{e}}^1 = a_{\mathrm{e}}^1 \otimes \mathbf{1} + \mathbf{1} \otimes a_{\mathrm{e}}^1$. The next step is to define Δ on any a_η^i with $1 \le i \le \bar{m}$ and $|\eta| > 0$ by specifying intertwining relations between the maps $\tilde{\theta}_{x_i}$ and the coproduct

$$\Delta \circ \tilde{\theta}_{x_i} := \left(\tilde{\theta}_{x_i} \otimes \mathrm{id} + \mathrm{id} \otimes \tilde{\theta}_{x_i} + \sum_{j=1}^{i-1} \tilde{\theta}_{x_j} \otimes A_{\mathrm{e}}^{(i-j)} \right) \circ \Delta. \tag{20}$$

The map $A_{\mathrm{e}}^{(k)}$ is defined by

$$A_{\mathrm{e}}^{(k)} a_\eta^i := a_\eta^i a_{\mathrm{e}}^k.$$

The following notation is used, $\Delta \circ \tilde{\theta}_{x_i} = \tilde{\Theta}_{x_i} \circ \Delta$, where

$$\tilde{\Theta}_{x_i} := \tilde{\theta}_{x_i} \otimes \mathrm{id} + \mathrm{id} \otimes \tilde{\theta}_{x_i} + \sum_{j=1}^{i-1} \tilde{\theta}_{x_j} \otimes A_{\mathrm{e}}^{(i-j)},$$

and $\tilde{\Theta}_\eta := \tilde{\Theta}_{x_{i_n}} \cdots \tilde{\Theta}_{x_{i_1}}$ for $\eta = x_{i_1} \cdots x_{i_n} \in X^*$. In this setting, $a_{x_1}^1$ is primitive since

$$\Delta a_{x_1}^1 = \Delta \circ \tilde{\theta}_{x_1} a_{\mathrm{e}}^1 = \tilde{\Theta}_{x_1} \circ \Delta a_{\mathrm{e}}^1 = (\tilde{\theta}_{x_1} \otimes \mathrm{id} + \mathrm{id} \otimes \tilde{\theta}_{x_1})(a_{\mathrm{e}}^1 \otimes \mathbf{1} + \mathbf{1} \otimes a_{\mathrm{e}}^1) = a_{x_1}^1 \otimes \mathbf{1} + \mathbf{1} \otimes a_{x_1}^1,$$

which follows from $\tilde{\theta}_{x_j} \mathbf{1} = 0$. The coproduct of $a_{x_2}^l$ is

$$\Delta a_{x_2}^l = \Delta \circ \tilde{\theta}_{x_2} a_{\mathrm{e}}^l = \tilde{\Theta}_{x_2} \circ \Delta a_{\mathrm{e}}^l = \left(\tilde{\theta}_{x_2} \otimes \mathrm{id} + \mathrm{id} \otimes \tilde{\theta}_{x_2} + \tilde{\theta}_{x_1} \otimes A_{\mathrm{e}}^{(1)} \right) \circ \Delta a_{\mathrm{e}}^l$$

$$= a_{x_2}^l \otimes \mathbf{1} + \mathbf{1} \otimes a_{x_2}^l + a_{x_1}^l \otimes a_{\mathrm{e}}^1 + \sum_{k=1}^{l-1} a_{x_2}^k \otimes a_{\mathrm{e}}^{l-k} + \sum_{k=1}^{l-1} a_{\mathrm{e}}^k \otimes a_{x_2}^{l-k} + \sum_{k=1}^{l-1} a_{x_1}^k \otimes a_{\mathrm{e}}^1 a_{\mathrm{e}}^{l-k}.$$

The coproduct of a general $a^l_{x_i}$ is

$$\begin{aligned}\Delta a^l_{x_i} &= \Delta \circ \tilde{\theta}_{x_i} a^l_{\mathrm{e}} = \tilde{\Theta}_{x_i} \circ \Delta a^l_{\mathrm{e}} = \big(\tilde{\theta}_{x_i} \otimes \mathrm{id} + \mathrm{id} \otimes \tilde{\theta}_{x_i} + \sum_{j=1}^{i-1} \tilde{\theta}_{x_j} \otimes A_{\mathrm{e}}^{(i-j)}\big) \circ \Delta a^l_{\mathrm{e}} \\ &= a^l_{x_i} \otimes \mathbf{1} + \mathbf{1} \otimes a^l_{x_i} + \sum_{j=1}^{i-1} a^l_{x_j} \otimes a_{\mathrm{e}}^{i-j} + \sum_{k=1}^{l-1} a^k_{x_i} \otimes a_{\mathrm{e}}^{l-k} + \sum_{k=1}^{l-1} a^k_{\mathrm{e}} \otimes a^{l-k}_{x_i} \\ &\quad + \sum_{j=1}^{i-1} \sum_{k=1}^{l-1} a^k_{x_j} \otimes a_{\mathrm{e}}^{i-j} a_{\mathrm{e}}^{l-k}. \end{aligned} \tag{21}$$

Observe that the grading is preserved. A few examples may be helpful

$$\begin{aligned}\Delta' a^2_{x_1} &= a^1_{x_1} \otimes a^1_{\mathrm{e}} + a^1_{\mathrm{e}} \otimes a^1_{x_1} \\ \Delta' a^1_{x_2} &= a^1_{x_1} \otimes a^1_{\mathrm{e}} \\ \Delta' a^2_{x_2} &= a^2_{x_1} \otimes a^1_{\mathrm{e}} + a^1_{x_2} \otimes a^1_{\mathrm{e}} + a^1_{\mathrm{e}} \otimes a^1_{x_2} + a^1_{x_1} \otimes a^1_{\mathrm{e}} a^1_{\mathrm{e}} \\ \Delta' a^1_{x_3} &= a^1_{x_1} \otimes a^2_{\mathrm{e}} + a^1_{x_2} \otimes a^1_{\mathrm{e}} \\ \Delta' a^2_{x_3} &= a^2_{x_1} \otimes a^2_{\mathrm{e}} + a^2_{x_2} \otimes a^1_{\mathrm{e}} + a^1_{x_3} \otimes a^1_{\mathrm{e}} + a^1_{\mathrm{e}} \otimes a^1_{x_3} + a^1_{x_1} \otimes a^2_{\mathrm{e}} a^1_{\mathrm{e}} + a^1_{x_2} \otimes a^1_{\mathrm{e}} a^1_{\mathrm{e}},\end{aligned}$$

where $\Delta' a^l_\eta := \Delta a^l_\eta - a^l_\eta \otimes \mathbf{1} - \mathbf{1} \otimes a^l_\eta$ is the reduced coproduct. For the element $a^l_{x_2 x_1}$ one finds the following coproduct

$$\begin{aligned}\Delta a^l_{x_2x_1} &= \Delta \circ \tilde{\theta}_{x_1} \tilde{\theta}_{x_2} a^l_{\mathrm{e}} = \tilde{\Theta}_{x_1} \tilde{\Theta}_{x_2} \circ \Delta a^l_{\mathrm{e}} \\ &= a^l_{x_2x_1} \otimes \mathbf{1} + \mathbf{1} \otimes a^l_{x_2x_1} + a^l_{x_1} \otimes a^1_{x_1} + a^l_{x_1x_1} \otimes a^1_{\mathrm{e}} + \sum_{k=1}^{l-1} a^k_{x_2x_1} \otimes a_{\mathrm{e}}^{l-k} \\ &\quad + \sum_{k=1}^{l-1} a^k_{x_1} \otimes a^{l-k}_{x_2} + \sum_{k=1}^{l-1} a^k_{x_1x_1} \otimes a^1_{\mathrm{e}} a_{\mathrm{e}}^{l-k} + \sum_{k=1}^{l-1} a^k_{x_2} \otimes a^{l-k}_{x_1} + \sum_{k=1}^{l-1} a^k_{\mathrm{e}} \otimes a^{l-k}_{x_2x_1} \\ &\quad + \sum_{k=1}^{l-1} a^k_{x_1} \otimes a^1_{x_1} a_{\mathrm{e}}^{l-k} + \sum_{k=1}^{l-1} a^k_{x_1} \otimes a^1_{\mathrm{e}} a^{l-k}_{x_1}.\end{aligned}$$

The general formula for words of length two is

$$\begin{aligned}
&\Delta a^l_{x_jx_i} = \Delta\circ\tilde\theta_{x_i}\tilde\theta_{x_j}a^l_{\rm e} = \tilde\Theta_{x_i}\tilde\Theta_{x_j}\circ\Delta a^l_{\rm e}\\
&= a^l_{x_jx_i}\otimes\mathbf{1}+\mathbf{1}\otimes a^l_{x_jx_i}+\sum_{n=1}^{j-1}a^l_{x_nx_i}\otimes a^{j-n}_{\rm e}+\sum_{n=1}^{j-1}a^l_{x_n}\otimes a^{j-n}_{x_i}+\sum_{s=1}^{i-1}a^l_{x_jx_s}\otimes a^{i-s}_{\rm e}\\
&+\sum_{s=1}^{i-1}\sum_{n=1}^{j-1}a^l_{x_nx_s}\otimes a^{i-s}_{\rm e}a^{j-n}_{\rm e}+\sum_{k=1}^{l-1}a^k_{x_jx_i}\otimes a^{l-k}_{\rm e}+\sum_{k=1}^{l-1}a^k_{x_i}\otimes a^{l-k}_{x_j}+\sum_{n=1}^{j-1}\sum_{k=1}^{l-1}a^k_{x_nx_i}\otimes a^{j-n}_{\rm e}a^{l-k}_{\rm e}\\
&+\sum_{k=1}^{l-1}a^k_{x_j}\otimes a^{l-k}_{x_i}+\sum_{k=1}^{l-1}a^k_{\rm e}\otimes a^{l-k}_{x_jx_i}+\sum_{n=1}^{j-1}\sum_{k=1}^{l-1}a^k_{x_n}\otimes a^{j-n}_{x_i}a^{l-k}_{\rm e}+\sum_{n=1}^{j-1}\sum_{k=1}^{l-1}a^k_{x_n}\otimes a^{j-n}_{\rm e}a^{l-k}_{x_i}\\
&+\sum_{s=1}^{i-1}\sum_{k=1}^{l-1}a^k_{x_jx_s}\otimes a^{i-s}_{\rm e}a^{l-k}_{\rm e}+\sum_{s=1}^{i-1}\sum_{k=1}^{l-1}a^k_{x_s}\otimes a^{i-s}_{\rm e}a^{l-k}_{x_j}+\sum_{s=1}^{i-1}\sum_{n=1}^{j-1}\sum_{k=1}^{l-1}a^k_{x_nx_s}\otimes a^{i-s}_{\rm e}a^{j-n}_{\rm e}a^{l-k}_{\rm e}.
\end{aligned}$$

The coproduct is then extended multiplicatively to all of $H^{(\tilde m)}$ and $\Delta(\mathbf{1}) := \mathbf{1}\otimes\mathbf{1}$.

Theorem 6 *$H^{(\tilde m)}$ is a connected graded commutative non-cocommutative Hopf algebra with unit map $u:\mathbb{K}\to H^{(\tilde m)}$, counit $\epsilon: H^{(\tilde m)}\to\mathbb{K}$ and coproduct $\Delta: H^{(\tilde m)}\to H^{(\tilde m)}\otimes H^{(\tilde m)}$*

$$\Delta a^k_\eta = \tilde\Theta_\eta\circ\Delta a^k_{\rm e}. \tag{22}$$

Proof $H^{(\tilde m)} = \bigoplus_{n\geq 0}H^{(\tilde m)}_n$ is connected graded and commutative by construction. In addition, it is clear that the coproduct is non-cocommutative. What is left to be shown is coassociativity. This is done by first proving the claim for $a^l_{\rm e}$, which follows from the identity

$$\sum_{k=1}^{l-1}\sum_{p=1}^{k-1}a^p_{\rm e}\otimes a^{k-p}_{\rm e}\otimes a^{l-k}_{\rm e} = \sum_{k=1}^{l-1}\sum_{p=1}^{l-k-1}a^k_{\rm e}\otimes a^p_{\rm e}\otimes a^{l-k-p}_{\rm e}.$$

From $\Delta(a^k_{\eta x_i}) = \Delta\circ\tilde\theta_{x_i}(a^k_\eta) = \tilde\Theta_{x_i}\circ\Delta(a^k_\eta)$ it follows that

$$\begin{aligned}
&(\Delta\otimes\mathrm{id})\circ\Delta(a^k_{\eta x_i}) = (\Delta\otimes\mathrm{id})\circ\tilde\Theta_{x_i}\circ\Delta(a^k_\eta)\\
&= \Big(\Delta\circ\tilde\theta_{x_i}\otimes\mathrm{id}+\mathrm{id}\otimes\mathrm{id}\otimes\tilde\theta_{x_i}+\sum_{j=1}^{i-1}\Delta\circ\tilde\theta_{x_j}\otimes A^{(i-j)}_{\rm e}\Big)\circ\Delta(a^k_\eta)\\
&= \Big(\tilde\Theta_{x_i}\otimes\mathrm{id}+\mathrm{id}\otimes\mathrm{id}\otimes\tilde\theta_{x_i}+\sum_{j=1}^{i-1}\tilde\Theta_{x_j}\otimes A^{(i-j)}_{\rm e}\Big)(\Delta\otimes\mathrm{id})\circ\Delta(a^k_\eta)\\
&= \Big(\tilde\theta_{x_i}\otimes\mathrm{id}\otimes\mathrm{id}+\mathrm{id}\otimes\tilde\theta_{x_i}\otimes\mathrm{id}+\mathrm{id}\otimes\mathrm{id}\otimes\tilde\theta_{x_i}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes A_{\mathrm{e}}^{(i-j)}\otimes\mathrm{id}+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes\mathrm{id}\otimes A_{\mathrm{e}}^{(i-j)}\\
&+\sum_{j=1}^{i-1}\mathrm{id}\otimes\tilde{\theta}_{x_j}\otimes A_{\mathrm{e}}^{(i-j)}+\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\tilde{\theta}_{x_k}\otimes A_{\mathrm{e}}^{(j-k)}\otimes A_{\mathrm{e}}^{(i-j)}\Big)(\mathrm{id}\otimes\Delta)\circ\Delta(a_\eta^k)\\
&=\Big(\tilde{\theta}_{x_i}\otimes\mathrm{id}\otimes\mathrm{id}+\mathrm{id}\otimes\tilde{\Theta}_{x_i}+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes\Big(A_{\mathrm{e}}^{(i-j)}\otimes\mathrm{id}+\mathrm{id}\otimes A_{\mathrm{e}}^{(i-j)}\Big)\\
&+\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\tilde{\theta}_{x_k}\otimes A_{\mathrm{e}}^{(j-k)}\otimes A_{\mathrm{e}}^{(i-j)}\Big)(\mathrm{id}\otimes\Delta)\circ\Delta(a_\eta^k).
\end{aligned}
$$

As noted above, the last sum can be rewritten as

$$
\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\tilde{\theta}_{x_k}\otimes A_{\mathrm{e}}^{(j-k)}\otimes A_{\mathrm{e}}^{(i-j)}=\sum_{j=1}^{i-2}\sum_{k=1}^{i-j-1}\tilde{\theta}_{x_j}\otimes A_{\mathrm{e}}^{(k)}\otimes A_{\mathrm{e}}^{(i-j-k)}
$$

so that

$$
\begin{aligned}
&\Big(\tilde{\theta}_{x_i}\otimes\mathrm{id}\otimes\mathrm{id}+\mathrm{id}\otimes\tilde{\Theta}_{x_i}+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes\big(A_{\mathrm{e}}^{(i-j)}\otimes\mathrm{id}+\mathrm{id}\otimes A_{\mathrm{e}}^{(i-j)}\big)\\
&+\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\tilde{\theta}_{x_k}\otimes A_{\mathrm{e}}^{(j-k)}\otimes A_{\mathrm{e}}^{(i-j)}\Big)(\mathrm{id}\otimes\Delta)\circ\Delta(a_\eta^k)\\
&=\Big(\tilde{\theta}_{x_i}\otimes\mathrm{id}\otimes\mathrm{id}+\mathrm{id}\otimes\tilde{\Theta}_{x_i}\\
&+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes\Big(A_{\mathrm{e}}^{(i-j)}\otimes\mathrm{id}+\mathrm{id}\otimes A_{\mathrm{e}}^{(i-j)}+\sum_{k=1}^{i-j-1}A_{\mathrm{e}}^{(k)}\otimes A_{\mathrm{e}}^{(i-j-k)}\Big)\Big)(\mathrm{id}\otimes\Delta)\circ\Delta(a_\eta^k)\\
&=(\mathrm{id}\otimes\Delta)\circ\Big(\tilde{\theta}_{x_i}\otimes\mathrm{id}+\mathrm{id}\otimes\tilde{\theta}_{x_i}+\sum_{j=1}^{i-1}\tilde{\theta}_{x_j}\otimes A_{\mathrm{e}}^{(i-j)}\Big)\circ\Delta(a_\eta^k)\\
&=(\mathrm{id}\otimes\Delta)\circ\Delta(a_{\eta x_i}^k).
\end{aligned}
$$

The following was also used in the calculation above

$$
\Delta\circ A_{\mathrm{e}}^{(i-j)}=\Big(A_{\mathrm{e}}^{(i-j)}\otimes\mathrm{id}+\mathrm{id}\otimes A_{\mathrm{e}}^{(i-j)}+\sum_{k=1}^{i-j-1}A_{\mathrm{e}}^{(k)}\otimes A_{\mathrm{e}}^{(i-j-k)}\Big)\circ\Delta,
$$

which follows from $A_{\mathrm{e}}^{(l)}a_\eta^k=a_{\mathrm{e}}^l a_\eta^k$ together with the multiplicativity of Δ.

In the following, a variant of Sweedler's notation [40] is used for the reduced coproduct, i.e., $\Delta'(a^l_\eta) = \sum a^{l'}_{\eta'} \otimes a^{l''}_{\eta''}$, as well as for the full coproduct

$$\Delta(a^l_\eta) = \sum a^{l'}_{\eta_{(1)}} \otimes a^{l''}_{\eta_{(2)}} = a^l_\eta \otimes \mathbf{1} + \mathbf{1} \otimes a^l_\eta + \Delta'(a^l_\eta).$$

Connectedness of $H^{(\tilde{m})}$ implies for the antipode $S : H^{(\tilde{m})} \to H^{(\tilde{m})}$ the well known recursions

$$Sa^l_\eta = -a^l_\eta - \sum S(a^{l'}_{\eta'}) a^{l''}_{\eta''} = -a^l_\eta - \sum a^{l'}_{\eta'} S(a^{l''}_{\eta''}). \tag{23}$$

A few examples are given first. Coproduct (19) implies for the elements a^k_{e} that

$$Sa^k_{\mathrm{e}} = -a^k_{\mathrm{e}} + \sum_{i=2}^{k} (-1)^i \sum_{\substack{p_1+\cdots+p_i=k \\ p_j>0}} a^{p_1}_{\mathrm{e}} \cdots a^{p_i}_{\mathrm{e}}. \tag{24}$$

For example,

$$Sa^1_{\mathrm{e}} = -a^1_{\mathrm{e}}, \qquad Sa^2_{\mathrm{e}} = -a^2_{\mathrm{e}} + a^1_{\mathrm{e}} a^1_{\mathrm{e}}, \qquad Sa^3_{\mathrm{e}} = -a^3_{\mathrm{e}} + 2a^1_{\mathrm{e}} a^2_{\mathrm{e}} - a^1_{\mathrm{e}} a^1_{\mathrm{e}} a^1_{\mathrm{e}}.$$

The following examples are given for comparison with (12):

$$\begin{aligned}
Sa^1_{x_1} &= -a^1_{x_1} \\
Sa^1_{x_2} &= -a^1_{x_2} + a^1_{x_1} a^1_{\mathrm{e}} \\
Sa^1_{x_3} &= -a^1_{x_3} + a^1_{x_1} a^2_{\mathrm{e}} - a^1_{x_1} a^1_{\mathrm{e}} a^1_{\mathrm{e}} + a^1_{x_2} a^1_{\mathrm{e}} \\
Sa^2_{x_1} &= -a^2_{x_1} + 2a^1_{x_1} a^1_{\mathrm{e}} \\
Sa^2_{x_2} &= -a^2_{x_2} + a^2_{x_1} a^1_{\mathrm{e}} - 2a^1_{x_1} a^1_{\mathrm{e}} a^1_{\mathrm{e}} + 2a^1_{x_2} a^1_{\mathrm{e}} \\
Sa^2_{x_3} &= -a^2_{x_3} + 2a^1_{x_3} a^1_{\mathrm{e}} - 2a^1_{x_2} a^1_{\mathrm{e}} a^1_{\mathrm{e}} + a^2_{x_2} a^1_{\mathrm{e}} - a^2_{x_1} a^1_{\mathrm{e}} a^1_{\mathrm{e}} + a^2_{x_1} a^2_{\mathrm{e}} \\
&\quad - 2a^1_{x_1} a^1_{\mathrm{e}} a^2_{\mathrm{e}} + 2a^1_{x_1} a^1_{\mathrm{e}} a^1_{\mathrm{e}} a^1_{\mathrm{e}}
\end{aligned}$$

The next theorem uses the coproduct formula (20) to provide a simple formula for the antipode of $H^{(\tilde{m})}$.

Theorem 7 *For any nonempty word $\eta = x_{i_1} \cdots x_{i_l} \in X^*$, the antipode $S : H^{(\tilde{m})} \to H^{(\tilde{m})}$ can be written as*

$$Sa^k_\eta = \tilde{\Theta}'_\eta(Sa^k_{\mathrm{e}}), \tag{25}$$

where $\tilde{\Theta}'_\eta := \tilde{\theta}'_{x_{i_l}} \circ \cdots \circ \tilde{\theta}'_{x_{i_1}}$ *with*

$$\tilde{\theta}'_{x_l} := \tilde{\theta}_{x_l} + \sum_{j=1}^{l-1} S(a_{\mathrm{e}}^{l-j}) \tilde{\theta}_{x_j}.$$

For instance, calculating

$$\tilde{\Theta}'_{x_1}(Sa_{\mathrm{e}}^3) = \tilde{\theta}_{x_1}(-a_{\mathrm{e}}^3 + 2a_{\mathrm{e}}^1 a_{\mathrm{e}}^2 - a_{\mathrm{e}}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1) = -a_{x_1}^3 + 2a_{x_1}^1 a_{\mathrm{e}}^2 + 2a_{\mathrm{e}}^1 a_{x_1}^2 - 3a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1,$$

which coincides with $Sa_{x_1}^3$. Another example is

$$\tilde{\Theta}'_{x_2}(Sa_{\mathrm{e}}^2) = (\tilde{\theta}_{x_2} + S(a_{\mathrm{e}}^1)\tilde{\theta}_{x_1})(-a_{\mathrm{e}}^2 + a_{\mathrm{e}}^1 a_{\mathrm{e}}^1) = -a_{x_2}^2 + 2a_{x_2}^1 a_{\mathrm{e}}^1 + a_{x_1}^2 a_{\mathrm{e}}^1 - 2a_{x_1}^1 a_{\mathrm{e}}^1 a_{\mathrm{e}}^1.$$

Proof The proof follows by a nested induction using the weight of the root index and word length. First, formula (25) is shown to hold for words of length one. Note that the recursions (23) can be written in terms of the convolution product, i.e., $-S = P * S = S * P$, which is defined in terms of the coproduct (22)

$$S = -m_{H^{(\bar{m})}} \circ \big(P \otimes S\big) \circ \Delta = -m_{H^{(\bar{m})}} \circ \big(S \otimes P\big) \circ \Delta.$$

Here $m_{H^{(\bar{m})}}$ denotes the product in $H^{(\bar{m})}$ and $P := \mathrm{id} - u \circ \epsilon$ is the projector that maps the unit $\mathbf{1}$ in $H^{(\bar{m})}$ to zero and reduces to the identity on $H_+^{(\bar{m})} = \bigoplus_{n>0} H_n^{(\bar{m})}$. Formula (25) applied to $a_{x_l}^1$ gives

$$\tilde{\Theta}'_{x_l}(Sa_{\mathrm{e}}^1) = \Big(\tilde{\theta}_{x_l} + \sum_{j=1}^{l-1} S(a_{\mathrm{e}}^{l-j})\tilde{\theta}_{x_j}\Big) Sa_{\mathrm{e}}^1 = -a_{x_l}^1 - \sum_{j=1}^{l-1} S(a_{\mathrm{e}}^{l-j}) a_{x_j}^1,$$

where (24) was used. This coincides with

$$\begin{aligned} Sa_{x_l}^1 &= -m_{H^{(\bar{m})}} \circ \big(P \otimes S\big) \circ \Delta a_{x_l}^1 \\ &= -m_{H^{(\bar{m})}} \circ \big(P \otimes S\big)\Big(a_{x_l}^1 \otimes \mathbf{1} + \mathbf{1} \otimes a_{x_l}^1 + \sum_{j=1}^{l-1} a_{x_j}^1 \otimes a_{\mathrm{e}}^{l-j}\Big) \\ &= -a_{x_l}^1 - \sum_{j=1}^{l-1} S(a_{\mathrm{e}}^{l-j}) a_{x_j}^1. \end{aligned}$$

Now (25) applied to $a^k_{x_l}$ gives

$$\begin{aligned}\tilde{\Theta}'_{x_l}(Sa^k_{\mathrm{e}}) &= \Big(\tilde{\theta}_{x_l} + \sum_{j=1}^{l-1} S(a^{l-j}_{\mathrm{e}})\tilde{\theta}_{x_j}\Big)Sa^k_{\mathrm{e}}\\ &= \Big(\tilde{\theta}_{x_l} + \sum_{j=1}^{l-1} S(a^{l-j}_{\mathrm{e}})\tilde{\theta}_{x_j}\Big)\Big(-a^k_{\mathrm{e}} - \sum_{w=1}^{k-1} a^w_{\mathrm{e}} Sa^{k-w}_{\mathrm{e}}\Big)\\ &= -a^k_{x_l} - \sum_{j=1}^{l-1} S(a^{l-j}_{\mathrm{e}})a^k_{x_j} - \sum_{w=1}^{k-1} a^w_{x_l} Sa^{k-w}_{\mathrm{e}} - \sum_{j=1}^{l-1}\sum_{w=1}^{k-1} a^w_{x_j} S(a^{l-j}_{\mathrm{e}})S(a^{k-w}_{\mathrm{e}})\\ &\quad - \sum_{w=1}^{k-1} a^w_{\mathrm{e}}\tilde{\Theta}'_{x_l} Sa^{k-w}_{\mathrm{e}}.\end{aligned}$$

Using the induction hypothesis on the last term, namely, $\tilde{\Theta}'_{x_l} Sa^{k-w}_{\mathrm{e}} = Sa^{k-w}_{x_l}$, gives

$$\begin{aligned}\tilde{\Theta}'_{x_l}(Sa^k_{\mathrm{e}}) &= -a^k_{x_l} - \sum_{j=1}^{l-1} S(a^{l-j}_{\mathrm{e}})a^k_{x_j} - \sum_{w=1}^{k-1} a^w_{x_l} Sa^{k-w}_{\mathrm{e}} - \sum_{j=1}^{l-1}\sum_{w=1}^{k-1} a^w_{x_j} S(a^{l-j}_{\mathrm{e}})S(a^{k-w}_{\mathrm{e}})\\ &\quad - \sum_{w=1}^{k-1} a^w_{\mathrm{e}} Sa^{k-w}_{x_l}.\end{aligned}$$

This coincides with the antipode computed via the coproduct in (21) since

$$\begin{aligned}Sa^k_{x_l} &= -m_{H^{(\tilde{m})}}\circ\big(P\otimes S\big)\circ\Delta a^k_{x_l}\\ &= -m_{H^{(\tilde{m})}}\circ\big(P\otimes S\big)\Big(a^k_{x_l}\otimes\mathbf{1} + \mathbf{1}\otimes a^k_{x_l} + \sum_{j=1}^{l-1} a^k_{x_j}\otimes a^{l-j}_{\mathrm{e}}\\ &\quad + \sum_{w=1}^{k-1} a^w_{x_l}\otimes a^{k-w}_{\mathrm{e}} + \sum_{j=1}^{l-1}\sum_{w=1}^{k-1} a^w_{x_j}\otimes a^{l-j}_{\mathrm{e}}a^{k-w}_{\mathrm{e}} + \sum_{w=1}^{k-1} a^w_{\mathrm{e}}\otimes a^{k-w}_{x_l}\Big).\end{aligned}$$

Now suppose (25) holds for all words $\nu\in X^*$ up to length $|\nu| = n-1$. The final step is to consider a^l_η, where $\eta = x_{i_1}\cdots x_{i_n} = \bar{\eta}x_{i_n}$, i.e., $|\eta| = n$, and $l\in[\bar{m}]$. Observe

$$\begin{aligned}\tilde{\Theta}'_\eta(a^l_{\mathrm{e}}) &= \tilde{\Theta}'_{x_{i_n}} S(a^l_{\bar{\eta}})\\ &= -\tilde{\Theta}'_{x_{i_n}} m_{H^{(\tilde{m})}}\circ\big(P\otimes S\big)\circ\Delta a^l_{\bar{\eta}}\\ &= -m_{H^{(\tilde{m})}}\circ\big((\tilde{\Theta}'_{x_{i_n}}\otimes\mathrm{id} + \mathrm{id}\otimes\tilde{\Theta}'_{x_{i_n}})\circ(P\otimes S)\big)\circ\Delta a^l_{\bar{\eta}}\end{aligned}$$

$$
\begin{aligned}
&= -m_{H^{(\tilde{m})}} \circ \big(P \circ \tilde{\Theta}'_{x_{i_n}} \otimes S + P \otimes \tilde{\Theta}'_{x_{i_n}} \circ S\big) \circ \Delta a^l_{\tilde{\eta}} \\
&= -m_{H^{(\tilde{m})}} \circ \Big(P \circ \tilde{\theta}_{x_{i_n}} \otimes S + P \circ \sum_{j=1}^{i_n-1} S(a_{\mathrm{e}}^{i_n-j}) \tilde{\theta}_{x_j} \otimes S + P \otimes S \circ \tilde{\theta}_{x_{i_n}}\Big) \circ \Delta a^l_{\tilde{\eta}} \\
&= -m_{H^{(\tilde{m})}} \circ \Big((P \otimes S) \circ \big(\tilde{\theta}_{x_{i_n}} \otimes \mathrm{id} + \sum_{j=1}^{i_n-1} \tilde{\theta}_{x_j} \otimes A_{\mathrm{e}}^{(i_n-j)} + \mathrm{id} \otimes \tilde{\theta}_{x_{i_n}}\big)\Big) \circ \Delta a^l_{\tilde{\eta}} \\
&= -m_{H^{(\tilde{m})}} \circ \big(P \otimes S\big) \circ \tilde{\Theta}_{x_{i_n}} \circ \Delta a^l_{\tilde{\eta}} \\
&= -m_{H^{(\tilde{m})}} \circ \big(P \otimes S\big) \circ \Delta a^l_{\eta} = S a^l_{\eta}.
\end{aligned}
$$

The third equality above came from the fact that $\tilde{\Theta}'_{x_{i_n}}$ is a sum of derivations. The fourth equality is a consequence of the identity $P \circ \tilde{\theta}_{x_{i_n}} = \tilde{\theta}_{x_{i_n}} \circ P$. The step from the fourth to the fifth equality used the induction hypothesis to get $P \otimes \tilde{\Theta}'_{x_{i_n}} \circ S = P \otimes S \circ \tilde{\theta}_{x_{i_n}}$, which holds due to the projector P being on the left-hand side. In addition, the following identity was used:

$$
m_{H^{(\tilde{m})}} \circ \Big(P \circ \sum_{j=1}^{i_n-1} S(a_{\mathrm{e}}^{i_n-j}) \tilde{\theta}_{x_j} \otimes S\Big) \circ \Delta = m_{H^{(\tilde{m})}} \circ \Big((P \otimes S) \circ \sum_{j=1}^{i_n-1} \tilde{\theta}_{x_j} \otimes A_{\mathrm{e}}^{(i_n-j)}\Big) \circ \Delta,
$$

which holds due to S being an algebra morphism.

The final result is evident from the fact that the feedback structures in Figs. 1 and 2 coincide when condition (15) holds with $m = 2$.

Corollary 3 *For the alphabet $X := \{x_1, x_2\}$ the Hopf algebra $H^{(1)}$ coincides with the Faà di Bruno-type Hopf algebra for single-input, single-output (SISO) output feedback given in* [20, 23, 26].

5 Sufficient Condition for a Center of the Abel Equation

Consider first a new sufficient condition for a center inspired by viewing the Abel equation in terms of a feedback connection as described in Sect. 3.

Theorem 8 *Let $v_1, v_2, \ldots, v_m \in L_1[0, \omega]$ and $m \geq 2$ be fixed. Then the $m+1$ degree Abel equation (4) has a center at $z = 0$ if there exists an $R > 0$ such that for*

every $r < R$ the system of equations

$$v_1(t) = u_1(t) - \frac{ru_2(t)}{1 - rE_{x_1}[u](t)} \tag{26a}$$

$$v_2(t) = u_2(t) - \frac{ru_3(t)}{1 - rE_{x_1}[u](t)} \tag{26b}$$

$$\vdots$$

$$v_{m-1}(t) = u_{m-1}(t) - \frac{ru_m(t)}{1 - rE_{x_1}[u](t)} \tag{26c}$$

$$v_m(t) = u_m(t), \tag{26d}$$

has a solution $u_1, u_2, \ldots, u_m \in L_1[0, \omega]$ with $E_{x_1}[u](t) := \int_0^t u_1(\tau)\, d\tau < 1/r$ on the interval $[0, \omega]$ and $E_{x_1}[u](\omega) = 0$.

Proof The claim is proved by showing that if the system (26) has the solution u_1, $u_2, \ldots, u_m$ then the Abel equation (4) with $z(0) = r < R$ has the solution

$$z(t) = \frac{r}{1 - rE_{x_1}[u](t)}.$$

In which case, $z(0) = z(\omega) = r$ for all $r < R$ so that $z = 0$ is a center.

Consider the case where $m = 2$ for simplicity. The proposed solution for (4) can be checked by direct substitution. That is,

$$\dot{z}(t) = \frac{r^2}{\left(1 - rE_{x_1}[u](t)\right)^2} u_1(t),$$

so that

$$\begin{aligned} v_1(t)z^2(t) + v_2(t)z^3(t) &= \left[u_1(t) - \frac{ru_2(t)}{1 - rE_{x_1}[u](t)}\right]\left[\frac{r}{1 - rE_{x_1}[u](t)}\right]^2 + \\ &\quad u_2(t)\left[\frac{r}{1 - rE_{x_1}[u](t)}\right]^3 \\ &= \frac{r^2}{(1 - rE_{x_1}[u](t))^2} u_1(t) \end{aligned}$$

as expected.

Recall it was shown in Theorem 5 where $z(0) = 1$ that $z(t) = F_{c_{A,m}}[v](t) = 1/(1 - E_{x_1}[u](t))$. So for sufficiently small $R > 0$ and given any $r < R$ the solution

to Eq. (4) with $z(0) = r$ can be written in the form

$$z(t) = r \sum_{n=1}^{\infty} F_{c_{A,m}(n)}[v](t) r^n = r \sum_{n=1}^{\infty} F_{r^n c_{A,m}(n)}[v](t) =: r \sum_{n=1}^{\infty} F_{c'_{A,m}(n)}[v](t).$$

So letting $c'_{A,m} := \sum_{n=1}^{\infty} c'_{A,m}(n)$, the composition condition (5) ensures periodic solutions because

$$\begin{aligned} z(\omega) &= r F_{c'_{A,m}}[v](\omega) = r \sum_{\eta \in X^*} \langle c'_{A,m}, \eta \rangle E_\eta[v](\omega) \\ &= r \sum_{\eta \in X^*} \langle c'_{A,m}, \eta \rangle E_\eta[\bar{v}](q(\omega)) = r \sum_{\eta \in X^*} \langle c'_{A,m}, \eta \rangle E_\eta[\bar{v}](q(0)) \\ &= r E_\emptyset[\bar{v}](q(0)) = r = z(0), \end{aligned}$$

using the fact that $E_\eta[\bar{v}](q(0)) = 0$ for all $\eta \neq \emptyset$. Put another way, the composition condition gives periodic solutions by simply ensuring that $E_\eta[v](\omega) = 0$ for every nonempty word $\eta \in X^*$. In which case, it is immediate from the shuffle identity $x_i^{\sqcup\!\sqcup k} = k! x_i^k$ that the *moment conditions* with respect to v

$$\int_0^\omega v_i(\tau) E^k_{x_1}[v](\tau)\, d\tau = k! E_{x_i x_1^k}[v](\omega) = 0, \quad i = 2, 3, \ldots, m, \quad k \geq 0$$

are satisfied. It is known for polynomial v_i, however, that the moment conditions do not imply the composition condition [22]. The following theorem indicates a condition under which the two conditions are satisfied with respect to the u_i functions.

Theorem 9 *Suppose the $v_1, v_2, \ldots, v_m \in L_1[0, \omega]$ satisfy the composition condition. Let $u_1, u_2, \ldots, u_m \in L_1[0, \omega]$ be any solution to (26) with $E_{x_1}[u](t) := \int_0^t u_1(\tau)\, d\tau < 1/r$ on the interval $[0, \omega]$. Then the composition condition and the moment conditions with respect to the u_i are equivalent.*

Proof Integrating both sides of (26) over $[0, \omega]$ gives

$$\begin{aligned} E_{x_i}[v](\omega) &= E_{x_i}[u](\omega) - r \sum_{k=0}^{\infty} r^k \int_0^\omega u_{i+1}(t) E^k_{x_1}[u](\tau)\, d\tau \\ &= E_{x_i}[u](\omega) - r \sum_{k=0}^{\infty} r^k k! E_{x_{i+1} x_1^k}[u](\omega) \end{aligned}$$

for $i = 1, 2, \ldots, m-1$ with $E_{x_m}[v](\omega) = E_{x_m}[u](\omega)$. Therefore, if the v_i satisfy the composition condition then the left-hand side of this equation is zero. In which case, the claim follows immediately.

References

1. Alwash, M.A.M.: On a condition for a center of cubic non-autonomous equations. Proc. R. Soc. Edinb. **113**, 289–291 (1989)
2. Alwash, M.A.M.: The composition conjecture for Abel equation. Expo. Math. **27**, 241–250 (2009)
3. Alwash, M.A.M., Lloyd, N.G.: Nonautonomous equations related to polynomial two-dimensional systems. Proc. R. Soc. Edinb. **105A**, 129–152 (1987)
4. Berlin, L., Gray, W.S., Duffaut Espinosa, L.A., Ebrahimi-Fard, K.: On the performance of antipode algorithms for the multivariable output feedback Hopf algebra. In: Proceedings of the 51st Conference on Information Sciences and Systems, Baltimore (2017)
5. Briskin, M., Roytvarf, N., Yomdin, Y.: Center conditions at infinity for Abel differential equation. Ann. Math. **172**, 437–483 (2010)
6. Briskin, M., Yomdin, Y.: Tangential version of Hilbert 16th problem for the Abel equation. Moscow Math. J. **5**, 23–53 (2005)
7. Brudnyi, A.: Some algebraic aspects of the center problem for ordinary differential equations. Qual. Theory Dyn. Syst. **9**, 9–28 (2010)
8. Brudnyi, A.: Shuffle and Faà di Bruno Hopf algebras in the center problem for ordinary differential equations. Bull. Sci. Math. **140**, 830–863 (2016)
9. Cherkas, L.: Number of limit cycles of an autonomous second-order system. Differ. Uravn. **12**, 944–946 (1976)
10. Devlin, J.: Word problems related to periodic solutions of a nonautonomous system. Math. Proc. Camb. Philos. Soc. **108**, 127–151 (1990)
11. Devlin, J.: Word problems related to derivatives of the displacement map. Math. Proc. Camb. Philos. Soc. **110**, 569–579 (1991)
12. Duffaut Espinosa, L.A., Ebrahimi-Fard, K., Gray, W.S.: A combinatorial Hopf algebra for nonlinear output feedback control systems. J. Algebra **453**, 609–643 (2016)
13. Duffaut Espinosa, L.A., Gray, W.S.: Integration of output tracking and trajectory generation via analytic left inversion. In: Proceedings of the 21st International Conference on System Theory, Control and Computing, Sinaia, pp. 802–807 (2017)
14. Ebrahimi-Fard, K., Gray, W.S.: Center problem, Abel equation and the Faà di Bruno Hopf algebra for output feedback. Int. Math. Res. Not. **2017**, 5415–5450 (2017)
15. Ferfera, A.: Combinatoire du Monoïde Libre Appliquée à la Composition et aux Variations de Certaines Fonctionnelles Issues de la Théorie des Systèmes. Doctoral dissertation, University of Bordeaux I (1979)
16. Ferfera, A.: Combinatoire du monoïde libre et composition de certains systèmes non linéaires. Astérisque **75**–**76**, 87–93 (1980)
17. Figueroa, H., Gracia-Bondía, J.M.: Combinatorial Hopf algebras in quantum field theory I. Rev. Math. Phys. **17**, 881–976 (2005)
18. Fliess, M.: Fonctionnelles causales non linéaires et indéterminées non commutatives. Bull. Soc. Math. France **109**, 3–40 (1981)
19. Fliess, M.: Réalisation locale des systèmes non linéaires, algèbres de Lie filtrées transitives et séries génératrices non commutatives. Invent. Math. **71**, 521–537 (1983)
20. Foissy, L.: The Hopf algebra of Fliess operators and its dual pre-Lie algebra. Commun. Algebra **43**, 4528–4552 (2015)
21. Frabetti, A., Manchon, D.: Five interpretations of Faà di Bruno's formula. In: Ebrahimi-Fard, K., Fauvet, F. (eds.) Faà di Bruno Hopf Algebras, Dyson-Schwinger Equations, and Lie-Butcher Series. IRMA Lectures in Mathematics and Theoretical Physics, vol. 21, pp. 91–147. European Mathematical Society, Zürich (2015)
22. Giné, J., Grau, M., Santallusia, X.: The center problem and composition condition for Abel differential equations. Expo. Math. **34**, 210–222 (2016)
23. Gray, W.S., Duffaut Espinosa, L.A.: A Faà di Bruno Hopf algebra for a group of Fliess operators with applications to feedback. Syst. Control Lett. **60**, 441–449 (2011)

24. Gray, W.S., Wang, Y.: Fliess operators on L_p spaces: convergence and continuity. Syst. Control Lett. **46**, 67–74 (2002)
25. Gray, W.S., Wang, Y.: Formal Fliess operators with applications to feedback interconnections. In: Proceedings of the 18th International Symposium Mathematical Theory of Networks and Systems, Blacksburg (2008)
26. Gray, W.S., Duffaut Espinosa, L.A.: A Faà di Bruno Hopf algebra for analytic nonlinear feedback control systems. In: Ebrahimi-Fard, K., Fauvet, F., (eds.) Faà di Bruno Hopf Algebras, Dyson-Schwinger Equations, and Lie-Butcher Series. IRMA Lectures in Mathematics and Theoretical Physics, vol. 21, pp. 149–217. European Mathematical Society, Zürich (2015)
27. Gray, W.S., Duffaut Espinosa, L.A., Ebrahimi-Fard, K.: Recursive algorithm for the antipode in the SISO feedback product. In: Proceedings of the 21st International Symposium on the Mathematical Theory of Networks and Systems, Groningen, pp. 1088–1093 (2014)
28. Gray, W.S., Ebrahimi-Fard, K.: SISO affine feedback transformation group and its Faà di Bruno Hopf algebra. SIAM J. Control Optim. **55**, 885–912 (2017)
29. Gray, W.S., Duffaut Espinosa, L.A., Ebrahimi-Fard, K.: Faà di Bruno Hopf algebra of the output feedback group for multivariable Fliess operators. Syst. Control Lett. **74**, 64–73 (2014)
30. Gray, W.S., Duffaut Espinosa, L.A., Ebrahimi-Fard, K.: Analytic left inversion of multivariable Lotka-Volterra models. In: Proceedings of the 54nd IEEE Conference on Decision and Control, Osaka, pp. 6472–6477 (2015)
31. Gray, W.S., Duffaut Espinosa, L.A., Thitsa, M.: Left inversion of analytic nonlinear SISO systems via formal power series methods. Automatica **50**, 2381–2388 (2014)
32. Isidori, A.: Nonlinear Control Systems, 3rd edn. Springer, London (1995)
33. Kawski, M., Sussmann, H.J., Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory. In: Helmke, U., Pratzel-Wolters, D., Zerz, E. (eds.) Operators, Systems, and Linear Algebra: Three Decades of Algebraic Systems Theory, pp. 111–128. Teubner B.G, Stuttgart (1997)
34. Lijun, Y., Yun, T.: Some new results on Abel equations. J. Math. Anal. Appl. **261**, 100–112 (2001)
35. Lloyd, N.G.: Small amplitude limit cycles of polynomial differential equations. In: Everitt, W.N., Lewis, R.T. (eds.) Ordinary Differential Equations and Operators. Lecture Notes in Mathematics, vol. 1032, pp. 346–357. Springer, Berlin (1982)
36. Manchon, D.: Hopf algebras and renormalisation. In: Hazewinkel, M. (ed.) Handbook of Algebra, vol. 5, pp. 365–427. Elsevier, Amsterdam (2008)
37. Nijmeijer, H., van der Schaft, A.J.: Nonlinear Dynamical Control Systems. Springer, New York (1990)
38. Poincaré, H.: Mémoire sur les courbes définies par une équation différentielle. Journal de Mathématiques, Series 3. **7**, 375–422 (1881); **8**, 251–296 (1882)
39. Reutenauer, C.: Free Lie Algebras. Oxford University Press, New York (1993)
40. Sweedler, M.E.: Hopf Algebras. Benjamin, W.A., Inc., New York (1969)
41. Thitsa, M., Gray, W.S.: On the radius of convergence of interconnected analytic nonlinear input-output systems. SIAM J. Control Optim. **50**, 2786–2813 (2012)
42. Wang, Y.: Differential Equations and Nonlinear Control Systems, Doctoral dissertation, Rutgers University, New Brunswick (1990)
43. Wang, Y.: Analytic constraints and realizability for analytic input/output operators. J. Math. Control Inf. **12**, 331–346 (1995)
44. Wang, Y., Sontag, E.D.: Generating series and nonlinear systems: analytic aspects, local realizability and i/o representations. Forum Math. **4**, 299–322 (1992)
45. Wang, Y., Sontag, E.D.: Algebraic differential equations and rational control systems. SIAM J. Control Optim. **30**, 1126–1149 (1992)
46. Wang, Y., Sontag, E.D.: Orders of input/output differential equations and state-space dimensions. SIAM J. Control Optim. **33**, 1102–1126 (1995)
47. Yomdin, Y.: The center problem for the Abel equations, compositions of functions, and moment conditions. Moscow Math. J. **3**, 1167–1195 (2003)

Continuous-Time Autoregressive Moving-Average Processes in Hilbert Space

Fred Espen Benth and André Süss

Abstract We introduce the class of continuous-time autoregressive moving-average (CARMA) processes in Hilbert spaces. As driving noises of these processes we consider Lévy processes in Hilbert space. We provide the basic definitions, show relevant properties of these processes and establish the equivalents of CARMA processes on the real line. Finally, CARMA processes in Hilbert space are linked to the stochastic wave equation and functional autoregressive processes.

1 Introduction

Continuous-time autoregressive moving-average processes, or CARMA for short, play an important role in modelling the stochastic dynamics of various phenomena like wind speed, temperature variations and economic indices. For example, based on such models, in [1] the author analyses fixed-income markets while in [8] and [15] the dynamics of weather factors at various locations in Europe and Asia are modelled. Finally, in [5, 7] and [19] continuous-time autoregressive models for commodity markets like power and oil are studied. The versatile class of CARMA processes can flexibly model stationarity, memory and non-Gaussian effects in data in many areas in natural science, engineering and economics.

CARMA processes constitute the continuous-time version of autoregressive moving-average time series models. In this paper we generalize these processes to a Hilbert space context. Hilbert-valued CARMA processes will form a continuous-time version of functional autoregressive processes studied by [10]. The area of functional data analysis, or the statistics of curves and surfaces, has gained attention

F. E. Benth (✉)
Department of Mathematics, University of Oslo, Oslo, Norway
e-mail: fredb@math.uio.no

A. Süss
Facultat de Matematiques, Universitat de Barcelona, Barcelona, Spain
e-mail: suess.andre@web.de

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_11

in recent years (see for example [25] for a thorough review with references). CARMA processes in Hilbert space can be attractive for modeling futures price curves in finance or weather dynamics in continuous space and time. These processes also provide an interesting theoretical tool linking higher-order stochastic partial differential equations to a "multivariate" infinite dimensional dynamics.

The crucial ingredient in the extension of the CARMA dynamics to infinite dimensions is a "multivariate" Ornstein-Uhlenbeck process with values in a Hilbert space. There already exists an analysis of infinite dimensional Lévy-driven Ornstein-Uhlenbeck processes, and we refer the reader to the survey [3]. Moreover, matrix-valued operators and their semigroups play an important role. In [14] a detailed semigroup theory for such operators is developed. We review some of the results from [3] and [14] in the context of Hilbert-valued CARMA processes, as well as providing some new results for these processes.

Let us recall the definition of a real-valued CARMA process. We follow [11] and first introduce the multivariate Ornstein-Uhlenbeck process $\{\mathbf{Z}(t)\}_{t\geq 0}$ with values in $\mathbb{R}^p$ for $p \in \mathbb{N}$ by

$$\mathrm{d}\mathbf{Z}(t) = C_p\mathbf{Z}(t)\mathrm{d}t + \mathbf{e}_p\mathrm{d}L(t),\ \mathbf{Z}(0) = \mathbf{Z}_0 \in \mathbb{R}^p. \tag{1}$$

Here, L is a one-dimensional square integrable Lévy process with zero mean defined on a complete probability space $(\Omega, \mathscr{F}, P)$ with filtration $\mathscr{F} = \{\mathscr{F}_t\}_{t\geq 0}$, satisfying the usual hypotheses. Furthermore, $\mathbf{e}_i$ is the ith canonical unit vector in $\mathbb{R}^p$, $i = 1, \ldots, p$. The $p \times p$ matrix C_p takes the particular form

$$C_p = \begin{bmatrix} 0 & 1 & 0 \ldots & 0 \\ 0 & 0 & 1\,0 \ldots & 0 \\ \cdot & \cdot & \cdot \ldots & \cdot \\ \cdot & \cdot & \cdot \ldots & \cdot \\ 0 & \cdot & \cdot \ldots & 1 \\ -\alpha_p & -\alpha_{p-1} & \cdot \ldots & -\alpha_1 \end{bmatrix}, \tag{2}$$

for constants $\alpha_i > 0, i = 1, \ldots, p$.[1]

We define a continuous-time autoregressive process of order p by

$$X(t) = \mathbf{e}_1^\top \mathbf{Z}(t),\ t \geq 0, \tag{3}$$

where $\mathbf{x}^\top$ means the transpose of $\mathbf{x} \in \mathbb{R}^p$. We say that $\{X(t)\}_{t\geq 0}$ is a CAR(p)-process. For $q \in \mathbb{N}$ with $p > q$, we define a CARMA(p, q)-process by

$$X(t) = \mathbf{b}^\top \mathbf{Z}(t),\ t \geq 0, \tag{4}$$

[1]The odd labelling of these constants stems from an interpretation of CARMA processes as solutions to higher-order linear stochastic differential equations.

where $\mathbf{b} \in \mathbb{R}^p$ is the vector $\mathbf{b} = (b_0, b_1, \ldots, b_{q-1}, 1, 0, .., 0)^\top \in \mathbb{R}^p$, where $b_q = 1$ and $b_i = 0, i = q+1, \ldots, p-1$. Note that $\mathbf{b} = \mathbf{e}_1$ yields a CAR(p)-process. Sampling the CARMA(p, q)-process $\{X(t)\}_{t\geq 0}$ on an equidistant discretized time grid we get an (weak) autoregressive moving average time series process (see [8, Eq. (4.17)] for an Euler-Maryuama approximation, yielding an autoregressive moving average time series of order p, q.). An explicit dynamics of the CARMA(p, q)-process $\{X(t)\}_{t\geq 0}$ are (see e.g. [9, Lemma 10.1])

$$X(t) = \mathbf{b}^\top \exp(tC_p)\mathbf{Z}_0 + \int_0^t \mathbf{b}^\top \exp((t-s)C_p)\mathbf{e}_p \mathrm{d}L(s), \tag{5}$$

where $\exp(tC_p)$ is the matrix exponential of tC_p, the matrix C_p multiplied by time t.

If C_p has only eigenvalues with negative real part, then the process X admits a limiting distribution μ_X with characteristic exponent (see [11])

$$\widehat{\mu}_X(z) := \lim_{t\to\infty} \log \mathbb{E}\left[\mathrm{e}^{\mathrm{i}zX(t)}\right] = \int_0^\infty \psi_L\left(\mathbf{b}^\top \exp(sC_p)\mathbf{e}_p z\right) \mathrm{d}s.$$

Here, Ψ_L is the log-characteristic function of $L(1)$ (see e.g. [2]) and log the distinguished logarithm (see e.g. [21, Lemma 7.6]). In particular, if $L = \sigma B$ with $\sigma > 0$ constant and B a standard Brownian motion, we find

$$\widehat{\mu}_X(z) = -\frac{1}{2}z^2\sigma^2 \int_0^\infty (\mathbf{b}^\top \exp(sC_p)\mathbf{e}_p)^2 \mathrm{d}s\,,$$

and thus X has a Gaussian limiting distribution μ_X with zero mean and variance $\sigma^2 \int_0^\infty (\mathbf{b}^\top \exp(sC_p)\mathbf{e}_p)^2 \mathrm{d}s$.

When X admits a limiting distribution, we have a stationary representation of the process X such that $X(t) \sim \mu_X$ for all $t \in \mathbb{R}$, namely,

$$X(t) = \int_{-\infty}^t \mathbf{b}^\top \exp((t-s)C_p)\mathbf{e}_p \mathrm{d}L(s), \tag{6}$$

where L is now a two-sided Lévy process. This links CARMA(p, q)-processes to the more general class of Lévy semistationary (LSS) processes, defined in [5] as

$$X(t) := \int_{-\infty}^t g(t-s)\sigma(s)\mathrm{d}L(s), \tag{7}$$

for $g : \mathbb{R}_+ \to \mathbb{R}$ being a measurable function and σ a predictable process such that $s \mapsto g(t-s)\sigma(s)$ for $s \leq t$ is integrable with respect to L. Indeed, LSS processes are again a special case of so-called *ambit fields*, which are spatio-temporal stochastic processes originally developed in [4] for modelling turbulence. An ambit field in our context can be defined as a real-valued space-time random field $\{X(t, x)\}_{t\geq 0, x\in D}$

of the form

$$X(t,x) = \int_{-\infty}^{t} \int_{D} g(t,s,x,y)\sigma(s,y)L(\mathrm{d}y,\mathrm{d}s), \tag{8}$$

where $D \subset \mathbb{R}^d$ is a Borel-measurable subset, g is a measurable real-valued function on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^d$ and σ is a real-valued predictable random field on $\mathbb{R}_+ \times \mathbb{R}^d$. Furthermore, L is a so-called Lévy basis, which means that it is an independently scattered infinitely divisible random measure on $\mathscr{B}_b(\mathbb{R}^{d+1})$, the set of bounded Borel sets on $\mathbb{R}^{d+1}$. Under appropriate conditions on g, σ and L (see [4] and [24]), the stochastic integral in (7) makes sense as an Itô-type integral.

The infinite dimensional CARMA processes that we are going to define in this paper will form a subclass of ambit fields, as we will see in Sect. 4. We note that CARMA processes with values in $\mathbb{R}^n$ have been defined and analysed by [18, 22] and recently in [16]. In [12] we find a definition of multivariate CARMA processes which is related to our infinite dimensional approach.

2 Definition of CARMA Processes in Hilbert Space

Given $p \in \mathbb{N}$, and let H_i for $i = 1, \ldots, p$ be separable Hilbert spaces with inner products denoted by $\langle \cdot, \cdot \rangle_i$ and associated norms $|\cdot|_i$. We define the product space $H := H_1 \times \ldots \times H_p$, which is again a separable Hilbert space equipped with the inner product $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{p} \langle x_i, y_i \rangle_i$ and the induced norm denoted $|\cdot| = \sum_{i=1}^{n} |\cdot|_i$ for $\mathbf{x} = (x_1, \ldots, x_p), \mathbf{y} = (y_1, \ldots, y_p) \in H$. The projection operator $\mathscr{P}_i : H \to H_i$ is defined as $\mathscr{P}_i \mathbf{x} = x_i$ for $\mathbf{x} \in H, i = 1, \ldots, p$. It is straightforward to see that its adjoint $\mathscr{P}_i^* : H_i \to H$ is given by $\mathscr{P}_i^* x = (0, \ldots, 0, x, 0, \ldots, 0)$ for $x \in H_i$, where the x appears in the ith coordinate of the vector consisting of p elements. If U and V are two separable Hilbert spaces, we denote $L(U, V)$ the Banach space of bounded linear operators from U to V, equipped with the operator norm $\|\cdot\|_{\mathrm{op}}$. The Hilbert-Schmidt norm for operators in $L(U, V)$ is denoted $\|\cdot\|_{\mathrm{HS}}$, and $L_2(U, V)$ denotes the space of Hilbert-Schmidt operators. If $U = V$, we simply write $L(U)$ for $L(U, U)$.

Let $A_i : H_{p+1-i} \to H_p, i = 1, \ldots, p$ be p (unbounded) densely defined linear operators, and $I_i : H_{p+2-i} \to H_{p+1-i}, i = 2, \ldots, p$ be another $p - 1$ (unbounded) densely defined linear operators. Define the linear operator $\mathscr{C}_p : H \to H$ represented as a $p \times p$ matrix of operators

$$\mathscr{C}_p = \begin{bmatrix} 0 & I_p & 0 & \ldots & 0 \\ 0 & 0 & I_{p-1} & 0 \ldots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \cdot & \ldots & I_2 \\ A_p & A_{p-1} & \cdot & \ldots & A_1 \end{bmatrix}. \tag{9}$$

Since the A_i's and I_i's are densely defined, $\mathscr{C}_p$ has domain

$$Dom(\mathscr{C}_p) = Dom(A_p) \times (Dom(A_{p-1}) \cap Dom(I_p)) \times \ldots \times (Dom(A_1) \cap Dom(I_2)),$$

which we suppose is dense in H. We note in passing that typically, $H_1 = H_2 = \ldots = H_p$ and $I_i = \mathrm{Id}$, the identity operator on H_i, $i = 1, \ldots, p$. Then $Dom(\mathscr{C}_p) = Dom(A_p) \times Dom(A_{p-1}) \times \ldots \times Dom(A_1)$, which is dense in H.

A family $\{\mathscr{S}(t)\}_{t\geq 0} \subset L(H)$ of operators is said to be a C_0-*semigroup* if $\mathscr{S}(0) = \mathrm{Id}$, $\mathscr{S}(t)\mathscr{S}(s) = \mathscr{S}(t+s)$ for any $t, s \geq 0$ and $\mathscr{S}(t)\mathbf{x} \to \mathbf{x}$ in H whenever $t \downarrow 0$ for all $\mathbf{x} \in H$. From [14, Ch. II. Thm. 1.4], we know that there exists a densely defined linear operator $\mathscr{C}$ on H such that

$$\mathscr{C}\mathbf{x} = \lim_{t\downarrow 0} \frac{1}{t} \left(\mathscr{S}(t)\mathbf{x} - \mathbf{x}\right),$$

for all $\mathbf{x} \in \mathrm{Dom}(\mathscr{C})$, where the limit is taken in H. One says that $\mathscr{C}$ is the *generator* of the C_0-semigroup $\{\mathscr{S}(t)\}_{t\geq 0}$. The question of when a densely defined operator $\mathscr{C}$ is a generator of a C_0-semigroup can be answered by the generation theorem of Hille and Yoshida (see [14, Ch. II, Thm. 3.5]) in the contractive case: if $R(\lambda, \mathscr{C})$ denotes the resolvent of $\mathscr{C}$, then $\mathscr{C}$ is a generator if and only if $\mathscr{C}$ is a closed operator and for every $\lambda > 0$, λ is in the resolvent set and $\|\lambda R(\lambda, \mathscr{C})\|_{\mathrm{op}} \leq 1$. If the densely defined linear operator $\mathscr{C}_p$ in (9) is the generator of a C_0-semigroup, we denote this semigroup by $\{\mathscr{S}_p(t)\}_{t\geq 0}$ from now on.

On a complete probability space $(\Omega, \mathscr{F}, P)$ with filtration $\mathscr{F} = \{\mathscr{F}_t\}_{t\geq 0}$ satisfying the usual hypotheses, denote by $L := \{L(t)\}_{t\geq 0}$ a zero-mean square-integrable H_p-valued Lévy process with covariance operator Q (i.e., a symmetric non-negative definite trace class operator), defined as follows (see e.g. [20, Sect. 4.9]):

Definition 1 An H_p-valued stochastic process $L = \{L(t)\}_{t\geq 0}$ is called a Lévy process if $L(0) = 0$, L is stochastically continuous, $L(t) - L(s)$ is independent of $L(u) - L(v)$ for all $t > s \geq u > v \geq 0$ and the law of $L(t) - L(s)$ depends only on $t - s$.

An H_p-valued Lévy process is thus, in short, a stochastically continuous process on H_p starting at zero which has independent and stationary increments. The process $L = \{L(t)\}_{t\geq 0}$ is square-integrable whenever $\mathbb{E}[|L(t)|_p^2] < \infty$ for all $t \geq 0$. For a square-integrable Lévy process $L = \{L(t)\}_{t\geq 0}$ with zero mean, it follows from [20, Thm. 4.44] that there exists a linear operator $Q \in L(H_p)$ being symmetric, non-negative definite trace class such that

$$\mathbb{E}[\langle L(t), x\rangle_p \langle L(s), y\rangle_p] = t \wedge s \, \langle Qx, y\rangle_p, \quad x, y \in H_p, \quad t, s \geq 0.$$

One refers to Q as the covariance operator of L.

Consider the following stochastic differential equation. For $t \geq 0$,

$$\mathrm{d}\mathbf{Z}(t) = \mathscr{C}_p \mathbf{Z}(t)\mathrm{d}t + \mathscr{P}_p^* \mathrm{d}L(t), \ \mathbf{Z}(0) := \mathbf{Z}_0 \in H. \tag{10}$$

This H-valued Ornstein-Uhlenbeck process is a special case of a more general stochastic differential equation in H of the form

$$\mathrm{d}\mathbf{Z}(t) = \mathscr{C}_p\mathbf{Z}(t)\mathrm{d}t + F(\mathbf{Z}(t))\mathrm{d}t + G(\mathbf{Z}(t))\mathrm{d}L(t),\ \mathbf{Z}(0) := \mathbf{Z}_0 \in H, \tag{11}$$

where $F : H \to H$ and $G : H \to L_2(H_p, H)$ are Lipschitz continuous operators. In the case $\mathscr{C}_p$ is the generator of a C_0-semigroup, a *mild solution* of (11) is defined according to [20, Def. 9.5] (see also Remark 9.4 in [20]) as follows:

Definition 2 A predictable H-valued stochastic process $\mathbf{Z} = \{\mathbf{Z}(t)\}_{t\geq 0}$ is said to be a *mild solution* of (11) if $\sup_{t\in[0,T]} \mathbb{E}[|\mathbf{Z}(t)|^2] < \infty$ for all $0 < T < \infty$ and

$$\mathbf{Z}(t) = \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{S}_p(t-s)F(\mathbf{Z}(s))\mathrm{d}s + \int_0^t \mathscr{S}_p(t-s)G(\mathbf{Z}(s))\mathrm{d}L(s),$$

for all $t \geq 0$.

The next proposition states an explicit expression for the mild solution of (10)

Proposition 1 *Assume that $\mathscr{C}_p$ defined in (9) is the generator of a C_0-semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$ on H. Then the H-valued stochastic process $\mathbf{Z}$ given by*

$$\mathbf{Z}(t) = \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{S}_p(t-s)\mathscr{P}_p^*\mathrm{d}L(s)$$

is the unique mild solution of (10).

Proof We have that $\mathscr{S}_p(t-s)\mathscr{P}_p^* \in L(H_p, H)$, and moreover, since $\|\mathscr{P}_p^*\|_{\mathrm{op}} = 1$ it follows

$$\|\mathscr{S}_p(t-s)\mathscr{P}_p^*Q^{1/2}\|_{\mathrm{HS}} \leq \|\mathscr{S}_p(t-s)\|_{\mathrm{op}}\|\mathscr{P}_p^*\|_{\mathrm{op}}\|Q^{1/2}\|_{\mathrm{HS}} \leq K\mathrm{e}^{c(t-s)}\|Q^{1/2}\|_{\mathrm{HS}}$$

by the general exponential growth bound on the operator norm of a C_0-semigroup (see e.g. [14, Prop. I.5.5]). Thus, for all $t \geq 0$,

$$\int_0^t \|\mathscr{S}_p(t-s)\mathscr{P}_p^*Q^{1/2}\|_{\mathrm{HS}}^2\mathrm{d}s \leq \frac{K}{2c}\mathrm{e}^{2ct}\|Q^{1/2}\|_{\mathrm{HS}}^2 < \infty$$

because Q is trace class by assumption. The stochastic integral with respect to L in the definition of $\mathbf{Z}$ is therefore well-defined. Hence, the result follows directly from the existence and uniqueness theorem of mild solutions in [20, Thm. 9.29].

From now on we restrict our attention to operators $\mathscr{C}_p$ in (9) which admit a C_0-semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$. We remark in passing that in the next section we will provide a recursive definition of the semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$ in a special situation where all involved operators are bounded except A_1.

A CARMA process with values in a Hilbert space is defined next:

Definition 3 Let U be a separable Hilbert space. For $\mathscr{L}_U \in L(H, U)$, define the U-valued stochastic process $X := \{X(t)\}_{t\geq 0}$ by

$$X(t) := \mathscr{L}_U \mathbf{Z}(t), t \geq 0,$$

for $\mathbf{Z}(t)$ defined in (10). We call $\{X(t)\}_{t\geq 0}$ a CARMA$(p, U, \mathscr{L}_U)$-process.

Note that we do not have any q-parameter present in the definition, as in the real-valued case (recall (4)). Instead we introduce a Hilbert space and a linear operator as the "second" parameters in the CARMA$(p, U, \mathscr{L}_U)$-process. Indeed, the vector $\mathbf{b}$ in the real-valued CARMA(p, q)-process defined in (4) can be viewed as a linear operator from $\mathbb{R}^p$ into $\mathbb{R}$ by the scalar product operation $\mathbb{R}^p \ni \mathbf{z} \mapsto \mathbf{b}'\mathbf{z} \in \mathbb{R}$, or, by choosing $U = H_i = \mathbb{R}$, $\mathscr{L}_{\mathbb{R}}\mathbf{z} = \mathbf{b}'\mathbf{z}$. This also demonstrates that any real-valued CARMA(p, q)-process is a CARMA$(p, \mathbb{R}, \mathbf{b}'\cdot)$-process according to Definition 3. We further remark that our definition of a CARMA process in Hilbert space can be viewed as a natural extension of the controller canonical representation of a multivariate (i.e., finite dimensional) CARMA process introduced in [12].

From Proposition 1 we find that the explicit representation of $\{X(t)\}_{t\geq 0}$ is

$$X(t) = \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{L}_U \mathscr{S}_p(t-s)\mathscr{P}_p^* \mathrm{d}L(s), \tag{12}$$

for $t \geq 0$. Note that by linearity of the stochastic integral we can move the operator $\mathscr{L}_U$ inside. Furthermore, the stochastic integral is well-defined since $\mathscr{L}_U \in L(H, U)$ and thus has a finite operator norm.

The continuous-time autoregressive (CAR) processes constitute a particularly interesting subclass of the CARMA$(p, U, \mathscr{L}_U)$-processes:

Definition 4 The CARMA$(p, H_1, \mathscr{P}_1)$-process $\{X(t)\}_{t\geq 0}$ from Definition 3 is called an H_1-valued CAR(p)-process.

The explicit dynamics of an H_1-valued CAR(p)-process becomes

$$X(t) = \mathscr{P}_1 \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{P}_1 \mathscr{S}_p(t-s)\mathscr{P}_p^* \mathrm{d}L(s), \tag{13}$$

for $t \geq 0$. In this paper we will be particularly focused on H_1-valued CAR(p)-processes.

Remark that the process $\mathbf{L} := \mathscr{P}_p^* L$ defines an H-valued Lévy process which has mean zero and is square integrable. Its covariance operator is easily seen to be $\mathscr{P}_p^* Q \mathscr{P}_p$.

It is immediate to see that an H_1-valued CAR(1) process is an Ornstein-Uhlenbeck process defined on H_1, with

$$\mathrm{d}X(t) = A_1 X(t)\mathrm{d}t + \mathrm{d}L(t),$$

and thus

$$X(t) = \mathscr{S}_1(t)Z_0 + \int_0^t \mathscr{S}_1(t-s)\mathrm{d}L(s),$$

being its mild solution.

An H_1-valued CAR(p) process for $p > 1$ can be viewed as a higher-order (indeed, a pth order) stochastic differential equation, as we now discuss.

Proposition 2 *Suppose that $Ran(A_q) \subset Dom(I_2)$ and $Ran(I_q) \subset Dom(I_{q+1})$, and assume that there exist $p-1$ linear (unbounded) operators $B_1, B_2, \ldots, B_{p-1} : H_1 \to H_1$ such that*

$$I_p \cdots I_2 A_q = B_q I_p I_{p-1} \cdots I_{q+1}. \tag{14}$$

for $q = 1, \ldots, p-1$. We suppose that $Dom(B_q)$ is so that $Dom(B_q I_p I_{p-1} \cdots I_{q+1}) = Dom(A_q)$. Furthermore we define the operator $B_p : H_1 \to H_1$ as

$$B_p := I_p \cdots I_2 A_p. \tag{15}$$

and suppose that B_p is a linear (possibly unbounded) operator with domain $Dom(B_p) = Dom(A_p)$. Then,

$$\mathrm{d}X^{(p-1)}(t) = \left(\sum_{q=1}^{p} B_q X^{(p-q)}(t) \right) \mathrm{d}t + I_p \cdots I_2 \mathrm{d}L(t), \tag{16}$$

where $X^{(q)}(t)$ denotes the qth derivative of $X(t)$, $q = 1, \ldots, p-1$.

Proof We note that $I_p \cdots I_2 : H_p \to H_1$ and hence $I_p \cdots I_2 A_q : H_{p+1-q} \to H_1$. Moreover, $I_p \cdots I_{q+1} : H_{p+1-q} \to H_1$, and therefore $B_q : H_1 \to H_1$ for $q = 1, \ldots, p-1$. We also observe that that $Dom(A_q)$ is the domain of the operator $I_p \cdots I_2 A_q$. We see further that the definition of B_p is consistent with the inductive relations in (14).

By definition, $X(t) = \mathscr{P}_1 \mathbf{Z}(t)$, which is the first coordinate in the vector $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_p(t))^\top \in H$. From (10) and the definition of the operator matrix $\mathscr{C}_p$ in (9), we find that $Z_1'(t) = I_p Z_2(t)$, $Z_2'(t) = I_{p-1} Z_3(t), \ldots, Z_{p-1}'(t) = I_2 Z_p(t)$ and finally

$$Z_p'(t) = A_p Z_1(t) + \cdots A_1 Z_1(t) + L'(t).$$

Here, $L'(t)$ is the formal time derivative of L. By iteration, we find that $Z_1^{(q)}(t) = I_p I_{p-1} \cdots I_{p-(q-1)} Z_{q+1}(t)$ for $q = 1, \ldots, p-1$. Thus,

$$\begin{aligned} Z_1^{(p)} &= \frac{d}{dt} Z_1^{(p-1)} = I_p \cdots I_2 Z_p'(t) \\ &= I_p \cdots I_2 A_p Z_1(t) + I_p \cdots I_2 A_{p-1} Z_2(t) + \cdots + I_p \cdots I_2 A_1 Z_p(t) + I_p \cdots I_2 L'(t) \\ &= B_p Z_1(t) + B_{p-1} Z_1'(t) + B_{p-2} Z_1^{(2)}(t) + \ldots + B_1 Z_1^{(p-1)}(t) + I_p \cdots I_2 L'(t). \end{aligned}$$

In the last equality we made use of (14) and (15). After multiplying both sides above with $\mathrm{d}t$, we find that an H_1-valued CAR(p) process $X(t) = \mathscr{P}_1 \mathbf{Z}(t)$ is the solution to the pth-order stochastic differential equation (16).

Let us introduce the operator-valued pth-order polynomial $Q_p(\lambda)$ for $\lambda \in \mathbb{C}$,

$$Q_p(\lambda) = \lambda^p - B_1 \lambda^{p-1} - B_2 \lambda^{p-2} - \cdots - B_{p-1}\lambda - B_p. \tag{17}$$

Inspecting the proof of the proposition above, we see that we can express informally the CAR(p) process $\{X(t)\}_{t \geq 0}$ as the solution of the pth-order differential equation,

$$Q_p\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) X(t) = I_p \cdots I_2 L'(t). \tag{18}$$

The form of the operator-valued polynomial Q_p is a consequence of the specification of the CARMA process by the matrix operator $\mathscr{C}_p$ in (9).

If $H_1 = \ldots = H_p$ and $\mathscr{C}_p$ is a bounded operator, then $B_q = I_q \cdots I_2 A_q$ in (14) whenever $I_q \cdots I_2 A_q$ commutes with $I_p \cdots I_{q+1}$. In this sense the condition (14) is a specific commutation relationship on A_q and the operators $I_2, \ldots, I_p$. In the particular case $I_i = \mathrm{Id}$ for $i = 2, \ldots, p$, then we trivially have $A_q = B_q$ for $q = 1, \ldots, p$. As a special case, let us for a moment suppose that $p = 2$, and consider a CARMA(2,H_1,$\mathscr{L}_{H_1}$)-process $\{X(t)\}_{t \geq 0}$. As $H = H_1^{\times 2}$ and $U = H_1$ in this case, we represent $\mathscr{L}_{H_1}$ as a vector-valued operator $\mathscr{L}_{H_1} := (M_1, M_2)$, where $M_i \in L(H_1)$, $i = 1, 2$. We assume that M_i commutes with A_j for all $i, j = 1, 2$ (recall that A_1 and A_2 are now bounded). By definition, $X(t) = M_1 Z_1(t) + M_2 Z_2(t)$. Doing an informal calculation, we find, using the relationships for Z_1 and Z_2 and the commutation assumptions,

$$\begin{aligned} &Q_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) X(t) \\ &\quad = X''(t) - A_1 X'(t) - A_2 X(t) \\ &\quad = M_1 Z_1''(t) + M_2 Z_2''(t) - A_1 M_1 Z_1'(t) - A_1 M_2 Z_2'(t) - A_2 M_1 Z_1(t) - A_2 M_2 Z_2(t) \\ &\quad = (M_1 A_2 + M_2 A_1 A_2 - A_1 M_2 A_2 - A_2 M_2)\, Z_1(t) \end{aligned}$$

$$
\begin{aligned}
&+\left(M_1A_1+M_2A_2+M_2A_1^2-A_1M_1-A_1M_2A_1-A_2M_2\right)Z_2(t)\\
&+(M_2A_1-A_1M_2+M_1)L'(t)+M_2L''(t)\\
&=M_1L'(t)+M_2L''(t).
\end{aligned}
$$

Indeed, for a general $p \in \mathbb{N}$ and under the assumption that M_i commutes with A_j for all $i, j = 1, \ldots, p$, we can extend the above derivation to

$$
Q_p\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)X(t)=R_{p-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)L'(t)
$$

for the operator-valued $(p-1)$th-order polynomial $R_{p-1}(\lambda)$, $\lambda \in \mathbb{C}$,

$$
R_{p-1}(\lambda)=M_p\lambda^{p-1}+M_{p-1}\lambda^{p-2}+\cdots+M_2\lambda+M_1. \tag{19}
$$

Hence, informally, a CARMA$(p, H_1, \mathscr{L}_{H_1})$-process $\{X(t)\}_{t\geq 0}$ can, under rather strong conditions on commutativity, be represented by an "autoregressive" polynomial operator Q_p and a "moving average" polynomial operator R_{p-1}. This is a representation that we also find for multivariate CARMA processes, see [12].

Although the choice of $I_i = \mathrm{Id}$ for $i = 2, \ldots, p$ (with $H_1 = \ldots = H_p$) is the canonical choice from the point of view of the finite dimensional CARMA processes (see [18, 22]), it may be convenient with more flexibility in the Hilbert-valued case. For example, with our generality, we may choose the state space H_p of the noise $\{L(t)\}_{t\geq 0}$ to be different than the state space H_1 of the process $\{X(t)\}_{t\geq 0}$. This can accommodate situations where there is a finite-dimensional noise, but with the process $\{X(t)\}_{t\geq 0}$ taking values in an infinite dimensional space. This is the case for many models of forward rates in fixed-income markets in finance (see [13]). The operators I_i may also be viewed as a "volatility" which scales the noise in the sense that it acts on the Hilbert-structure of L (recall (18)).

We end this section with showing that the stochastic wave equation can be viewed as an example of a Hilbert-valued CAR(2)-process. To this end, let $H_2 := L^2(0, 1)$, the space of square-integrable functions on the unit interval, and consider the stochastic partial differential equation

$$
\frac{\partial^2 Y(t,x)}{\partial t^2}=\frac{\partial^2 Y(t,x)}{\partial x^2}+\frac{\partial L(t,x)}{\partial t}, \tag{20}
$$

with $t \geq 0$ and $x \in (0, 1)$. We can rephrase this wave equation as

$$
\mathrm{d}\begin{bmatrix} Y(t,x) \\ \frac{\partial Y(t,x)}{\partial t} \end{bmatrix}=\begin{bmatrix} 0 & \mathrm{Id} \\ \Delta & 0 \end{bmatrix}\begin{bmatrix} Y(t,x) \\ \frac{\partial Y(t,x)}{\partial t} \end{bmatrix}\mathrm{d}t+\begin{bmatrix} 0 \\ \mathrm{d}L(t,x) \end{bmatrix}, \tag{21}
$$

with $\Delta = \partial^2/\partial x^2$ being the Laplace operator. The eigenvectors $e_n(x) := \sqrt{2}\sin(\pi n x)$, $n \in \mathbb{N}$, for Δ form an orthonormal basis of $L^2(0,1)$. Introduce the Hilbert space H_1 as the subspace of $L^2(0,1)$ for which $|f|_1^2 := \pi^2 \sum_{n=1}^{\infty} n^2 \langle f, e_n\rangle_2^2 < \infty$. Following Example B.13 in [20],

$$\mathscr{C}_2 = \begin{bmatrix} 0 & \mathrm{Id} \\ \Delta & 0 \end{bmatrix}$$

generates a C_0-semigroup $\mathscr{S}_2(t)$ on $H := H_1 \times H_2$. The Laplace operator Δ is a self-adjoint negative definite operator on H_1. The semigroup $\mathscr{S}_2(t)$ can be represented as

$$\mathscr{S}_2(t) = \begin{bmatrix} \cos((-\Delta)^{1/2}t) & (-\Delta)^{-1/2}\sin((-\Delta)^{1/2}t) \\ -(-\Delta)^{1/2}\sin((-\Delta)^{1/2}t) & \cos((-\Delta)^{1/2}t) \end{bmatrix}. \tag{22}$$

In the previous equality, we define for a real-valued function g the linear operator $g(\Delta)$ using functional calculus, i.e., $g(\Delta)f = \sum_{n=1}^{\infty} g(-\pi^2 n^2)\langle f, e_n\rangle_2 e_n$ whenever this sum converges. These considerations show that the wave equation is a specific example of a CAR(2)-process.

3 Analysis of CARMA Processes

In this section we derive some fundamental properties of CARMA processes in Hilbert spaces.

3.1 Distributional Properties

We state the conditional characteristic functional of a CARMA$(p, U, \mathscr{L}_U)$-process in the next proposition.

Proposition 3 *Assume X is a CARMA$(p, U, \mathscr{L}_U)$-process. Then, for $x \in U$,*

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle X(t),x\rangle_U} \mid \mathscr{F}_s\right] = \exp\left(\mathrm{i}\langle \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0, x\rangle_U + \int_0^{t-s} \psi_L\left(\mathscr{P}_p \mathscr{S}_p^*(u)\mathscr{L}_U^* x\right)\mathrm{d}u\right)$$
$$\times \exp\left(\mathrm{i}\langle \int_0^s \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u), x\rangle_U\right),$$

for $0 \le s \le t$. Here, ψ_L is the characteristic exponent of the Lévy process L.

Proof From (12) it holds for $0 \leq s \leq t$,

$$X(t) = \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^s \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u) + \int_s^t \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u).$$

The Lévy process has independent increments, and $\mathscr{F}_s$-measurability of the first stochastic integral thus yields

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle X(t),x\rangle_U} \mid \mathscr{F}_s\right] = \exp\left(\mathrm{i}\langle \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0, x\rangle_U + \mathrm{i}\langle \int_0^s \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u), x\rangle_U\right)$$
$$\times \mathbb{E}\left[\exp\left(\mathrm{i}\langle \int_s^t \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u), x\rangle_U\right)\right].$$

The result follows from [20, Chapter 4].

Suppose now that $L = W$, an H_p-valued Wiener process. Then the characteristic exponent is

$$\psi_W(h) = -\frac{1}{2}\langle Qh, h\rangle_p,$$

for $h \in H_p$. Hence, from Proposition 3 it follows that,

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle X(t),x\rangle_U} \mid \mathscr{F}_s\right] = \exp\left(\mathrm{i}\langle \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0, x\rangle_U + \mathrm{i}\langle \int_0^s \mathscr{L}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}W(u), x\rangle_U\right)$$
$$\times \exp\left(-\frac{1}{2}\int_0^{t-s} \langle \mathscr{L}_U \mathscr{S}_p(u)\mathscr{P}_p^* Q \mathscr{P}_p \mathscr{S}_p^*(u)\mathscr{L}_U^* x, x\rangle_U \mathrm{d}u\right)$$

We find that $X(t)|\mathscr{F}_s$ for $s \leq t$ is a Gaussian process in H_1, with mean

$$\mathbb{E}\left[X(t) \mid \mathscr{F}_s\right] = \mathscr{L}_U \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^s \mathscr{W}_U \mathscr{S}_p(t-u)\mathscr{P}_p^* \mathrm{d}L(u)$$

and covariance operator

$$\mathrm{Var}(X(t)|\mathscr{F}_s) = \int_0^{t-s} \mathscr{L}_U \mathscr{S}_p(u)\mathscr{P}_p^* Q \mathscr{P}_p \mathscr{S}_p^*(u)\mathscr{L}_U^* \mathrm{d}u,$$

where the integral is interpreted in the Bochner sense. If the semigroup $\mathscr{S}_p(u)$ is exponentially stable, then $X(t)|\mathscr{F}_s$ admits a Gaussian limiting distribution with mean zero and covariance operator

$$\lim_{t\to\infty} \mathrm{Var}(X(t)|\mathscr{F}_s) = \int_0^\infty \mathscr{L}_U \mathscr{S}_p(u)\mathscr{P}_p^* Q \mathscr{P}_p \mathscr{S}_p^*(u)\mathscr{L}_U^* \mathrm{d}u.$$

This is the invariant measure of X. We remark in passing that measures on H are defined on its Borel σ-algebra.

In [3] there is an analysis of invariant measures of Lévy-driven Ornstein-Uhlenbeck processes. We discuss this here in the context of the Ornstein-Uhlenbeck process $\{\mathbf{Z}(t)\}_{t\geq 0}$ defined in (10). Assume $\mu_{\mathbf{Z}}$ is the invariant measure of $\{\mathbf{Z}(t)\}_{t\geq 0}$, and recall the definition of its characteristic exponent $\widehat{\mu}_{\mathbf{Z}}(\mathbf{x})$,

$$\widehat{\mu}_{\mathbf{Z}}(\mathbf{x}) = \log \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle \mathbf{x}, \mathbf{Z}(t)\rangle}\right]. \tag{23}$$

Here, $\mathbf{x} \in H$ and log is the distinguished logarithm (see e.g. [21, Lemma 7.6]). If $\mathbf{Z}_0 \sim \mu_{\mathbf{Z}}$, then, in distribution, $\mathbf{Z}_0 = \mathbf{Z}(t)$ for all $t \geq 0$ and it follows that the characteristic exponent of $\mu_{\mathbf{Z}}$ satisfies,

$$\widehat{\mu}_{\mathbf{Z}}(\mathbf{x}) = \widehat{\mu}_{\mathbf{Z}}(\mathscr{S}_p^*(t)\mathbf{x}) + \int_0^t \psi_L(\mathscr{P}_p \mathscr{S}_p^*(u)\mathbf{x})\mathrm{d}u \tag{24}$$

for any $\mathbf{x} \in H$ and $t \geq 0$. Following [3], $\mu_{\mathbf{Z}}$ becomes an operator self-decomposable distribution since,

$$\mu_{\mathbf{Z}} = \mathscr{S}_p(t)\mu_{\mathbf{Z}} \star \mu_t. \tag{25}$$

Here, μ_t is the distribution of $\int_0^t \mathscr{S}_p(u)\mathscr{P}_p^* \mathrm{d}L(u)$, $\star$ is the convolution product of measures and $\mathscr{S}_p(t)\mu_{\mathbf{Z}} := \mu_{\mathbf{Z}} \circ \mathscr{S}_p(t)^{-1}$ is a probability distribution on H, given by

$$\int_H f(\mathbf{x})(\mathscr{S}_p(t)\mu_{\mathbf{Z}})(\mathrm{d}\mathbf{x}) = \int_H f(\mathscr{S}_p(t)^*\mathbf{x})\mu_{\mathbf{Z}}(\mathrm{d}\mathbf{x}),$$

for any bounded measurable function $f : H \to \mathbb{R}$. If $\mathbf{Z}(t) \sim \mu_{\mathbf{Z}}$, then since

$$\log \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle \mathscr{L}_U \mathbf{Z}(t), x\rangle_U}\right] = \log \mathbb{E}\left[\mathrm{e}^{\langle \mathbf{Z}(t), \mathscr{L}_U^* x\rangle}\right] = \widehat{\mu}_{\mathbf{Z}}(\mathscr{L}_U^* x),$$

it follows that $\{X(t)\}_{t\geq 0}$ is stationary with distribution μ_X having characteristic exponent $\widehat{\mu}_X(x) = \widehat{\mu}_{\mathbf{Z}}(\mathscr{L}_U^* x)$ for $x \in U$.

We notice that $\mathscr{C}_p$ is a bounded operator on H if and only if A_i, $i = 1, \ldots, p$ and I_j, $j = 2, \ldots, p$ are bounded operators. In the case of $\mathscr{C}_p$ being bounded, we know from Thm. I.3.14 in [14] that the semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$ is exponentially stable if and only if $\mathrm{Re}(\lambda) < 0$ for all $\lambda \in \sigma(\mathscr{C}_p)$, where $\sigma(\mathscr{C}_p)$ denotes the spectrum of the bounded operator $\mathscr{C}_p$. Recall from Sect. 1 that a real-valued CARMA process admits a limiting distribution if and only if all the eigenvalues of C_p in Eq. 2 have negative real part. In general, by Thm. V.1.11 in [14], the semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$ is exponentially stable if and only if $\{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) > 0\}$ is a subset of the resolvent set $\rho(\mathscr{C}_p)$ of $\mathscr{C}_p$ and $\sup_{\mathrm{Re}(\lambda)>0} \|R(\lambda, \mathscr{C}_p)\| < \infty$. Here, $R(\lambda, \mathscr{C}_p)$ is the resolvent of $\mathscr{C}_p$ for $\lambda \in \rho(\mathscr{C}_p)$.

3.2 Path Regularity

Let us study the regularity of the paths of the CAR(p) process. We have the following proposition:

Proposition 4 *For $p \in \mathbb{N}$ with $p > 1$, assume that $\mathscr{C}_p$ defined in (9) is the generator of a C_0-semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$. Then the H_1-valued CAR(p) process X given in Definition 4 has the representation*

$$X(t) = \mathscr{P}_1 \mathscr{S}_p(t)\mathbf{Z}_0 + \mathscr{P}_1 \mathscr{C}_p \int_0^t \int_0^u \mathscr{S}_p(u-s)\mathscr{P}_p^* \mathrm{d}L(s)\mathrm{d}u,$$

for all $t \geq 0$.

Proof From [14, Ch. II, Lemma 1.3], we have that

$$\mathscr{S}_p(t) = \mathrm{Id} + \mathscr{C}_p \int_0^t \mathscr{S}_p(s)\mathrm{d}s.$$

But for any $\mathbf{x} \in H$, it is simple to see that $\mathscr{P}_1 \mathrm{Id} \mathscr{P}_p^* \mathbf{x} = 0$ when $p > 1$. Therefore it holds

$$\mathscr{P}_1 \mathscr{S}_p(t)\mathscr{P}_p^* = \mathscr{P}_1 \mathscr{C}_p \int_0^t \mathscr{S}_p(s)\mathscr{P}_p^* \mathrm{d}s.$$

The integral on the right-hand side is in Bochner sense as an integral of operators. After appealing to the stochastic Fubini theorem, see [20, Thm. 8.14], it follows from the explicit expression of $X(t)$ in (13)

$$\begin{aligned} X(t) &= \mathscr{P}_1 \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{P}_1 \mathscr{S}_p(t-s)\mathscr{P}_p^* \mathrm{d}L(s) \\ &= \mathscr{P}_1 \mathscr{S}_p(t)\mathbf{Z}_0 + \int_0^t \mathscr{P}_1 \mathscr{C}_p \int_0^{t-s} \mathscr{S}_p(u)\mathscr{P}_p^* \mathrm{d}u\mathrm{d}L(s) \\ &= \mathscr{P}_1 \mathscr{S}_p(t)\mathbf{Z}_0 + \mathscr{P}_1 \int_0^t \mathscr{C}_p \int_s^t \mathscr{S}_p(u-s)\mathscr{P}_p^* \mathrm{d}u\mathrm{d}L(s). \end{aligned}$$

We know from [14, Ch. II, Lemma 1.3] that $\int_s^t \mathscr{S}_p(u-s)\mathscr{P}_p^* \mathrm{d}u \in Dom(\mathscr{C}_p)$. We demonstrate that $\int_0^t \int_s^t \mathscr{S}_p(u-s)\mathscr{P}_p^* \mathrm{d}u\mathrm{d}L(s) \in Dom(\mathscr{C}_p)$: First we recall that $\mathbf{L} = \mathscr{P}_p^* L$ is an H-valued square-integrable Lévy process with mean zero. From

the semigroup property,

$$\begin{aligned}
&\frac{1}{h}\left(\mathscr{S}_p(h)\int_0^t\int_s^t \mathscr{S}_p(u-s)\mathrm{d}u\mathrm{d}\mathbf{L}(s) - \int_0^t\int_s^t \mathscr{S}_p(u-s)\mathrm{d}u\mathrm{d}\mathbf{L}(s)\right)\\
&\quad= \frac{1}{h}\int_0^t\int_s^t \mathscr{S}_p(u+h-s)\mathrm{d}u\mathrm{d}\mathbf{L}(s) - \frac{1}{h}\int_0^t\int_s^t \mathscr{S}_p(u-s)\mathrm{d}u\mathrm{d}\mathbf{L}(s)\\
&\quad= \int_0^t \frac{1}{h}\int_{s+h}^{t+h} \mathscr{S}_p(v-s)\mathrm{d}v - \frac{1}{h}\int_s^t \mathscr{S}_p(v-s)\mathrm{d}s\mathrm{d}\mathbf{L}(s)\\
&\quad= \int_0^t \frac{1}{h}\int_{t}^{t+h} \mathscr{S}_p(v-s)\mathrm{d}v - \frac{1}{h}\int_s^{s+h} \mathscr{S}_p(v-s)\mathrm{d}s\mathrm{d}\mathbf{L}(s)\\
&\quad= \int_0^t \frac{1}{h}\int_0^h \mathscr{S}_p(u)\mathrm{d}u\mathscr{S}_p(t-s) - \frac{1}{h}\int_0^h \mathscr{S}_p(u)\mathrm{d}u\mathrm{d}\mathbf{L}(s)\\
&\quad= \frac{1}{h}\int_0^h \mathscr{S}_p(u)\mathrm{d}u\left(\int_0^t \mathscr{S}_p(t-s)\mathrm{d}\mathbf{L}(s) - \mathbf{L}(t)\right).
\end{aligned}$$

By the fundamental theorem of calculus for Bochner integrals, $(1/h)\int_0^h \mathscr{S}_p(u)\mathrm{d}u \to \mathrm{Id}$ when $h \downarrow 0$. Therefore, the limit exists and the claim follows. From this we find that

$$\begin{aligned}
X(t) &= \mathscr{P}_1\mathscr{S}_p(t)\mathbf{Z}_0 + \mathscr{P}_1\mathscr{C}_p\int_0^t\int_s^t \mathscr{S}_p(u-s)\mathscr{P}_p^*\mathrm{d}u\mathrm{d}L(s)\\
&= \mathscr{P}_1\mathscr{S}_p(t)\mathbf{Z}_0 + \mathscr{P}_1\mathscr{C}_p\int_0^t\int_0^u \mathscr{S}_p(u-s)\mathscr{P}_p^*\mathrm{d}L(s)\mathrm{d}u.
\end{aligned}$$

In the last equality, we applied the stochastic Fubini Theorem (see e.g. [20, Thm. 8.14]). Hence, the result follows.

Note that if $\mathbf{Z}_0 \in Dom(\mathscr{C}_p)$, then by [14, Ch. II, Lemma 1.3] $t \mapsto \mathscr{P}_1\mathscr{S}_p(t)\mathbf{Z}_0$ are differentiable. Assuming that $\int_0^t \mathscr{S}_p(t-s)\mathscr{P}_p^*\mathrm{d}L(s) \in Dom(\mathscr{C}_p)$, it follows from the Proposition above that the paths $t \mapsto X(t), t \geq 0$ of X are absolutely continuous, with weak derivative

$$X'(t) = \mathscr{P}_1\mathscr{C}_p\mathscr{S}_p(t)\mathbf{Z}_0 + \mathscr{P}_1\mathscr{C}_p\int_0^t \mathscr{S}_p(t-s)\mathscr{P}_p^*\mathrm{d}L(s), \tag{26}$$

for $t \geq 0$. The stochastic integral in (26) has RCLL (cadlag) paths when $\{\mathscr{S}_p(t)\}_{t\geq 0}$ is contractive (see [20, Prop. 9.18]), and therefore the H_1-valued CAR(p)-processes for $p > 1$ have weakly differentiable paths being RCLL. If $L = W$, an H_p-valued Wiener process, then the stochastic integral has continuous paths in the case the semigroup is contractive and the paths of X become continuously differentiable. We

point out that $p > 1$ is very different from $p = 1$ in this respect, as the Ornstein-Uhlenbeck process

$$X(t) = \mathscr{S}_1(t)Z_0 + \int_0^t \mathscr{S}_1(t-s)\mathrm{d}L(s)$$
$$= \mathscr{S}_1(t)Z_0 + L(t) + \int_0^t \int_0^u \mathscr{S}_1(u-s)\mathrm{d}L(s)\mathrm{d}u,$$

does not have absolutely continuous paths except in the trivial case when the Lévy process is simply a drift. It is straightforward to define an H_p-valued Lévy process L for which $\int_0^t \mathscr{S}_p(t-s)\mathscr{P}_p^*\mathrm{d}L(s) \in Dom(\mathscr{C}_p)$. For example, let $\widetilde{L}$ be an $\mathbb{R}$-valued square-integrable Lévy process with zero mean, and define $L = \widetilde{L}g$ for $g \in Dom(A_1) \cap Dom(I_2)$. Then $\mathscr{P}_p^* L = (\mathscr{P}_p^* g)\widetilde{L} \in Dom(\mathscr{C}_p)$, and therefore $\int_0^t \mathscr{S}_p(t-s)(\mathscr{P}_p^* g)\mathrm{d}\widetilde{L}(s) \in Dom(\mathscr{C}_p)$ from [14, Ch. II, Lemma 1.3]. If we consider the particular case of the wave equation, as presented at the end of Sect. 1, we have $A_1 = 0$ and $I_2 = \mathrm{Id}$, and thus we can choose any $g \in H_2$. In this case we can conclude that the paths of the solution of the wave equation are absolutely continuous with weak derivative as in (26).

3.3 Semigroup Representation

We study a recursive representation of the C_0-semigroup $\{\mathscr{S}_p(t)\}_{t\geq 0}$ with $\mathscr{C}_p$ as generator, where we recall $\mathscr{C}_p$ from (9). The following result is known as the variation-of-constants formula (see e.g. [20, Appendix B.1.1 and Thm. B.5]) and turns out to be convenient when expressing the semigroup for $\mathscr{C}_p$.

Proposition 5 *Let $\mathscr{A}$ be a linear operator on H being the generator of a C_0-semigroup $\{\mathscr{S}_{\mathscr{A}}(t)\}_{t\geq 0}$. Assume that $\mathscr{B} \in L(H)$. Then the operator $\mathscr{A} + \mathscr{B} : Dom(\mathscr{A}) \to H$ is the generator of the C_0-semigroup $\{\mathscr{S}(t)\}_{t\geq 0}$ defined by*

$$\mathscr{S}(t) = \mathscr{S}_{\mathscr{A}}(t) + \mathscr{R}(t),$$

where

$$\mathscr{R}(t) = \sum_{n=1}^{\infty} \mathscr{R}_n(t),$$

and

$$\mathscr{R}_{n+1}(t) = \int_0^t \mathscr{S}_{\mathscr{A}}(t-s)\mathscr{B}\mathscr{R}_n(s)\mathrm{d}s,$$

for $n = 0, 1, 2, \ldots$, with $\mathscr{R}_0(t) = \mathscr{S}_{\mathscr{A}}(t)$.

We apply the proposition above to give a recursive description of the C_0-semigroup of $\mathscr{C}_p$.

Proposition 6 *Given the operator $\mathscr{C}_p$ defined in (9) for $p \in \mathbb{N}$, where $\mathscr{C}_1 = A_1$ is a densely defined linear operator on H_p (possibly unbounded) with C_0-semigroup $\{\mathscr{S}_1(t)\}_{t\geq 0}$. For $p > 1$, assume that $I_p \in L(H_2, H_1)$, $A_p \in L(H_1, H_p)$ and $\mathscr{C}_{p-1}$ is a densely defined operator on $H_2 \times \ldots \times H_p$ with a C_0-semigroup $\{\mathscr{S}_{p-1}(t)\}_{t\geq 0}$, then*

$$\mathscr{S}_p(t) = \mathscr{S}^+_{p-1}(t) + \sum_{n=1}^{\infty} \mathscr{R}_{n,p}(t),$$

where $\mathscr{R}_{0,p}(t) = \mathscr{S}^+_{p-1}(t)$ and for $n = 1, 2, \ldots,$

$$\mathscr{R}_{n+1,p}(t) = \int_0^t \mathscr{S}^+_{p-1}(t-s)\mathscr{B}_p \mathscr{R}_{n,p}(s)\mathrm{d}s.$$

Here, $\mathscr{B}_p \in L(H)$ is

$$\mathscr{B}_p = \begin{bmatrix} 0 & I_p & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ A_p & 0 & . & \ldots & 0 \end{bmatrix}$$

and $\mathscr{S}^+_{p-1} \in L(H)$

$$\mathscr{S}^+_{p-1} = \begin{bmatrix} Id & 0 & \ldots & 0 \\ 0 & & & \\ . & & & \\ . & & \mathscr{S}_{p-1}(t) & \\ 0 & & & \end{bmatrix}$$

for Id being the identity operator on H_1.

Proof By assumption, $I_p \in L(H_2, H_1)$ and $A_p \in L(H_1, H_p)$, and thus $\mathscr{B}_p \in L(H)$. Define

$$\mathscr{A}_p = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & & & \\ . & & & \\ . & & \mathscr{C}_{p-1} & \\ 0 & & & \end{bmatrix}.$$

Then, $\mathscr{A}_p + \mathscr{B}_p = \mathscr{C}_p$. Moreover, $\{\mathscr{S}^+_{p-1}(t)\}_{t\geq 0}$ is the C_0-semigroup of $\mathscr{A}_p$. Hence, the result follows from Proposition 5.

As an example, consider $p = 3$. Then we have

$$\mathscr{C}_3 = \begin{bmatrix} 0 & I_3 & 0 \\ 0 & 0 & I_2 \\ A_3 & A_2 & A_1 \end{bmatrix}.$$

First, $\mathscr{C}_1 = A_1$ is a (possibly unbounded) operator on H_3, with C_0-semigroup $\{\mathscr{S}_1(t)\}_{t\geq 0} \subset L(H_3)$. Next, let

$$\mathscr{B}_2 = \begin{bmatrix} 0 & I_2 \\ A_2 & 0 \end{bmatrix}$$

where we assume $I_2 \in L(H_3, H_2)$ and $A_2 \in L(H_2, H_3)$ to have $\mathscr{B}_2 \in L(H_2 \times H_3)$. With

$$\mathscr{S}^+_1(t) = \begin{bmatrix} \mathrm{Id} & 0 \\ 0 & \mathscr{S}_1(t) \end{bmatrix},$$

which defines a C_0-semigroup on $L(H_2 \times H_3)$ with generator

$$\mathscr{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & A_1 \end{bmatrix},$$

we obtain

$$\mathscr{S}_2(t) = \mathscr{S}^+_1(t) + \sum_{n=1}^{\infty} \mathscr{R}_{n,2}(t)$$

for $\mathscr{R}_{0,2} = \mathscr{S}^+_1(t)$ and

$$\mathscr{R}_{n+1,2}(t) = \int_0^t \mathscr{S}^+_1(t_s)\mathscr{B}_2\mathscr{R}_{n,2}(s)\mathrm{d}s, n = 1, 2, \ldots.$$

We note that $\{\mathscr{S}_2\}_{t\geq 0} \subset L(H_2 \times H_3)$ is the C_0-semigroup with generator $\mathscr{C}_2$ densely defined on $H_2 \times H_3$. Finally, let

$$\mathscr{B}_3 = \begin{bmatrix} 0 & I_3 & 0 \\ 0 & 0 & 0 \\ A_3 & 0 & 0 \end{bmatrix}$$

which is a bounded operator on H after assuming $I_3 \in L(H_2, H_1)$ and $A_3 \in L(H_1, H_3)$. With

$$\mathscr{S}_2^+(t) = \begin{bmatrix} \mathrm{Id}\ 0 & & 0 \\ 0 & & \\ & \mathscr{S}_2(t) & \\ 0 & & \end{bmatrix},$$

which is a C_0-semigroup on $L(H)$ with generator

$$\mathscr{A}_3 = \begin{bmatrix} 0\ 0 & & 0 \\ 0 & & \\ & \mathscr{C}_2 & \\ 0 & & \end{bmatrix},$$

we conclude with

$$\mathscr{S}_3(t) = \mathscr{S}_2^+(t) + \sum_{n=1}^{\infty} \mathscr{R}_{n,3}(t),$$

where $\mathscr{R}_{0,3}(t) = \mathscr{S}_2^+(t)$ and

$$\mathscr{R}_{n+1,3}(t) = \int_0^t \mathscr{S}_2^+(t-s)\mathscr{B}_3\mathscr{R}_{n,3}(s)\mathrm{d}s.$$

From this example we see that A_2, A_3, I_2 and I_3 must all be bounded operators, while only A_1 is allowed to be unbounded. By recursion in Proposition 6, we see that we must have $I_i \in L(H_{p+2-i}, H_{p+1-i})$ and $A_i \in L(H_{p+1-i}, H_p)$, $i = 2, 3, \ldots, p$, and $A_1 : Dom(A_1) \to H_p$ can be an unbounded operator with densely defined domain $Dom(A_1) \subset H_p$.

We remark that Ch. III in [14] presents a deep theory for perturbations of generators $\mathscr{A}$ by operators $\mathscr{B}$. Matrix operators of the kind $\mathscr{C}_p$ for $p = 2$ has been analysed in, for example [23], where conditions for analyticity of the semigroup $\{\mathscr{S}_2(t)\}_{t\geq 0}$ is studied.

4 Applications of CARMA Processes

In this Section we will look at an Euler discretization of the Hilbert-valued CAR(p) dynamics, and relate this to the functional autoregressive processes studied by [10]. Next, we discuss the wave equation in our context of our analysis, and establish a relationship to ambit fields. In many applications, like in finance or turbulence

say, continuous-time models are often preferred. The infinite dimensional CARMA processes may be an attractive class in this respect. In particular, these processes may provide insight into analysis of data monitored continuously in time, such as traffic flow or weather variables. We remark that this aspect has been mentioned as a future perspective of functional data analysis by [25, Section 6].

Recall Proposition 1, and let $t_i := i \cdot \delta$ for $i = 0, 1, \ldots$ and a given $\delta > 0$. Define further $\mathbf{z}_i := \mathbf{Z}(t_i)$. By the semigroup property of $\{\mathscr{S}_p(t)\}_{t\geq 0}$ it holds,

$$\begin{aligned}\mathbf{z}_{i+1} &= \mathscr{S}_p(t_{i+1})\mathbf{z}_0 + \int_0^{t_{i+1}} \mathscr{S}_p(t_{i+1}-s)\mathscr{P}_p^* \mathrm{d}L(s) \\ &= \mathscr{S}_p(\delta)\mathscr{S}_p(t_i)\mathbf{z}_i + \mathscr{S}_p(\delta)\int_0^{t_i} \mathscr{S}_p(t_i-s)\mathscr{P}_p^* \mathrm{d}L(s) \\ &\quad + \int_{t_i}^{t_{i+1}} \mathscr{S}_p(t_{i+1}-s)\mathscr{P}_p^* \mathrm{d}L(s) \\ &= \mathscr{S}_p(\delta)\mathbf{z}_i + \boldsymbol{\epsilon}_i,\end{aligned}$$

with

$$\boldsymbol{\epsilon}_i := \int_{t_i}^{t_{i+1}} \mathscr{S}_p(t_{i+1}-s)\mathscr{P}_p^* \mathrm{d}L(s).$$

The process above has the form of a discrete-time AR(1) process. Obviously, $\mathscr{S}_p(\delta) \in L(H)$ and by the independent increment property of the H_p-valued Lévy process L, $\{\boldsymbol{\epsilon}_i\}_{i=0}^{\infty}$ is a sequence of independent H-valued random variables. Furthermore, $\mathbb{E}[\boldsymbol{\epsilon}_i] = 0$ due to the zero-mean hypothesis of L. Finally, we can compute the covariance operators of $\boldsymbol{\epsilon}_i$ by appealing to the Itô isometry (cf. [20, Cor. 8.17])

$$\begin{aligned}\mathbb{E}[\langle \boldsymbol{\epsilon}_i, \mathbf{x}\rangle\langle \boldsymbol{\epsilon}_i, \mathbf{y}\rangle] &= \int_{t_i}^{t_{i+1}} \langle Q\mathscr{P}_p\mathscr{S}_p^*(t_{i+1}-s)\mathscr{P}_1^*\mathbf{x}, \mathscr{P}_p\mathscr{S}_p^*(t_{i+1}-s)\mathscr{P}_1^*\mathbf{y}\rangle \mathrm{d}s \\ &= \int_0^{\delta} \langle \mathscr{P}_1\mathscr{S}_p(s)\mathscr{P}_p^* Q\mathscr{P}_p\mathscr{S}_p^*(s)\mathscr{P}_1^*\mathbf{x}, \mathbf{y}\rangle \mathrm{d}s,\end{aligned}$$

where $\mathbf{x}, \mathbf{y} \in H$. Thus, $\boldsymbol{\epsilon}_i$ has covariance operator $\mathscr{Q}_{\boldsymbol{\epsilon}}$ independent of i given by

$$\mathscr{Q}_{\boldsymbol{\epsilon}} = \int_0^{\delta} \mathscr{P}_1\mathscr{S}_p(s)\mathscr{P}_p^* Q\mathscr{P}_p\mathscr{S}_p^*(s)\mathscr{P}_1^* \mathrm{d}s.$$

Therefore, $\{\boldsymbol{\epsilon}_i\}_{i=0}^{\infty}$ is an *iid* sequence of H-valued random variables. Hence, the H-valued time series $\{\mathbf{z}_i\}_{i=0}^{\infty}$ is a so-called *linear process* according to [10].

Let us now focus on the H_1-valued CAR(p) dynamics in Definition 4, and see how this process can be related to a times series in H_1. To this end, recall the

operator-valued polynomial $Q_p(\lambda)$ introduced in (17) and the formal pth-order stochastic differential equation in (18). Let Δ_δ be the forward differencing operator with time step $\delta > 0$. Moreover, we assume Δ_δ^n to be the nth order forward differencing, defined as

$$\Delta_\delta^n f(t) = \sum_{k=0}^{n} \binom{n}{k} (-1)^k f(t + (n-k)\delta)$$

for a function f and $n \in \mathbb{N}$. Obviously, $\Delta_\delta^1 = \Delta_\delta$. Introduce the discrete time grid $t_i := i\delta$, $i = 0, 1, 2, \ldots$, and observe that

$$\frac{1}{\delta}\Delta_\delta I_p \cdots I_2 L(t_i) = \frac{1}{\delta} I_p \cdots I_2 (L(t_{i+1}) - L(t_i)).$$

Assuming that the increments of L belongs to the domain of $I_p \cdots I_2$, we find that

$$\epsilon_i := \frac{1}{\delta} I_p \cdots I_2 (L(t_{i+1}) - L(t_i)) \tag{27}$$

for $i = 0, 1, 2, \ldots$ define an *iid* sequence of H_1-valued random variables. We remark that this follows from the stationarity hypothesis of a Lévy process saying that the increments $L(t_{i+1}) - L(t_i)$ are distributed as $L(\delta)$. The random variables $\epsilon_i, i = 0, 1, \ldots,$ will be the numerical approximation of the formal expression $I_p \cdots I_2 \dot{L}(t_i)$. Finally, we define (formally) a time series $\{x_i\}_{i=0}^\infty$ in H_1 by

$$Q_p\left(\frac{\Delta_\delta}{\delta}\right) x_i = \epsilon_i. \tag{28}$$

In this definition, we use the notation $x_i = x(t_i)$ when applying the forward differencing operator Δ_δ. The polynomial Q_p involves the linear operators $B_1, \ldots, B_p$ that may not be everywhere defined. We define the domain $Dom(B) \subset H_1$ by

$$Dom(B) := Dom(B_1) \cap \cdots \cap Dom(B_p), \tag{29}$$

which we assume to be non-empty. This will form the natural domain for the time series $\{x_i\}_{i=0}^\infty$.

Proposition 7 *Assume that for any $y_1, \ldots, y_p \in Dom(B)$, $B_1 y_1 + \cdots B_p y_p \in Dom(B)$. If $\{\epsilon_i\}_{i=0}^\infty \subset Dom(B)$ with ϵ_i defined in (27) and $x_0, \ldots, x_{p-1} \in Dom(B)$, then $\{x_i\}_{i=0}^\infty$ is an AR(p) process in H_1 with dynamics*

$$x_{i+p} = \sum_{q=1}^{p} \widetilde{B}_q x_{i+(p-q)} + \delta^p \epsilon_i$$

where

$$\widetilde{B}_q = (-1)^{q+1}\binom{p}{q} Id + \sum_{k=1}^{q} \delta^k B_k (-1)^{q-k} \binom{p-k}{q-k}, q = 1, \ldots, p,$$

and Id is the identity operator on H_1.

Proof First we observe that the assumption $B_1 y_1 + \cdots + B_p y_p \in Dom(B)$ for any $y_1, \ldots, y_p \in Dom(B)$ is equivalent with $\widetilde{B}_1 y_1 + \cdots + \widetilde{B}_p y_p \in Dom(B)$ for any $y_1, \ldots, y_p \in Dom(B)$ since $\widetilde{B}_q$ is a linear combination of $B_1, \ldots, B_q$. Thus, by the assumptions, we see that $x_i \in Dom(B)$ for all $i = 0, 1, 2 \ldots$ and the recursion for the time series dynamics is well-defined.

We next show that the time series $\{x_i\}_{i=0}^{\infty}$ is indeed given by the recursion in the Proposition. From the definition of Q_p and the forward differencing operators, we find after isolating x_{i+p} on the left hand side and the remaining terms on the right hand side in the definition in Eq. (28) that

$$x_{i+p} = -\sum_{q}^{p} (-1)^q \binom{p}{q} x_{i+(p-q)} + \sum_{q=1}^{p-1} \delta^q B_q \left(\sum_{k=0}^{p-q} (-1)^k \binom{p-q}{k} x_{i+(p-q-k)} \right) + \delta^p B_p x_i + \delta^p \epsilon_i.$$

Identifying terms for $x_{i+(p-1)}, x_{i+(p-2)}, \ldots, x_i$ yields the result.

The time series $\{x_i\}_{i=0}^{\infty}$ defined in (28) can be viewed as the numerical approximation of the H_1-valued CAR(p) process $X(t)$. Notice that for small δ we find that $\mathscr{S}_p(\delta) \approx \delta \mathscr{C}_p + \mathrm{Id}$. Using this approximation in the explicit representation of $\mathbf{Z}(t)$ in Proposition 1 will yield the same conclusion as in our discussion above.

We remark that if the operators $B_1, \ldots, B_p$ are bounded, then $Dom(B) = H_1$. In this case, the time series $\{x_i\}_{i=0}^{\infty}$ will be everywhere defined on H_1, and we do not need to impose any additional "domain preservation" hypothesis.

Let us consider an example where $p = 3$, and $H_1 = H_2 = H_3$. Suppose that $I_i = \mathrm{Id}$ for $i = 1, 2, 3$ and recall from the discussion in Sect. 2 that in this case $B_q = A_q$ for $q = 1, 2, 3$. Using Proposition 7 yields that

$$x_{i+3} = (3\mathrm{Id} + A_1)x_{i+2} + (A_2 - 2A_1 - 3\mathrm{Id})x_{i+1} + (\mathrm{Id} + A_1 - A_2 + A_3)x_i + \epsilon_i$$

when $\delta = 1$. Here, $\epsilon_i = L(t_{i+1}) - L(t_i)$ and thus being distributed as $L(1)$. This formula is the analogy of Ex. 10.2 in [9]. Indeed, Proposition 7 is the generalization of [8, Eq. (4.17)] to Hilbert space.

The H_1-valued AR(p)-process in Proposition 7 is called a *functional autoregressive* process of order p (or, in short-hand notation, FAR(p)-process) by [10]. For example, [17] apply such models in a functional data analysis of Eurodollar futures, where they find statistical evidence for a FAR(2) dynamics. We remark that [17] defines FAR(p) processes using the observer canonical form rather that the

controller canonical form as we use. At this point, we would also like to mention that the stochastic wave equation considered in Sect. 1 will be an AR(2) process with values in H_1 (or a FAR(2)-process). Indeed, since in this case $A_1 = 0$, $I_2 = \mathrm{Id}$ and $A_2 = \Delta$, the Laplacian, we find that $B_1 = 0$ and $B_2 = \Delta$, and hence,

$$x_{i+2} = 2\mathrm{Id}x_{i+1} - (\mathrm{Id} - \delta^2\Delta)x_i + \delta^2\epsilon_i,$$

for $i = 0, 1, 2 \ldots$ Obviously, this recursion is obtained by approximating the wave equation by the discrete second derivative in time.

Recalling from (22) the semigroup $\{\mathscr{S}_2(t)\}_{t\geq 0}$ of the wave equation, we see from (13) that it has the representation (with initial condition $\mathbf{Z}_0 = \mathbf{0}$)

$$X(t) = \int_0^t (-\Delta)^{-1/2} \sin((-\Delta)^{1/2}(t-s))\mathrm{d}L(s).$$

Following the analysis in [6], X will be a Hilbert-valued ambit field. Ambit fields have attracted a great deal of attention as random fields in time and space suitable for modelling turbulence, as we recall from the definition and discussion in Sect. 1. As L is a $L^2(0, 1)$-valued Lévy process, one can represent it in terms of the basis $\{e_n\}_{n=1}^{\infty}$, where $e_n(x) = \sqrt{2}\sin(\pi n x)$, as

$$L(t, x) = \sum_{n=1}^{\infty} \ell_n(t)e_n(x),$$

with $\ell_n(t) := \langle L(t, \cdot), e_n\rangle_2, n = 1, 2, \ldots$ being real-valued square-integrable Lévy processes with zero mean (see [20, Sect. 4.8]). Thus, the stochastic wave equation has the representation

$$X(t, x) = \sum_{n=1}^{\infty} \frac{\sqrt{2}}{\pi n} \int_0^t \sin(\pi n(t-s))\mathrm{d}\ell_n(s) \sin(\pi n x).$$

But with $\widetilde{L}(\mathrm{d}y, \mathrm{d}s) := \langle L(\mathrm{d}s), e_n\rangle_2 \mathrm{d}y$, we obtain an expression for $X(t, x)$ similar to the definition of an ambit field from Sect. 1. Note that $\widetilde{L}$ is not necessarily a Lévy basis in this context. Hilbert-valued CARMA$(p, U, \mathscr{L}_U)$-processes provide us with a rich class of ambit fields, as real-valued CARMA processes are specific cases of Lévy semistationary processes (see e.g. [5, 8, 9]).

Acknowledgements Financial support from the project FINEWSTOCH, funded by the Norwegian Research Council, is gratefully acknowledged. Two anonymous referees are thanked for their positive and constructive critics.

References

1. Andresen, A., Benth, F.E., Koekebakker, S., Zakamouline, V.: The CARMA interest rate model. Int. J. Theor. Appl. Finance **17**, 1–27 (2014)
2. Applebaum, D.: Lévy Processes and Stochastic Calculus. Cambridge University Press, Cambridge (2009)
3. Applebaum, D.: Infinite dimensional Ornstein-Uhlenbeck processes driven by Lévy processes. Probab. Surv. **12**, 33–54 (2015)
4. Barndorff-Nielsen, O.E., Schmiegel, J.: Ambit processes; with applications to turbulence and tumour growth. In: Benth, F.E., Di Nunno, G., Lindstrøm, T., Øksendal, B., Zhang, T.-S. (eds.) Stochastic Analysis and Applications, pp. 93–124. Springer, Heidelberg (2007)
5. Barndorff-Nielsen, O.E., Benth, F.E., Veraart, A.: Modelling energy spot prices by volatility modulated Lévy-driven Volterra processes. Bernoulli **19**, 803–845 (2013)
6. Benth, F.E., Eyjolfsson, H.: Representation and approximation of ambit fields in Hilbert space. Stochastics **89**, 311–347 (2017)
7. Benth, F.E., Müller, G., Klüppelberg, C., Vos, L.: Futures pricing in electricity markets based on stable CARMA spot models. Energy Econ. **44**, 392–406 (2014)
8. Benth, F.E., Šaltytė Benth, J.: Modeling and Pricing in Financial Markets for Weather Derivatives. World Scientific, Singapore (2013)
9. Benth, F.E., Šaltytė Benth, J., Koekebakker, S.: Stochastic Modelling of Electricity and Related Markets. World Scientific, Singapore (2008)
10. Bosq, D.: Linear Processes in Function Spaces. Springer, New York (2000)
11. Brockwell, P.J.: Lévy-driven CARMA processes. Ann. Inst. Statist. Math. **53**, 113–124 (2001)
12. Brockwell, P.J., Schlemm, E.: Parametric estimation of the driving Lévy process of multivariate CARMA processes from discrete observations. J. Multivariate Anal. **115**, 217–251 (2013)
13. Carmona, R., Tehranchi, M.: Interest Rate Models: An Infinite Dimensional Stochastic Analysis Perspective. Springer, Berlin/Heidelberg (2006)
14. Engel, K.-J., Nagel, R.: One-Parameter Semigroups for Linear Evolution Equations. Springer, New York (2000)
15. Härdle, W.K., Lopez Cabrera, B.: The implied market price of weather risk. Appl. Math. Finance **19**, 59–95 (2012)
16. Kevei, P.: High-frequency sampling of multivariate CARMA processes (2015). Available at arXiv:1509.03485v1
17. Kokoszka, P., Reimherr, M.: Determining the order of the functional autoregressive model. J. Time Series Anal. **34**, 116–129 (2013)
18. Marquardt, T., Stelzer, R.: Multivariate CARMA processes. Stoch. Process. Appl. **117**, 96–120 (2007)
19. Paschke, R., Prokopczuk, M.: Commodity derivatives valuation with autoregressive and moving average components in the price dynamics. J. Bank. Financ. **34**, 2742–2752 (2010)
20. Peszat, S., Zabczyk, J.: Stochastic Partial Differential Equations with Lévy Noise. Cambridge University Press, Cambridge (2007)
21. Sato, K.-I.: Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge (1999)
22. Schlemm, E., Stelzer, R.: Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes. Bernoulli **18**, 46–63 (2012)
23. Trunk, C.: Analyticity of semigroups related to a class of block operator matrices. Oper. Theory Adv. Appl. **195**, 257–271 (2009)
24. Walsh, J.: An introduction to stochastic partial differential equations. In: Carmona, R., Kesten, H., Walsh, J. (eds.) Lecture Notes in Mathematics, vol. 1180. Ecole dete de Probabilites de Saint-Flour XIV. Springer, Heidelberg (1984)
25. Wang, J.-L., Chiou, J.-M., Müller, H.-G.: Functional Data Analysis. Ann. Rev. Stat. Appl. **3**, 257–295 (2016)

Pre- and Post-Lie Algebras: The Algebro-Geometric View

Gunnar Fløystad and Hans Munthe-Kaas

Abstract We relate composition and substitution in pre- and post-Lie algebras to algebraic geometry. The Connes-Kreimer Hopf algebras and MKW Hopf algebras are then coordinate rings of the infinite-dimensional affine varieties consisting of series of trees, resp. Lie series of ordered trees. Furthermore we describe the Hopf algebras which are coordinate rings of the automorphism groups of these varieties, which govern the substitution law in pre- and post-Lie algebras.

1 Introduction

Pre-Lie algebras were first introduced in two different papers from 1963. Murray Gerstenhaber [13] studies deformations of algebras and Ernest Vinberg [29] problems in differential geometry. The same year John Butcher [2] published the first in a series of papers studying algebraic structures of numerical integration, culminating in his seminal paper [3] where B-series, the convolution product and the antipode of the Butcher–Connes–Kreimer Hopf algebra are introduced.

Post-Lie algebras are generalisations of pre-Lie algebras introduced in the last decade. Bruno Vallette [28] introduced the post-Lie operad as the Koszul dual of the commutative trialgebra operad. Simultaneously post-Lie algebras appear in the study of numerical integration on Lie groups and manifolds [21, 25]. In a differential geometric picture a pre-Lie algebra is the algebraic structure of the flat and torsion free connection on a locally Euclidean space, whereas post-Lie algebras appear naturally as the algebraic structure of the flat, constant torsion connection given by the Maurer–Cartan form on a Lie group [24]. Recently it is shown that the sections of an anchored vector bundle admits a post-Lie structure if and only if the bundle is an action Lie algebroid [22].

G. Fløystad (✉) · H. Munthe-Kaas
Matematisk Institutt, Realfagbygget, Bergen, Norway
e-mail: Gunnar.Floystad@uib.no; Hans.Munthe-Kaas@uib.no

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_12

B-series is a fundamental tool in the study of flow-maps (e.g. numerical integration) on Euclidean spaces. The generalised Lie-Butcher LB-series are combining B-series with Lie series and have been introduced for studying integration on Lie groups and manifolds.

In this paper we study B-series and LB-series from an algebraic geometry point of view. The space of B-series and LB-series can be defined as completions of the free pre- and post-Lie algebras. We study (L)B-series as an algebraic variety, where the coordinate ring has a natural Hopf algebra structure. In particular we are interested in the so-called substitution law. Substitutions for pre-Lie algebras were first introduced in numerical analysis [6]. The algebraic structure of pre-Lie substitutions and the underlying substitution Hopf algebra were introduced in [4]. For the post-Lie case, recursive formulae for substitution were given in [18]. However, the corresponding Hopf algebra of substitution for post-Lie algebras was not understood at that time.

In the present work we show that the algebraic geometry view gives a natural way to understand both the Hopf algebra of composition and the Hopf algebra of substitution for pre- and post-Lie algebras.

The paper is organised as follows. In Part 1 we study fundamental algebraic properties of the enveloping algebra of Lie-, pre-Lie and post-Lie algebras for the general setting that these algebras A are endowed with a decreasing filtration $A = A^1 \supseteq A^2 \supseteq \cdots$. This seems to be the general setting where we can define the exponential and logarithm maps, and define the (generalised) Butcher product for pre- and post-Lie algebras. Part 2 elaborates an algebraic geometric setting, where the pre- or post-Lie algebra forms an algebraic variety and the corresponding coordinate ring acquires the structure of a Hopf algebra. This yields the Hopf algebra of substitutions in the free post-Lie algebra. Finally, we provide a recursive formula for the coproduct in this substitution Hopf algebra.

Part 1: The Non-algebro Geometric Setting

In this part we have no type of finiteness condition on the Lie algebras, and pre- and post-Lie algebras. Especially in the first Sect. 2 the material will be largely familiar to the established reader.

2 The Exponential and Logarithm Maps for Lie Algebras

We work in the most general setting where we can define the exponential and logarithm maps. In Sect. 2.2 we assume the Lie algebra comes with a decreasing filtration, and is complete with respect to this filtration. We define the completed enveloping algebra, and discuss its properties. This is the natural general setting for the exponential and logarithm maps which we recall in Sect. 2.3.

2.1 The Euler Idempotent

The setting in this subsection is any Lie algebra L, finite or infinite dimensional over a field k of characteristic zero. Let $U(L)$ be its enveloping algebra. This is a Hopf algebra with unit η, counit ϵ and coproduct

$$\Delta : U(L) \to U(L) \otimes_k U(L)$$

defined by $\Delta(\ell) = 1 \otimes \ell + \ell \otimes 1$ for any $\ell \in L$, and extended to all of $U(L)$ by requiring Δ to be an algebra homomorphism.

For any algebra A with multiplication map $\mu_A : A \otimes A \to A$, we have the convolution product $\star$ on $\mathrm{Hom}_k(U(L), A)$. For $f, g \in \mathrm{Hom}_k(U(L), A)$ it is defined as

$$f \star g = \mu_A \circ (f \otimes g) \circ \Delta_{U(L)}.$$

Let $\mathbf{1}$ be the identity map on $U(L)$, and $J = \mathbf{1} - \eta \circ \epsilon$. The *Eulerian idempotent* $e : U(L) \to U(L)$ is defined by

$$e = \log^\star(\mathbf{1}) = \log^\star(\eta \circ \epsilon + J) = J - \frac{J^{\star 2}}{2} + \frac{J^{\star 3}}{3} - \cdots .$$

Proposition 2.1 *The image of $e : U(L) \to U(L)$ is $L \subseteq U(L)$, and e is the identity restricted to L.*

Proof This is a special case of the canonical decomposition stated in 0.4.3 in [27]. See also Proposition 3.7, and part (i) of its proof in [27]. □

Let $\mathrm{Sym}^c(L)$ be the free cocommutative conilpotent coalgebra on L. It is the subcoalgebra of the tensor coalgebra $T^c(L)$ consisting of the symmetrized tensors

$$\sum_{\sigma \in S_n} l_{\sigma(1)} \otimes l_{\sigma(2)} \otimes \cdots l_{\sigma(n)} \in L^{\otimes n}, \quad l_1, \ldots l_n \in L. \tag{1}$$

The above proposition gives a linear map $U(L) \xrightarrow{e} L$. Since $U(L)$ is a cocommutative coalgebra, there is then a homomorphism of cocommutative coalgebras

$$U(L) \xrightarrow{\alpha} \mathrm{Sym}^c(L). \tag{2}$$

We now have the following strong version of the Poincaré-Birkhoff-Witt theorem.

Proposition 2.2 *The map $U(L) \xrightarrow{\alpha} Sym^c(L)$ is an isomorphism of coalgebras.*

In order to show this we expand more on the Euler idempotent.

Again for $l_1, \ldots, l_n \in L$ denote by $(l_1, \ldots, l_n)$ the symmetrized product in $U(L)$:

$$\frac{1}{n!} \sum_{\sigma \in S_n} l_{\sigma(1)} l_{\sigma(2)} \cdots l_{\sigma(n)}, \tag{3}$$

and let $U_n(L) \subseteq U(L)$ be the subspace generated by all these symmetrized products.

Proposition 2.3 *Consider the map given by convolution of the Eulerian idempotent:*

$$\frac{e^{\star p}}{p!} : U(L) \to U(L).$$

a. *The map above is zero on $U_q(L)$ when $q \neq p$ and the identity on $U_p(L)$.*
b. *The sum of these maps*

$$\exp^{\star p}(e) = \eta \circ \epsilon + e + \frac{e^{\star 2}}{2} + \frac{e^{\star 3}}{3!} + \cdots$$

is the identity map on $U(L)$. (Note that the map is well defined since the maps $e^{\star p}/p!$ vanish on any element in $U(L)$ for p sufficiently large.)

From the above we get a decomposition

$$U(L) = \bigoplus_{n \geq 0} U_n(L).$$

Proof This is the canonical decomposition stated in 0.4.3 in [27], see also Proposition 3.7 and its proof in [27]. □

Proof of Proposition 2.2 Note that since e vanishes on $U_n(L)$ for $n \geq 2$, by the way one constructs the map α, it sends the symmetrizer $(l_1, \ldots, l_n) \in U_n(L)$ to the symmetrizer (3) in $\mathrm{Sym}_n^c(L)$. This shows α is surjective. But there is also a linear map, the surjective section $\beta : \mathrm{Sym}_n^c(L) \to U_n(L)$ sending the symmetrizer (3) to the symmetric product $(l_1, \ldots, l_n)$. This shows that α must also be injective. □

2.2 Filtered Lie Algebras

Now the setting is that the Lie algebra L comes with a filtration

$$L = L^1 \supseteq L^2 \supseteq L^3 \supseteq \cdots$$

such that $[L^i, L^j] \subseteq L^{i+j}$. Examples of such may be derived from any Lie algebra over k:

1. The lower central series gives such a filtration with $L^2 = [L, L]$ and $L^{p+1} = [L^p, L]$.
2. The polynomials $L[h] = \oplus_{n\geq 1} Lh^n$.
3. The power series $L[\![h]\!] = \Pi_{n\geq 1} Lh^n$.

Let $\mathrm{Sym}_n(L)$ be the symmetric product of L, that is the natural quotient of $L^{\otimes n}$ which is the coinvariants $(L^{\otimes n})^{S_n}$ for the action of the symmetric group S_n. By the definition of $\mathrm{Sym}^c(L)$ in (1) there are maps

$$\mathrm{Sym}_n^c(L) \hookrightarrow L^{\otimes n} \to \mathrm{Sym}_n(L),$$

and the composition is a linear isomorphism. We get a filtration on $\mathrm{Sym}_n(L)$ by letting

$$F^p(\mathrm{Sym}_n(L)) = \sum_{i_1+\cdots+i_n\geq p} L^{i_1} \cdots L^{i_n}.$$

The filtration on L gives an associated graded Lie algebra $\mathrm{gr}\, L = \oplus_{i\geq 1} L_i/L_{i+1}$. The filtration on $\mathrm{Sym}_n(L)$ also induces an associated graded vector space.

Lemma 2.4 *There is an isomorphism of associated graded vector spaces*

$$Sym_n(gr\, L) \xrightarrow{\cong} gr\, Sym_n(L). \tag{4}$$

Proof Note first that there is a natural map (where d denotes the grading induced by the graded Lie algebra $\mathrm{gr}\, L$)

$$\mathrm{Sym}_n(\mathrm{gr}\, L)_d \to F^d \mathrm{Sym}_n(L)/F^{d+1}\mathrm{Sym}_n(L). \tag{5}$$

It is also clear by how the filtration is defined that any element on the right may be lifted to some element on the left, and so this map is surjective. We must then show that it is injective.

Choose splittings $L/L^{i+1} \xrightarrow{s_i} L$ of $L \to L/L^{i+1}$ for $i = 1, \ldots p$, and let $L_i = s_i(L^i/L^{i+1})$. Then we have a direct sum decomposition

$$L = L_1 \oplus L_2 \oplus \cdots \oplus L_p \oplus \cdots .$$

This gives an isomorphism $L \xrightarrow{\cong} \mathrm{gr}\, L$ which again gives a graded isomorphism

$$\mathrm{Sym}_n(L) \xrightarrow{\cong} \mathrm{Sym}_n(\mathrm{gr}\, L). \tag{6}$$

Since in general $\mathrm{Sym}_n(A \oplus B)$ is equal to $\oplus_i \mathrm{Sym}_i(A) \otimes \mathrm{Sym}_{n-i}(B)$ we get that

$$\mathrm{Sym}_n(L) = \oplus_{i_1,\dots,i_p} S_{i_1}(L_1) \otimes \cdots \otimes S_{i_p}(L_p), \tag{7}$$

where we sum over all compositions where $\sum i_j = n$.

Claim

$$F^d S_n(L) = \oplus_{i_1,\dots,i_p} S_{i_1}(L_1) \otimes \cdots \otimes S_{i_p}(L_p),$$

where we sum over all $\sum i_j = n$ and $\sum j \cdot i_j \geq d$.

This shows that the composition of (6) and (5) is an isomorphism. Therefore the map in (5) is an isomorphism.

Proof of Claim. Clearly we have an inclusion $\supseteq$. Conversely let $a \in F^d \mathrm{Sym}_n(L)$. Then a is a sum of products $a_{r_1} \cdots a_{r_q}$ where $a_{r_j} \in L^{r_j}$ and $\sum r_j \geq d$. But then each $a_{r_j} \in \oplus_{t \geq r_j} L_t$, and so by the direct sum decomposition in (7), each $a_{r_1} \cdots a_{r_q}$ lives in the right side of the claimed equality, and so does a. □

We have the enveloping algebra $U(L)$ and the enveloping algebra of the associated graded algebra $U(\mathrm{gr}\, L)$. The augmentation ideal $U(L)_+$ is the kernel $\ker U(L) \xrightarrow{\epsilon} k$ of the counit. The enveloping algebra $U(L)$ now gets a filtration of ideals by letting $F^1 = U(L)_+$ and

$$F^{p+1} = F^p \cdot U(L)_+ + (L^{p+1}),$$

where (L^{p+1}) is the ideal generated by L^{p+1}. This filtration induces again a graded algebra

$$\mathrm{gr}\, U(L) = \bigoplus_i F^i / F^{i+1}.$$

There is also another version, the graded product algebra, which we will encounter later

$$\mathrm{gr}^{\Pi} U(L) = \prod_i F^i / F^{i+1}.$$

Proposition 2.5 *The natural map of graded algebras*

$$U(gr\, L) \xrightarrow{\cong} gr\, U(L),$$

is an isomorphism.

Proof The filtrations on each $\mathrm{Sym}^c_n(L)$ induces a filtration on $\mathrm{Sym}^c(L)$. Via the isomorphism α of (2) and the explicit form given in the proof of Proposition 2.2 the

filtrations on $U(L)$ and on $\mathrm{Sym}_n^c(L)$ correspond. Hence

$$\mathrm{gr}\,\alpha : \mathrm{gr}\,U(L) \xrightarrow{\cong} \mathrm{gr}\,\mathrm{Sym}_n^c(L)$$

is an isomorphism of vector spaces. There is also an isomorphism β and a commutative diagram

$$\begin{array}{ccc} U(\mathrm{gr}\,L) & \xrightarrow{\beta} & \mathrm{Sym}^c(\mathrm{gr}\,L) \\ \downarrow & & \downarrow \\ \mathrm{gr}\,U(L) & \xrightarrow{\mathrm{gr}\,\alpha} & \mathrm{gr}\,\mathrm{Sym}^c(L). \end{array}$$

By Lemma 2.4 the right vertical map is an isomorphism and so also the left vertical map. □

The cofiltration

$$\cdots \twoheadrightarrow U(L)/F^n \twoheadrightarrow U(L)/F^{n-1} \twoheadrightarrow \cdots$$

induces the completion

$$\hat{U}(L) = \varprojlim_p U(L)/F^p.$$

This algebra also comes with the filtration $\hat{F}^p$. Let $\hat{L} = \varprojlim_p L/L^p$.

Lemma 2.6 *The completed algebras are equal:*

$$\hat{U}(\hat{L}) = \hat{U}(L),$$

and so this algebra only depends on the completion $\hat{L}$.

Proof The natural map $L \to \hat{L}$ induces a natural map $U(L) \xrightarrow{\gamma} U(\hat{L})$. Since L and $\hat{L}$ have the same associated graded Lie algebras, the two downward maps in the commutative diagram

$$\begin{array}{ccc} \mathrm{gr}\,U(L) & \longrightarrow & \mathrm{gr}\,U(\hat{L}) \\ & \searrow \quad \swarrow & \\ & U(\mathrm{gr}\,L) & \end{array}$$

are isomorphisms, showing that the upper horizontal map is an isomorphism. But given the natural map γ this easily implies that the map of quotients

$$U(L)/F^{p+1}U(L) \xrightarrow{\gamma^p} U(\hat{L})/F^{p+1}U(\hat{L})$$

is an isomorphism, and so the completions are isomorphic. □

We denote the d'th graded part of the enveloping algebra $U(\mathrm{gr}\, L)$ by $U(\mathrm{gr}\, L)_d$. The following gives an idea of the "size" of $\hat{U}(L)$.

Lemma 2.7

$$gr^{\Pi}\hat{U}(L) = \hat{U}(gr\, L) = \prod_{d\in\mathbb{Z}} U(gr\, L)_d.$$

Proof The left graded product is

$$\mathrm{gr}^{\Pi}\hat{U}(L) = \prod_{p\geq 0} F^p/F^{p+1}.$$

But by Proposition 2.5 $F^p/F^{p+1} \cong U(\mathrm{gr}\, L)_p$ and so the above statement follows. □

Example 2.8 Let $V = \oplus_{i\geq 1} V_i$ be a graded vector space with V_i of degree i, and let $\mathrm{Lie}(V)$ be the free Lie algebra on V. It then has a grading $\mathrm{Lie}(V) = \oplus_{d\geq 1}\mathrm{Lie}(V)_d$ coming from the grading on V, and so a filtration $F^p = \oplus_{d\geq p}\mathrm{Lie}(V)_d$. The enveloping algebra $U(\mathrm{Lie}(V))$ is the tensor algebra $T(V)$. The completed enveloping algebra is

$$\hat{U}(\mathrm{Lie}(V)) = \hat{T}(V) := \prod_d T(V)_d.$$

Let L_p be the quotient L/L^{p+1}, which is a nilpotent filtered Lie algebra. We get enveloping algebras $U(L_p)$ with filtrations $F^j U(L_p)$ of ideals, and quotient algebras

$$U^j(L_p) = U(L_p)/F^{j+1}U(L_p).$$

Lemma 2.9

$$\hat{U}(L) = \varprojlim_{j,p} U^j(L_p).$$

Proof First note that if $j \leq p$ then $U^p(L_p) \twoheadrightarrow U^j(L_p)$ surjects. If $j \geq p$, then $U^j(L_j) \twoheadrightarrow U^j(L_p)$ surjects. Hence it is enough to show that the natural map

$$U(L)/F^{p+1} \to U(L_p)/F^{p+1}U(L_p) = U^p(L_p)$$

is an isomorphism. This follows since we have an isomorphism of associated graded vector spaces:

$$\begin{aligned}(\mathrm{gr}\,(U(L)/F^{p+1}))_{\leq p} &= (\mathrm{gr}\,U(L))_{\leq p} \cong U(\mathrm{gr}\,L)_{\leq p} \\ &= U(\mathrm{gr}\,L_p)_{\leq p} \cong (\mathrm{gr}\,U(L_p))_{\leq p} \\ &= (\mathrm{gr}\,U(L_p)/F^{p+1})_{\leq p}\end{aligned}$$

□

2.3 The Exponential and Logarithm

The coproduct Δ on $U(L)$ will send

$$F^p \xrightarrow{\Delta} 1 \otimes F^p + F^1 \otimes F^{p-1} + \cdots + F^p \otimes 1.$$

Thus we get a map

$$\hat{U}(L) \to U(L)/F^{2p-1} \xrightarrow{\Delta} U(L)/F^p \otimes U(L)/F^p.$$

Let

$$\hat{U}(L)\hat{\otimes}\hat{U}(L) := \varprojlim_p U(L)/F^p \otimes U(L)/F^p$$

be the completed tensor product We then get a *completed coproduct*

$$\hat{U}(L) \xrightarrow{\Delta} \hat{U}(L)\hat{\otimes}\hat{U}(L).$$

Note that the tensor product

$$\hat{U}(L) \otimes \hat{U}(L) \subseteq \hat{U}(L)\hat{\otimes}\hat{U}(L).$$

An element g of $\hat{U}(L)$ is *grouplike* if $\Delta(g) = g \otimes g$ in $\hat{U}(L) \otimes \hat{U}(L)$. We denote the set of grouplike elements by $G(\hat{U}(L))$. They are all of the form $1 + s$ where s is in the augmentation ideal

$$\hat{U}(L)_+ = \ker(\hat{U}(L) \xrightarrow{\epsilon} k).$$

The exponential map

$$\hat{U}(L)_+ \xrightarrow{exp} 1 + \hat{U}(L)_+$$

is given by

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

The logarithm map

$$1 + \hat{U}(L)_+ \xrightarrow{log} \hat{U}(L)_+$$

is defined by

$$\log(1 + s) = s - \frac{s^2}{2} + \frac{s^3}{3} - \cdots .$$

Proposition 2.10 *The maps*

$$\hat{U}(L)_+ \underset{\log}{\overset{\exp}{\rightleftarrows}} 1 + \hat{U}(L)_+$$

give inverse bijections. They restrict to inverse bijections

$$\hat{L} \underset{\log}{\overset{\exp}{\rightleftarrows}} G(\hat{U}(L))$$

between the completed Lie algebra and the grouplike elements.

Proof That $\log(\exp(x)) = x$ and $\exp(\log(1+s)) = 1+s$, are formal manipulations. If $\ell \in \hat{L}$ it is again a formal manipulation that

$$\Delta(\exp(\ell)) = \exp(\ell) \cdot \exp(\ell),$$

and so $\exp(\ell)$ is a grouplike element.

The maps exp and log can also be defined on the tensor products and give inverse bijections

$$\hat{U}(L)_+ \hat{\otimes} \hat{U}(L) + \hat{U}(L) \hat{\otimes} \hat{U}(L)_+ \underset{\log}{\overset{\exp}{\rightleftarrows}} 1 \otimes 1 + \hat{U}(L)_+ \hat{\otimes} \hat{U}(L) + \hat{U}(L) \hat{\otimes} \hat{U}(L)_+.$$

Now let $s \in G(\hat{U}(L))$ be a grouplike element. Since $\Delta = \Delta_{\hat{U}(L)}$ is an algebra homomorphism

$$\exp(\Delta(\log(s))) = \Delta(\exp(\log s)) = \Delta(s) = s \otimes s.$$

Since $1 \otimes s$ and $s \otimes 1$ are commuting elements we also have

$$\exp(\log(s) \otimes 1 + 1 \otimes \log(s)) = (\exp(\log(s)) \otimes 1) \cdot (1 \otimes \exp(\log(s))) = s \otimes s.$$

Taking logarithms of these two equations, we obtain

$$\Delta(\log(s)) = \log(s) \otimes 1 + 1 \otimes \log(s),$$

and so $\log(s)$ is in $\hat{L}$. □

3 Exponentials and Logarithms for Pre- and Post-Lie Algebras

For pre- and post-Lie algebras their enveloping algebra comes with two products $\bullet$ and $*$. This gives two possible exponential and logarithm maps. This is precisely the setting that enables us to define a map from formal vector fields to formal flows. It also gives the general setting for defining the Butcher product.

3.1 Filtered Pre- and Post-Lie Algebras

Given a linear binary operation on a k-vector space A

$$* : A \otimes_k A \to A$$

the associator is defined as:

$$a_*(x, y, z) = x * (y * z) - (x * y) * z.$$

Definition 3.1 A *post-Lie algebra* $(P, [,], \triangleright)$ is a Lie algebra $(P, [,])$ together with a linear binary map $\triangleright$ such that

- $x \triangleright [y, z] = [x \triangleright y, z] + [y, x \triangleright z]$
- $[x, y] \triangleright z = a_{\triangleright}(x, y, z) - a_{\triangleright}(y, x, z)$

It is then straightforward to verify that the following bracket

$$[\![x, y]\!] = x \triangleright y - y \triangleright x + [x, y]$$

defines another Lie algebra structure on P.

A *pre-Lie* algebra is a post-Lie algebra P such that bracket $[,]$ is zero, so P with this bracket is the abelian Lie algebra.

Example 3.2 Let $\mathcal{X}\mathbb{R}^n$ be the vector fields on the manifold $\mathbb{R}^n$. It comes with the natural Levi-Cevita connection ∇. Write $f = \sum_{i=1}^n f^i \partial_i$ and $g = \sum_{i=1}^n g^i \partial_i$ for two vector fields, where $\partial_i = \partial/\partial x_i$. Let

$$f \triangleright g = \nabla_f g = \sum_{i,j} f^j (\partial_j g^i) \partial_i.$$

Then $\mathcal{X}\mathbb{R}^n$ is a pre-Lie algebra with this operation. Hence also a post-Lie algebra with trivial Lie-bracket $[,]$ equal to zero.

Example 3.3 Let M be a manifold and $\mathcal{X}M$ the vector fields on M. Let $\mathfrak{g}$ be a finite dimensional Lie algebra and $\lambda : \mathfrak{g} \to \mathcal{X}M$ be a morphism of Lie algebras. Denote by $\Omega^0(M, \mathfrak{g})$ the space of smooth maps $M \to \mathfrak{g}$. This is a Lie algebra by

$$[x, y](u) = [x(u), y(u)].$$

The vector fields $\mathcal{X}M$ act on the functions $\Omega^0(M, k)$ by differentiation: For $f \in \mathcal{X}M$ and $\phi \in \Omega^0(M, k)$ we get $f\phi \in \Omega^0(M, k)$. Hence $\mathcal{X}M$ acts on $\Omega^0(M, \mathfrak{g}) = \Omega^0(M, k) \otimes_k \mathfrak{g}$.

Now define the operation

$$\begin{aligned} \Omega^0(M, \mathfrak{g}) \times \Omega^0(M, \mathfrak{g}) &\xrightarrow{\triangleright} \Omega^0(M, \mathfrak{g}) \\ x \triangleright y &\mapsto [u \mapsto (\lambda(x(u))y)(u)]. \end{aligned}$$

Then $\Omega^0(M, \mathfrak{g})$, $[,]$, $\triangleright$ becomes a post-Lie algebra by [24, Prop.2.10].

If $G \times M \to M$ is an action of a Lie group G on M then for each $u \in M$ we get a map $G \to M$ and on tangent spaces $\mathfrak{g} \to T_u M$. This gives a map to the tangent bundle of M: $\mathfrak{g} \times M \to TM$ and map of Lie algebras $\mathfrak{g} \to \mathcal{X}M$. Hence in this setting we get by the above a post-Lie algebra $\Omega^0(M, \mathfrak{g})$.

If $M = G$ and $G \times G \to G$ is the Lie group operation, then $\Omega^0(G, \mathfrak{g})$ naturally identifies with the vector fields $\mathcal{X}G$ by left multiplication, and so these vector fields becomes a post-Lie algebra. In the special case that $G = \mathbb{R}^n$ with group operation $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ sending $(a, b) \mapsto a + b$, we get the pre-Lie algebra of Example 3.2 above.

We now assume that P is a filtered post-Lie algebra: We have a decreasing filtration

$$P = P^1 \supseteq P^2 \supseteq \cdots,$$

such that

$$[P^p, P^q] \subseteq P^{p+q}, \quad P^p \rhd P^q \subseteq P^{p+q},$$

Then we will also have $[\![P^p, P^q]\!] \subseteq P^{p+q}$. If u and v are two elements of P such that $u - v \in P^{n+1}$, we say they are equal up to order n.

Again examples of this can be constructed for any post-Lie algebra over a field k by letting $P^1 = P$ and

$$P^{p+1} := P^p \rhd P + P \rhd P^p + [P, P^p].$$

Alternatively we may form the polynomials $P[h] = \oplus_{n\geq 1} Ph^n$, or the power series $P[\![h]\!] = \Pi_{n\geq 1} Ph^n$.

In [10] the enveloping algebra $U(P)$ of the post-Lie algebra was introduced. It is both the enveloping algebra for the Lie algebra $[,]$ and as such comes with associative product $\bullet$, and is the enveloping algebra for the Lie algebra $[\![,]\!]$ and as such comes with associative product $*$. The triangle product also extends to a product $\rhd$ on $U(P)$ but this is not associative.

3.2 The Map from Fields to Flows

By Example 3.2 above the formal power series of vector field $\mathcal{X}\mathbb{R}^n[\![h]\!]$ is a pre-Lie algebra, and from the last part of Example 3.3 we get a post-Lie algebra $\mathcal{X}G[\![h]\!]$ of series of vector fields. Using this perspective there are several natural ways to think of filtered post-Lie algebras and the related objects.

- The elements of P may be thought of as formal vector fields, in which case we write P_{field}.
- The grouplike elements of $\hat{U}(P)$ may be thought of as formal flows.
- The elements of P may be thought of as principal parts of formal flows, see below, in which case we write P_{flow}.

Let us explain how these are related. In the rest of this subsection we assume that $P = \hat{P}$ is complete with respect to the filtration. The exponential map

$$P_{field} \xrightarrow{\exp^*} \hat{U}(P) \tag{8}$$

sends a vector field to a formal flow, a grouplike element in $\hat{U}(P)$. (Note that the notion of a grouplike element in $\hat{U}(P)$ only depends on the shuffle coproduct.)

We may take the logarithm

$$G(\hat{U}(P)) \xrightarrow{\log^\bullet} P. \tag{9}$$

So if $B \in G(\hat{U}(P))$ we get $b = \log^\bullet(B)$. We think of b also as a formal flow, the *principal part* or *first order part* of the formal flow B. It determines B by $B = \exp^\bullet(b)$. Note that in (8) the exponential is with respect to the $*$ operation, while in (9) the logarithm is with respect to the $\bullet$ operation.

When P is a pre-Lie algebra A, then $\hat{U}(P)$ is the completed symmetric algebra $\widehat{\mathrm{Sym}}(A)$ and $\log^\bullet$ is simply the projection $\widehat{\mathrm{Sym}}(A) \to A$. If B is a Butcher series parametrized by forests (see Sect. 6.3), then b is the Butcher series parametrized by trees. Thus b determines the flow, but the full series B is necessary to compute pull-backs of functions along the flow.

We thus get a bijection

$$\Phi : P_{field} \xrightarrow{\log^\bullet \circ \exp^*} P_{flow} \tag{10}$$

which maps vector fields to principal part flows. This map is closely related to the *Magnus expansion* [8]. Magnus expresses the exact flow as $\exp^*(tv) = \exp^\bullet(\Phi(tv))$, from which a differential equation for $\Phi(tv)$ can be derived.

Example 3.4 Consider the manifold $\mathbb{R}^n$ and let $\mathcal{X}\mathbb{R}^n$ be the vector fields on $\mathbb{R}^n$. Let $f = \sum_{i\geq 0} f_i h^i$ on $\mathbb{R}^n$ be a power series of vector fields where each $f_i \in \mathcal{X}\mathbb{R}^n$. It induces the flow series $\exp^*(hf)$ in $\hat{U}(\mathcal{X}\mathbb{R}^n[\![h]\!])$. Since $\mathcal{X}\mathbb{R}^n$ is a pre-Lie algebra, the completed enveloping algebra is $\widehat{\mathrm{Sym}}(\mathcal{X}\mathbb{R}^n[\![h]\!])$. Thus the series

$$\exp^*(hf) = 1 + \sum_{i\geq d\geq 1} F_{i,d} h^i$$

where the $F_{i,d} \in \mathrm{Sym}_d(\mathcal{X}\mathbb{R}^n[\![h]\!])$ are d'th order differential operators. (Note that the principal part b is the $d = 1$ part.) It determines a flow $\Psi_h^f : \mathbb{R}^n \to \mathbb{R}^n$ sending a point P to $P(h)$. For any smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$ the *pullback* of ϕ along the flow is the composition $\phi \circ \Psi_h^f : \mathbb{R}^n \to \mathbb{R}$ and is given by

$$\exp^*(hf)\phi = 1 + \sum_{i\geq d\geq 1} F_{i,d}(\phi) h^i,$$

see [17, Section 4.1] or [23, Section 2.1]. In particular when ϕ is a coordinate function x_p we get the coordinate $x_p(h)$ of $P(h)$ as given by

$$x_p(h) = \exp^*(hf)x_p = \sum_{i\geq d\geq 1} F_{i,d} x_p h^i = x_p + \sum_i F_{i,1} x_p h^i$$

since higher derivatives of x_p vanish. This shows concretely geometrically why the flow is determined by its principal part.

For a given principal flow $b \in P_{flow}$ computing its inverse image by the map (10) above, which is the vector field $\log^* \circ \exp^\bullet(b)$ is called *backward error* in numerical analysis [14, 19].

For $a, a' \in P_{field}$ let

$$a * a' = \log^*(\exp^*(a) * \exp^*(a')),$$

a product which is computed using the Baker-Campbell-Hausdorff (BCH) formula for the Lie algebra $[\![,]\!]$. With this product P_{field} becomes a pro-unipotent group. Transporting this product to P_{flow} using the bijection Φ in (10), we get for $b, b' \in P_{flow}$ a product

$$b \natural b' = \log^\bullet(\exp^\bullet(b) * \exp^\bullet(b')),$$

the *composition* product for principal flows.

Example 3.5 We continue Example 3.4. Let $g = \sum_{i \geq 0} g_i h^i$ be another power series of vector fields, $\exp^*(hg)$ its flow series, and $\Psi_h^g : \mathbb{R}^n \to \mathbb{R}^n$ the flow it determines. Let c be the principal part of $\exp^*(hg)$. The composition of the flows $\Psi_h^g \circ \Psi_h^f$ is the flow sending ϕ to

$$\exp^*(hg)(\exp^*(hf)\phi) = (\exp^*(hg) * \exp^*(hf))\phi.$$

The principal part of the composed flow is

$$\log^\bullet(\exp^*(hg) * \exp^*(hf)) = \log^\bullet(\exp^\bullet(c) * \exp^\bullet(b)) = c \natural b,$$

the Butcher product of c and b.

Denote by $\dotplus$ the product in P_{flow} given by the BCH-formula for the Lie bracket $[,]$,

$$x \dotplus y := \log^\bullet(\exp^\bullet(x) \bullet \exp^\bullet(y)).$$

Proposition 3.6 *For x, y in the post-Lie algebra P_{flow} we have*

$$x \natural y = x \dotplus \left(\exp^\bullet(x) \rhd y\right).$$

Proof From [10, Prop.3.3] the product $A * B = \sum_{\Delta(A)} A_{(1)}(A_{(2)} \rhd B)$. Since $\exp^\bullet(x)$ is a group-like element it follows that:

$$\exp^\bullet(x) * \exp^\bullet(y) = \exp^\bullet(x) \bullet \left(\exp^\bullet(x) \rhd \exp^\bullet(y)\right).$$

By [10, Prop.3.1] $A \rhd BC = \sum_{\Delta(A)} (A_{(1)} \rhd B)(A_{(2)} \rhd C)$ and so again using that $\exp^\bullet(x)$ is group-like and the expansion of $\exp^\bullet(y)$:

$$\exp^\bullet(x) \rhd \exp^\bullet(y) = \exp^\bullet\left(\exp^\bullet(x) \rhd y\right).$$

Hence

$$x \sharp y = \log^\bullet\left(\exp^\bullet(x) \bullet \left(\exp^\bullet(x) \rhd \exp^\bullet(y)\right)\right) = \log^\bullet\left(\exp^\bullet(x) \bullet \exp^\bullet(\exp^\bullet(x) \rhd y)\right).$$

□

In the pre-Lie case $[,] = 0$, therefore $\dotplus = +$ and we obtain the formula derived in [9]

$$x \sharp y = x + \exp^\bullet(x) \rhd y.$$

3.3 Substitution

Let $\mathrm{End}_{\mathrm{postLie}}(P) = \mathrm{Hom}_{\mathrm{postLie}}(P, P)$ be the endomorphisms of P as a post-Lie algebra. (In the special case that P is a pre-Lie algebra, this is simply the endomorphisms of P as a pre-Lie algebra.) It is a monoid, but not generally a vector space. It acts on the post-Lie algebra P.

Since the action respects the brackets $[,]$, $[\![,]\!]$ and $\rhd$, it also acts on the enveloping algebra $U(P)$ and its completion $\hat{U}(P)$, and respects the products $*$ and $\bullet$. Hence the exponential maps $\exp^*$ and $\exp^\bullet$ are equivariant for this action, and similarly the logarithms $\log^*$ and $\log^\bullet$. So the formal flow map

$$\Phi : P_{\mathrm{field}} \longrightarrow P_{\mathrm{flow}}$$

is equivariant for the action. The action on P_{flow} (which is technically the same as the action on P_{field}), is called *substitution* and is usually studied in a more specific context, as we do in Sect. 7. An element $\phi \in \mathrm{End}_{\mathrm{postLie}}(P)$ comes from sending a field f to a perturbed field f', and one then sees how this affects the exact flow or approximate flow maps given by numerical algorithms.

Part 2: The Algebraic Geometric Setting

In this part we have certain finiteness assumptions on the Lie algebras and pre- and post-Lie algebras, and so may consider them and binary operations on them in the setting of varieties. The first three subsections of the next Sect. 4 will be quite familiar to the reader who knows basic algebraic geometry.

4 Affine Varieties and Group Actions

We assume the reader is familiar with basic algebraic geometry of varieties and morphisms, like presented in [16, Chap.1] or [7, Chap.1,5]. We nevertheless briefly recall basic notions. A notable and not so standard feature is that we in the last subsection define infinite dimensional varieties and morphisms between them.

4.1 Basics on Affine Varieties

Let k be a field and $S = k[x_1, \ldots, x_n]$ the polynomial ring. The *affine n-space* is

$$\mathbb{A}_k^n = \{(a_1, \ldots, a_n) \mid a_i \in k\}.$$

An ideal $I \subseteq S$ defines an *affine variety* in $\mathbb{A}_k^n$:

$$X = \mathcal{Z}(I) = \{p \in \mathbb{A}_k^n \mid f(p) = 0, \text{ for } f \in I\}.$$

Given an affine variety $X \subseteq \mathbb{A}_k^n$, its associated ideal is

$$\mathcal{I}(X) = \{f \in S \mid f(p) = 0, \text{ for } p \in X\}.$$

Note that if $X = \mathcal{Z}(I)$ then $I \subseteq \mathcal{I}(X)$, and $\mathcal{I}(X)$ is the largest ideal defining the variety X. The correspondence

$$\text{ideals in } k[x_1, \ldots, x_n] \underset{\mathcal{I}}{\overset{\mathcal{Z}}{\rightleftarrows}} \text{ subsets of } \mathbb{A}_k^n$$

is a Galois connection. Thus we get a one-to-one correspondence

$$\text{image of } \mathcal{I} \overset{1-1}{\longleftrightarrow} \text{image of } \mathcal{Z} \quad . \\ = \text{varieties in } \mathbb{A}_k^n$$

Remark 4.1 When the field k is algebraically closed, Hilbert's Nullstellensatz says that the image of $\mathcal{I}$ is precisely the radical ideals in the polynomial ring. In general however the image of $\mathcal{I}$ is only contained in the radical ideals.

The *coordinate ring* of a variety X is the ring $A(X) = k[x_1, \ldots, x_n]/\mathcal{I}(X)$. A morphism of affine varieties $f : X \to Y$ where $X \subseteq \mathbb{A}_k^n$ and $Y \subseteq \mathbb{A}_k^m$ is a a map sending a point $\mathbf{a} = (a_1, \ldots, a_n)$ to a point $(f_1(\mathbf{a}), \ldots, f_m(\mathbf{a}))$ where the f_i are polynomials in S. This gives rise to a homomorphism of coordinate rings

$$\begin{aligned} f^\sharp : A(Y) &\longrightarrow A(X) \\ \overline{y_i} &\longrightarrow f_i(\overline{\mathbf{x}}), \quad i = 1, \ldots, m \end{aligned}$$

In fact this is a one-one correspondence:

$$\{\text{morphisms } f : X \to Y\} \quad \overset{1-1}{\longleftrightarrow} \quad \{\text{algebra homomorphisms } f^\sharp : A(Y) \to A(X)\}.$$

The zero-dimensional affine space $\mathbb{A}_k^0$ is simply a point, and its coordinate ring is k. Therefore to give a point $p \in \mathbb{A}_k^n$ is equivalent to give an algebra homomorphism $k[x_1, \ldots, x_n] \to k$.

Remark 4.2 We may replace k by any commutative ring $\Bbbk$. The affine space $\mathbb{A}_\Bbbk^n$ is then $\Bbbk^n$. The coordinate ring of this affine space is $\Bbbk[x_1, \ldots, x_n]$. A point $p \in \mathbb{A}_\Bbbk^n$ still corresponds to an algebra homomorphism $\Bbbk[x_1, \ldots, x_n] \to \Bbbk$. Varieties in $\mathbb{A}_\Bbbk^n$ may be defined in the same way, and there is still a Galois connection between ideals in $\Bbbk[x_1, \ldots, x_n]$ and subsets of $\mathbb{A}_\Bbbk^n$, and a one-one correspondence between morphisms of varieties and coordinate rings.

The affine space $\mathbb{A}_k^n$ comes with the *Zariski topology*, whose closed sets are the affine varieties in $\mathbb{A}_k^n$ and whose open sets are the complements of these. This induces also the Zariski topology on any affine subvariety X in $\mathbb{A}_k^n$.

If X and Y are affine varieties in $\mathbb{A}_k^n$ and $\mathbb{A}_k^m$ respectively, their product $X \times Y$ is an affine variety in $\mathbb{A}_k^{n+m}$ whose ideal is the ideal in $k[x_1, \ldots, x_n, y_1, \ldots, y_m]$ generated by $\mathcal{I}(X) + \mathcal{I}(Y)$. Its coordinate ring is

$$A(X \times Y) = A(X) \otimes_k A(Y).$$

If A is a ring and $f \neq 0$ in A, we have the localized ring A_f whose elements are all a/f^n where $a \in A$. Two such elements a/f^n and b/f^m are equal if $f^k(f^m a - f^n b) = 0$ for some k. If A is an integral domain, this is equivalent to $f^m a - f^n b = 0$. Note that the localization A_f is isomorphic to the quotient ring $A[x]/(xf - 1)$. Hence if A is a finitely generated k-algebra, A_f is also a finitely generated k-algebra. A consequence of this is the following: Let X be an affine variety in $\mathbb{A}_k^n$ whose ideal is $I = \mathcal{I}(X)$ contained in $k[x_1, \ldots, x_n]$, and let f be a polynomial function. The open subset

$$D(f) = \{p \in X \mid f(p) \neq 0\} \subseteq X$$

is then in bijection to the variety $X' \in \mathbb{A}_k^{n+1}$ defined by the ideal $I + (x_{n+1} f - 1)$. This bijection is actually a homeomorphism in the Zariski topology. The coordinate ring

$$A(X') = A(X)[x_{n+1}]/(x_{n+1} f - 1) \cong A(X)_f.$$

Hence we identify A_f as the coordinate ring of the open subset $D(f)$ and can consider $D(f)$ as an affine variety. Henceforth we shall drop the adjective affine for a variety, since all our varieties will be affine.

4.2 Coordinate Free Descriptions of Varieties

For flexibility of argument, it may be desirable to consider varieties in a coordinate free context.

Let V and W be dual finite dimensional vector spaces. So $V = \mathrm{Hom}_k(W, k) = W^*$, and then W is naturally isomorphic to $V^* = (W^*)^*$. We consider V as an affine space (this means that we are forgetting the structure of vector space on V). Its coordinate ring is the symmetric algebra $\mathrm{Sym}(W)$. Note that any polynomial $f \in \mathrm{Sym}(W)$ may be evaluated on any point $\mathbf{v} \in V$, since $\mathbf{v} : W \to k$ gives maps $\mathrm{Sym}_d(W) \to \mathrm{Sym}_d(k) = k$ and thereby a map $\mathrm{Sym}(W) = \oplus_d \mathrm{Sym}_d(W) \to k$.

Given an ideal I in $\mathrm{Sym}(W)$, the associated affine variety is

$$X = \{\mathbf{v} \in V \mid f(\mathbf{v}) = 0, \text{ for } f \in I\} \subseteq V.$$

Given a variety $X \subseteq V$ we associate the ideal

$$\mathcal{I}(X) = \{f \in \mathrm{Sym}(W) \mid f(\mathbf{v}) = 0, \text{ for } \mathbf{v} \in X\} \subseteq \mathrm{Sym}(W).$$

The coordinate ring of X is $A(X) = \mathrm{Sym}(W)/\mathcal{I}(X)$.

Let W^1 and W^2 be two vector spaces, with dual spaces V^1 and V^2. A map $f : X^1 \to X^2$ between varieties in these spaces is a map which is given by polynomials once a coordinate system is fixed for V^1 and V^2. Such a map then gives a homomorphism of coordinate rings $f^\sharp : \mathrm{Sym}(W^2)/I(X^2) \to \mathrm{Sym}(W^1)/I(X^1)$, and this gives a one-one correspondence between morphisms f between X^1 and X^2 and algebra homomorphisms $f^\sharp$ between their coordinate rings.

4.3 Affine Spaces and Monoid Actions

The vector space of linear operators on V is denoted $\mathrm{End}(V)$. It is an affine space with $\mathrm{End}(V) \cong \mathbb{A}_k^{n \times n}$, and with coordinate ring $\mathrm{Sym}(\mathrm{End}(V)^*)$. We then have an action

$$\begin{aligned} \mathrm{End}(V) \times V &\to V \\ (\phi, v) &\mapsto \phi(v). \end{aligned} \tag{11}$$

This is a morphism of varieties. Explicitly, if V has basis $e_1, \ldots, e_n$ an element in $\mathrm{End}(V)$ may be represented by a matrix A and the map is given by:

$$(A, (v_1, \ldots, v_n)^t) \mapsto A \cdot (v_1, \ldots, v_n)^t,$$

which is given by polynomials.

The morphism of varieties (11) then corresponds to the algebra homomorphism on coordinate rings

$$\mathrm{Sym}(V^*) \to \mathrm{Sym}(\mathrm{End}(V)^*) \otimes_k \mathrm{Sym}(V^*).$$

With a basis for V, the coordinate ring $\mathrm{Sym}(\mathrm{End}(V)^*)$ is isomorphic to the polynomial ring $k[t_{ij}]_{i,j=1,\dots,n}$, where the t_{ij} are coordinate functions on $\mathrm{End}(V)$, and the coordinate ring $\mathrm{Sym}(V^*)$ is isomorphic to $k[x_1, \dots, x_n]$ where the x_i are coordinate functions on V. The map above on coordinate rings is then given by

$$x_i \mapsto \sum_j t_{ij} x_j.$$

We may also consider the set $\mathrm{GL}(V) \subseteq \mathrm{End}(V)$ of invertible linear operators. This is the open subset $D(\det(t_{ij}))$ of $\mathrm{End}(V)$ defined by the nonvanishing of the determinant. Hence, fixing a basis of V, its coordinate ring is the localized ring $k[t_{ij}]_{\det((t_{ij}))}$, by the last part of Sect. 4.1. The set $\mathrm{SL}(V) \subseteq \mathrm{End}(V)$ are the linear operators with determinant 1. This is a closed subset of $\mathrm{End}(V)$ defined by the polynomial equation $\det((t_{ij})) - 1 = 0$. Hence the coordinate ring of $\mathrm{SL}(V)$ is the quotient ring $k[t_{ij}]/(\det((t_{ij})) - 1)$.

Now given an affine monoid variety M, that is an affine variety with a product morphism $\mu : M \times M \to M$ which is associative and unital. Then we get an algebra homomorphism of coordinate rings

$$A(M) \xrightarrow{\Delta} A(M) \otimes_k A(M).$$

Since the following diagram commutes

$$\begin{array}{ccc} M \times M \times M & \xrightarrow{\mu \times \mathbf{1}} & M \times M \\ {\scriptstyle \mathbf{1} \times \mu}\downarrow & & \downarrow{\scriptstyle \mu} \\ M \times M & \xrightarrow{\mu} & M, \end{array}$$

we get a commutative diagram of coordinate rings:

$$\begin{array}{ccc} A(M) \otimes_k A(M) \otimes_k A(M) & \xleftarrow[\Delta \otimes \mathbf{1}]{} & A(M) \otimes A(M) \\ {\scriptstyle \mathbf{1} \otimes \Delta}\uparrow & & \uparrow{\scriptstyle \Delta} \\ A(M) \otimes_k A(M) & \xleftarrow{\Delta} & A(M). \end{array}$$

The zero-dimensional affine space $\mathbb{A}^0_k$ is simply a point, and its coordinate ring is k. A character on $A(M)$ is an algebra homomorphism $A(M) \to k$. On varieties this

gives a morphism $P = \mathbb{A}_k^0 \to M$, or a point in the monoid variety. In particular the unit in M corresponds to a character $A(M) \xrightarrow{\epsilon} k$, the counit. Thus the algebra $A(M)$ with Δ and ϵ becomes a bialgebra.

The monoid may act on a variety X via a morphism of varieties

$$M \times X \to X. \tag{12}$$

On coordinate rings we get a homomorphism of algebras,

$$A(X) \to A(M) \otimes_k A(X), \tag{13}$$

making $A(X)$ into a comodule algebra over the bialgebra $A(M)$.

In coordinate systems the morphism (12) may be written:

$$(m_1, \ldots, m_r) \times (x_1, \ldots, x_n) \mapsto (f_1(\mathbf{m}, \mathbf{x}), f_2(\mathbf{m}, \mathbf{x}), \ldots).$$

If X is an affine space V and the action comes from a morphism of monoid varieties $M \to \mathrm{End}(V)$, the action by M is linear on V. Then $f_i(\mathbf{m}, \mathbf{v}) = \sum_j f_{ij}(\mathbf{m}) v_j$. The homomorphism on coordinate rings (recall that $V = W^*$)

$$\mathrm{Sym}(W) \to A(M) \otimes_k \mathrm{Sym}(W)$$

is then induced from a morphism

$$W \to A(M) \otimes_k W$$
$$x_j \mapsto \sum_i f_{ij}(\mathbf{u}) \otimes_k x_i$$

where the x_j's are the coordinate functions on V and $\mathbf{u}$ are the coordinate functions on M.

We can also consider an affine group variety G with a morphism $G \to GL(V)$ and get a group action $G \times V \to V$. The inverse morphism for the group, induces an antipode on the coordinate ring $A(G)$ making it a commutative Hopf algebra.

4.4 Infinite Dimensional Affine Varieties and Monoid Actions

The infinite dimensional affine space $\mathbb{A}_k^\infty$ is $\prod_{i \geq 1} k$. Its elements are infinite sequences $(a_1, a_2, \ldots)$ where the a_i are in k. Its coordinate ring is the polynomial ring in infinitely many variables $S = k[x_i, i \in \mathbb{N}]$.

An ideal I in S, defines an affine variety

$$X = V(I) = \{\mathbf{a} \in \mathbb{A}_k^\infty \mid f(\mathbf{a}) = 0, \text{ for } f \in I\}.$$

Note that a polynomial f in S always involves only a finite number of the variables, so the evaluation $f(\mathbf{a})$ is meaningful. Given an affine variety X, let its ideal be:

$$\mathcal{I}(X) = \{f \in S \mid f(\mathbf{a}) = 0 \text{ for } \mathbf{a} \in X\}.$$

The coordinate ring $A(X)$ of X is the quotient ring $S/\mathcal{I}(X)$. The affine subvarieties of $\mathbb{A}_k^\infty$ form the closed subsets in the Zariski topology on $\mathbb{A}_k^\infty$, and this then induces the Zariski topology on any subvariety of $\mathbb{A}_k^\infty$.

A morphism $f : X \to Y$ of two varieties, is a map such that $f(\mathbf{a}) = (f_1(\mathbf{a}), f_2(\mathbf{a}), \ldots)$ where each f_i is a polynomial function (and so involves only a finite number of the coordinates of $\mathbf{a}$).

Letting $k[y_i, i \in \mathbb{N}]$ be the coordinate ring of affine space where Y lives, we get a morphism of coordinate rings

$$\begin{aligned} f^\sharp : A(Y) &\to A(X) \\ \overline{y_i} &\mapsto f_i(\overline{\mathbf{x}}) \end{aligned}$$

This gives a one-one correspondence

$$\{\text{morphisms } f : X \to Y\} \quad \leftrightarrow \quad \{\text{algebra homomorphisms } f^\sharp : A(Y) \to A(X)\}.$$

For flexibility of argument, it is desirable to have a coordinate free definition of these varieties also. The following includes then both the finite and infinite-dimensional case in a coordinate free way.

Let W be a vector space with a countable basis. We get the symmetric algebra $\mathrm{Sym}(W)$. Let $V = \mathrm{Hom}_k(W, k)$ be the dual vector space, which will be our affine space. Given an ideal I in $\mathrm{Sym}(W)$, the associated affine variety is

$$X = V(I) = \{\mathbf{v} \in V \mid f(\mathbf{v}) = 0, \text{ for } f \in I\}.$$

The evaluation of f on $\mathbf{v}$ is here as explained in Sect. 4.2. Given a variety X we associate the ideal

$$\mathcal{I}(X) = \{f \in \mathrm{Sym}(W) \mid f(\mathbf{v}) = 0, \text{ for } \mathbf{v} \in X\}.$$

Its coordinate ring is $A(X) = \mathrm{Sym}(W)/\mathcal{I}(X)$. We shall shortly define morphism between varieties. In order for these to be given by polynomial maps, we will need filtrations on our vector spaces. Given a filtration by finite dimensional vector spaces

$$\langle 0 \rangle = W_0 \subseteq W_1 \subseteq W_2 \subseteq \cdots \subseteq W.$$

On the dual space V we get a decreasing filtration by $V^i = \ker((W)^* \to (W_{i-1})^*)$. The affine variety $V/V^i \cong (W_{i-1})^*$ has coordinate ring $\mathrm{Sym}(W_{i-1})$. If X is a variety in V its image X_i in the finite affine space V/V^i need not be Zariski

closed. Let $\overline{X_i}$ be its closure. This is an affine variety in V/V^i whose ideal is $\mathcal{I}(X) \cap \mathrm{Sym}(W_{i-1})$.

A map $f : X_1 \to X_2$ between varieties in these spaces is a *morphism of varieties* if there exists decreasing filtrations

$$V_1 = V_1^1 \supseteq V_1^2 \supseteq \cdots, \quad V_2 = V_2^1 \supseteq V_2^2 \supseteq \cdots$$

with finite dimensional quotient spaces, such that for any i we have a commutative diagram

$$\begin{array}{ccc} X_1 & \xrightarrow{f} & X_2 \\ \downarrow & & \downarrow \\ \overline{X_{1,i}} & \longrightarrow & \overline{X_{2,i}} \end{array}$$

and the lower map is a morphism between varieties in V_1/V_1^i and V_2/V_2^i.

We then get a homomorphisms of coordinate rings

$$f_i^\sharp : \mathrm{Sym}(W_i^2)/\mathcal{I}(X_{2,i}) \to \mathrm{Sym}(W_i^1)/\mathcal{I}(X_{1,i}), \tag{14}$$

and the direct limit of these gives a homomorphism of coordinate rings

$$f^\sharp : \mathrm{Sym}(W^2)/\mathcal{I}(X_2) \to \mathrm{Sym}(W^1)/\mathcal{I}(X_1). \tag{15}$$

Conversely given an algebra homomorphism $f^\sharp$ above. Let

$$W_1^2 \subseteq W_2^2 \subseteq W_3^2 \subseteq \cdots$$

be a filtration. Write $W^1 = \oplus_{i\in\mathbb{N}} k w_i$ in terms of a basis. The image of W_i^2 will involve only a finite number of the w_i. Let W_i^1 be the f.d. subvector space generated by these w_i. Then we get maps (14), giving morphisms

$$\begin{array}{ccc} \overline{X_{1,i+1}} & \longrightarrow & \overline{X_{2,i+1}} \\ \downarrow & & \downarrow \\ \overline{X_{1,i}} & \longrightarrow & \overline{X_{2,i}}. \end{array}$$

In the limit we then get a morphism of varieties $f : X_1 \to X_2$. This gives a one-one correspondence between morphisms of varieties $f : X_1 \to X_2$ and algebra homomorphisms $f^\sharp$.

Let X^1 and X^2 be varieties in the affine spaces V^1 and V^2. Their product $X^1 \times X^2$ is a variety in the affine space $V^1 \times V^2$ which is the dual space of $W^1 \oplus W^2$. Its coordinate ring is $A(X^1 \times X^2) = A(X^1) \otimes_k A(X^2)$.

If M is an affine monoid variety (possibly infinite dimensional) its coordinate ring $A(M)$ becomes a commutative bialgebra. If M is an affine group variety, then $A(M)$ is a Hopf algebra. We can again further consider an action on the affine space

$$M \times V \to V.$$

It corresponds to a homomorphism of coordinate rings

$$\mathrm{Sym}(W) \to A(M) \otimes_k \mathrm{Sym}(W),$$

making $\mathrm{Sym}(W)$ into a comodule algebra over $A(M)$. If the action by M is linear on V, the algebra homomorphism above is induced by a linear map $W \to A(M) \otimes_k W$.

5 Filtered Algebras with Finite Dimensional Quotients

In this section we assume the quotients $L_p = L/L^{p+1}$ from Sect. 2.2 are finite dimensional vector spaces. This enables us to define the dual Hopf algebra $U^c(K)$ of the enveloping algebra $U(L)$. This Hopf algebra naturally identifies as the coordinate ring of the completed Lie algebra $\hat{L}$. In Sect. 5.3 the Baker-Campbell-Hausdorff product on the variety L is shown to correspond to the natural coproduct on the dual Hopf algebra $U^c(K)$. In the last Sect. 5.4 the Lie-Butcher product on a post-Lie algebra is also shown to correspond to the natural coproduct on the dual Hopf algebra.

5.1 Filtered Lie Algebras with Finite Dimensional Quotients

Recall that L_p is the quotient L/L^{p+1} from Sect. 2.2. The setting in this section is k is a field of characteristic zero, and that these quotients L_p are finite dimensional as k-vector spaces. We assume that the Lie algebra L is complete with respect to this cofiltration, so we have the inverse limit

$$L = \hat{L} = \varprojlim_p L_p.$$

The dual $K^p = \mathrm{Hom}_k(L_p, k)$ is a finite dimensional Lie coalgebra. Let $K = \varinjlim_p K^p$ be the direct limit. Recall that the quotient algebra

$$U^j(L_p) = U(L_p)/F^{j+1}U(L_p).$$

The dual $U^j(L_p)^*$ is a finite dimensional coalgebra $U^c_j(K^p)$, and we have inclusions

$$\begin{array}{ccc} U^c_j(K^p) & \xrightarrow{\subseteq} & U^c_{j+1}(K^p) \\ \subseteq\downarrow & & \downarrow\subseteq \\ U^c_j(K^{p+1}) & \xrightarrow{\subseteq} & U^c_{j+1}(K^{p+1}). \end{array}$$

We have the direct limits

$$U^c(K^p) := \varinjlim_j U^c_j(K^p), \quad U^c(K) := \varinjlim_{j,p} U^c_j(K^p).$$

Lemma 5.1 *Let $T^c(K)$ be the tensor coalgebra. It is a Hopf algebra with the shuffle product. Then $U^c(K)$ is a Hopf sub-algebra of $T^c(K)$.*

Proof $U^j(L_p)$ is a quotient algebra of $T(L_p)$ and $T(L)$, and so $U^c_j(K_p)$ is a subcoalgebra of $T^c(K_p)$ and $T^c(K)$. The coproduct on $U(L_p)$, the shuffle coproduct, does not descend to a coproduct on $U^j(L_p)$. But we have a well defined map

$$U^{2j}(L_p) \to U^j(L_p) \otimes U^j(L_p)$$

compatible with the shuffle coproduct on $T(L_p)$. Dualizing this we get

$$U^c_j(K_p) \otimes U^c_j(K_p) \to U^c_{2j}(K_p)$$

and taking colimits, we get $U^c(K)$ as a subalgebra of $T^c(K)$ with respect to the shuffle product. □

Proposition 5.2 *There are isomorphisms*

a. $L \cong Hom_k(K, k)$ *of Lie algebras,*
b. $\hat{U}(L) \cong Hom_k(U^c(K), k)$ *of algebras.*
c. *The coproduct on $U^c(K)$ is dual to the completed product on $\hat{U}(L)$*

$$U^c(K) \xrightarrow{\Delta_\bullet} U^c(K) \otimes U^c(K), \quad \hat{U}(L)\hat{\otimes}\hat{U}(L) \xrightarrow{\bullet} \hat{U}(L).$$

Proof

a. Since L is the completion of the L^p, it is clear that there is a map of Lie algebras $\mathrm{Hom}_k(K, k) \to L$. We need only show that this is an isomorphism of vector spaces.

It is a general fact that for any object N in a category $\mathcal{C}$ and any indexed diagram $F : J \to \mathcal{C}$ then

$$\mathrm{Hom}(\varinjlim F(-), N) \cong \varprojlim \mathrm{Hom}(F(-), N).$$

Applying this to the category of k-vector spaces enriched in k-vector spaces (meaning that the Hom-sets are k-vector spaces), we get

$$\mathrm{Hom}_k(K, k) = \mathrm{Hom}_k(\varinjlim K^p, L) = \varprojlim \mathrm{Hom}(K^p, k) = \varprojlim L^p = \hat{L}.$$

b. This follows as in a. above.
c. This follows again by the above. Since tensor products commute with colimits we have

$$U^c(K) \otimes U^c(K) = \varinjlim_{p,j} U^c_j(K^p) \otimes U^c_j(K^p).$$

Then

$$\begin{aligned}\mathrm{Hom}_k(U^c(K) \otimes U^c(K), k) =& \mathrm{Hom}_k(\varinjlim U^c_j(K^p) \otimes U^c_j(K^p), k)\\ =& \varprojlim_{p,j} U^j(L^p) \otimes U^j(L^p) = \hat{U}(L)\hat{\otimes}\hat{U}(L).\end{aligned}$$

□

The coalgebra $U^c(K)$ is a Hopf algebra with the shuffle product. It has unit η and counit ϵ. Denote by $\star$ the convolution product on this Hopf algebra, and by $\mathbf{1}$ the identity map. Write $\mathbf{1} = \eta \circ \epsilon + J$. The Euler idempotent

$$e : U^c(K) \to U^c(K)$$

is the convolution logarithm

$$e = \log^\star(\mathbf{1}) = \log^\star(\eta \circ \epsilon + J) = J - J^{\star 2}/2 + J^{\star 3}/3 - \cdots .$$

Proposition 5.3 *The image of $U^c(K) \xrightarrow{e} U^c(K)$ is K. This inclusion of $K \subseteq U^c(K)$ is a section of the natural map $U^c(K) \to K$.*

Proof This follows the same argument as Proposition 2.1. □

This gives a map $K \to U^c(K)$. Since $U^c(K)$ is a commutative algebra under the shuffle product, we get a map from the free commutative algebra $\mathrm{Sym}(K) \to U^c(K)$.

Proposition 5.4 *This map*

$$\psi : Sym(K) \xrightarrow{\cong} U^c(K) \tag{16}$$

is an isomorphism of commutative algebras. (We later denote the shuffle product by $\sqcup\!\sqcup$.*)*

Proof By Proposition 2.2 there is an isomorphism of coalgebras

$$U(L_p) \xrightarrow{\cong} \mathrm{Sym}^c(L_p)$$

and the filtrations on these coalgebras correspond. Hence we get an isomorphism

$$U^j(L_p) \xrightarrow{\cong} \mathrm{Sym}^{c,j}(L_p).$$

Dualizing this we get

$$\mathrm{Sym}_j(K^p) \xrightarrow{\cong} U^c_j(K^p).$$

Taking the colimits of this we get the statement. □

In $\mathrm{Hom}_k(U^c(K), k)$ there are two distinguished subsets. The *characters* are the algebra homomorphisms $\mathrm{Hom}_{Alg}(U^c(K), k)$. Via the isomorphism of Proposition 5.2 they corresponds to the grouplike elements of $\hat{U}(L)$. The *infinitesimal characters* are the linear maps $\alpha : U^c(K) \to k$ such that

$$\alpha(uv) = \epsilon(u)\alpha(v) + \alpha(u)\epsilon(v).$$

We denote these as $\mathrm{Hom}_{Inf}(U^c(K), k)$.

Lemma 5.5 *Via the isomorphism in Proposition* 5.2b. *these characters correspond naturally to the following:*

a. $Hom_{Inf}(U^c(K), k) \cong Hom_k(K, k) \cong L$.
b. $Hom_{Alg}(U^c(K), k) \cong G(\hat{U}(L))$.

Proof

a. The map $U^c(K) \xrightarrow{\phi} K$ from Proposition 5.3 has kernel $k \oplus U^c(K)^{\sqcup\!\sqcup 2}_+$, by Proposition 5.4 above, where $\sqcup\!\sqcup$ denotes the shuffle product. We then see that any linear map $K \to k$ induces by composition an infinitesimal character on $U^c(K)$. Conversely given an infinitesimal character $\alpha : U^c(K) \to k$ then both k and $U^c(K)^{\sqcup\!\sqcup 2}_+$ are seen to be in the kernel, and so such a map is induced from a linear map $K \to k$ by composition with ϕ.

b. That $s : U^c(K) \to k$ is an algebra homomorphism is equivalent to the commutativity of the diagram

$$\begin{array}{ccc} U^c(K) \otimes U^c(K) & \longrightarrow & U^c(K) \\ \downarrow{\scriptstyle s\otimes s} & & \downarrow{\scriptstyle s} \\ k \otimes_k k & \longrightarrow & k \end{array} \tag{17}$$

But this means that by the map

$$\begin{aligned} \hat{U}(L) &\to \hat{U}(L)\hat{\otimes}\hat{U}(L) \\ s &\mapsto s \otimes s. \end{aligned}$$

Conversely given a grouplike element $s \in \hat{U}(L)$, it corresponds by Proposition 5.2b. to $s : U^c(K) \to k$, and it being grouplike means precisely that the diagram (17) commutes. □

On $\mathrm{Hom}_k(U^c(K), k)$ we also have the convolution product, which we again denote by $\star$. Note that by the isomorphism in Proposition 5.2, this corresponds to the product on $\hat{U}(L)$. Let $\mathrm{Hom}_k(U^c(K), k)_+$ consist of the α with $\alpha(1) = 0$. We then get the exponential map (we write this map without a $\star$ superscript since it is a product on the dual space)

$$\mathrm{Hom}_k(U^c(K), k)_+ \xrightarrow{\exp} \epsilon + \mathrm{Hom}_k(U^c(K), k)_+$$

given by

$$\exp(\alpha) = \epsilon + \alpha + \alpha^{\star 2}/2! + \alpha^{\star 3}/3! + \cdots .$$

This is well defined since $U^c(K)$ is a conilpotent coalgebra and $\alpha(1) = 0$. Correspondingly we get

$$\epsilon + \mathrm{Hom}_k(U^c(K), k)_+ \xrightarrow{\log} \mathrm{Hom}_k(U^c(K), k)_+$$

given by

$$\log(\epsilon + \alpha) = \alpha - \frac{\alpha^{\star 2}}{2} + \frac{\alpha^{\star 3}}{3} - \cdots .$$

Lemma 5.6 *The maps*

$$Hom_k(U^c(K), k)_+ \underset{\log}{\overset{\exp}{\rightleftarrows}} \epsilon + Hom_k(U^c(K), k)_+$$

give inverse bijections. They restrict to the inverse bijections

$$Hom_{Inf}(U^c(K), k) \underset{\log}{\overset{\exp}{\rightleftarrows}} Hom_{Alg}(U^c(K), k).$$

Proof Using the identification of Proposition 5.2 the exp and log maps above correspond to the exp and log maps in Proposition 2.10. □

Since $\mathrm{Sym}(K)$ is the free symmetric algebra on K, there is a bijection $\mathrm{Hom}_{Alg}(\mathrm{Sym}(K), k) \xrightarrow{\cong} \mathrm{Hom}_k(K, k)$. The following shows that all the various maps correspond.

Proposition 5.7 *The following diagram commutes, showing that the various horizontal bijections correspond to each other:*

$$\begin{array}{ccc}
Hom_k(K,k) & \xrightarrow{\cong} & Hom_{Alg}(Sym(K),k) \\
\| & & \uparrow{\scriptstyle \psi^*} \\
Hom_k(K,k) & \xrightarrow{\exp} & Hom_{Alg}(U^c(K),k) \\
\cong\downarrow & & \downarrow\cong \\
L & \xrightarrow{\exp} & G(\hat{U}(L))
\end{array}$$

Proof That the lower diagram commutes is clear by the proof of Lemma 5.6. The middle (resp. top) map sends $K \to k$ to the unique algebra homomorphism ϕ (resp. ϕ') such that the following diagrams commute

$$\begin{array}{ccc} K & \longrightarrow & U^c(K), \\ & \searrow & \downarrow{\scriptstyle \phi} \\ & & k \end{array} \qquad \begin{array}{ccc} K & \longrightarrow & \mathrm{Sym}(K). \\ & \searrow & \downarrow{\scriptstyle \phi} \\ & & k \end{array}$$

Since the following diagram commutes where ψ is the isomorphism of algebras

$$\begin{array}{ccc} K & \longrightarrow & \mathrm{Sym}(K), \\ & \searrow & \downarrow{\scriptstyle \psi} \\ & & U^c(K) \end{array}$$

the commutativity of the upper diagram in the statement of the proposition follows. □

5.2 Actions of Endomorphisms

Let $E = \mathrm{End}_{Lie\ co}(K)$ be the endomorphisms of K as a Lie co-algebra, which also *respect the filtration* on K.

Proposition 5.8 *The Euler map in Proposition 5.3 is equivariant for the endomorphism action. Hence the isomorphism $\Psi : Sym(K) \to U^c(K)$ is equivariant for the action of the endomorphism group E.*

Proof The coproduct on $U^c(K)$ is clearly equivariant for E and similarly the product on $U^c(K)$ is equivariant, since $U^c(K)$ is a subalgebra of $T^c(K)$ for the shuffle product. Then if $f, g : U^c(K) \to U^c(K)$ are two equivariant maps, their convolution product $f \star g$ is also equivariant.

Since $\mathbf{1}$ and $\eta \circ \epsilon$ are equivariant for E, the difference $J = \mathbf{1} - \eta \circ \epsilon$ is so also. The Euler map $e = J - J^{\star 2}/2 + J^{\star 3}/3 - \cdots$ must then be equivariant for the action of E.

Since the image of the Euler map is K, the inclusion $K \hookrightarrow U^c(K)$ is equivariant also, and so is the map Ψ above. □

As a consequence of this the action of E on K induces an action on the dual Lie algebra L respecting its filtration. By Proposition 5.2 this again induces a diagram of actions of the following *sets*

$$\begin{array}{ccc} E \times \hat{U}(L) & \longrightarrow & \hat{U}(L) \\ \downarrow & & \downarrow \\ E \times L & \longrightarrow & L. \end{array} \tag{18}$$

5.2.1 The Free Lie Algebra

Now let $V = \oplus_{i \geq 1} V_i$ be a positively graded vector space with finite dimensional parts V_i. We consider the special case of the above that L is the completion $\widehat{\mathrm{Lie}}(V)$ of the free Lie algebra on V. Note that $\mathrm{Lie}(V)$ is a graded Lie algebra with finite dimensional graded parts. The enveloping algebra $U(\mathrm{Lie}(V))$ is the tensor algebra $T(V)$.

The graded dual vector space is $V^{\circledast} = \oplus V_i^*$ and the graded dual Lie co-algebra is $\mathrm{Lie}(V)^{\circledast}$. The Hopf algebra $U^c(\mathrm{Lie}(V)^{\circledast})$ is the shuffle Hopf algebra $T(V^{\circledast})$.

Since $\mathrm{Lie}(V)$ is the free Lie algebra on V, the endomorphisms E identifies as (note that here it is essential that we consider endomorphisms respecting the filtration)

$$\mathrm{End}_{\mathrm{Lie}\,co}(\mathrm{Lie}(V)^{\circledast}, \mathrm{Lie}(V)^{\circledast}) = \mathrm{Hom}_{\mathrm{Lie}}(\mathrm{Lie}(V), \widehat{\mathrm{Lie}}(V)). \tag{19}$$

This is a variety with coordinate ring $\mathcal{E}_V = \mathrm{Sym}(V \otimes \mathrm{Lie}(V)^{\circledast})$, which is a bialgebra. Furthermore the diagram (18) with $L = \widehat{\mathrm{Lie}}(V)$ in this case will be a

morphism of varieties: Both E, L and $\hat{U}(L)$ come with filtrations and all maps are given by polynomial maps. So we get a dual diagram of coordinate rings

$$\begin{array}{ccc} \mathrm{Sym}(\mathrm{Lie}(V)^{\circledast}) & \longrightarrow & \mathcal{E}_V \otimes \mathrm{Sym}(\mathrm{Lie}(V)^{\circledast}) \\ \downarrow & & \downarrow \\ \mathrm{Sym}(T(V)^{\circledast}) & \longrightarrow & \mathcal{E}_V \otimes \mathrm{Sym}(T(V)^{\circledast}) \end{array}.$$

But since the action of E is linear on $\widehat{\mathrm{Lie}}(V)$ and $\hat{T}(V)$, this gives a diagram

$$\begin{array}{ccc} \mathrm{Lie}(V)^{\circledast} & \longrightarrow & \mathcal{E}_V \otimes \mathrm{Lie}(V)^{\circledast} \\ \downarrow & & \downarrow \\ T^c(V^{\circledast}) & \longrightarrow & \mathcal{E}_V \otimes T^c(V^{\circledast}) \end{array},$$

and so the isomorphism $\mathrm{Sym}(\mathrm{Lie}(V)^{\circledast}) \xrightarrow{\cong} T^c(V^{\circledast})$ is an isomorphism of comodules over the algebra $\mathcal{E}_V$.

5.3 Baker-Campbell-Hausdorff on Coordinate Rings

The space K has a countable basis and so we may consider $\mathrm{Sym}(K)$ as the coordinate ring of the variety $L = \mathrm{Hom}_k(K, k)$. By the isomorphism $\psi : \mathrm{Sym}(K) \xrightarrow{\cong} U^c(K)$ of Proposition 5.4 we may think of $U^c(K)$ as this coordinate ring. Then also $U^c(K) \otimes_k U^c(K)$ is the coordinate ring of $L \times L$.

The coproduct (whose dual is the product on $\hat{U}(L)$)

$$U^c(K) \xrightarrow{\Delta_\bullet} U^c(K) \otimes_k U^c(K),$$

will then correspond to a morphism of varieties $L \times L \to L$. The following explains what it is.

Proposition 5.9 *The map $L \times L \to L$ given by*

$$(a, b) \mapsto \log^\bullet(\exp^\bullet(a) \bullet \exp^\bullet(b))$$

is a morphism of varieties, and on coordinate rings it corresponds to the coproduct

$$U^c(K) \xrightarrow{\Delta_\bullet} U^c(K) \otimes U^c(K).$$

This above product on L is the Baker-Campbell-Hausdorff product.

Example 5.10 Let $V = \oplus_{i\geq 1} V_i$ be a graded vector space with finite dimensional graded parts. Let $\mathrm{Lie}(V)$ be the free Lie algebra on V, which comes with a natural grading. The enveloping algebra $U(\mathrm{Lie}(V))$ is the tensor algebra $T(V)$. The dual Lie coalgebra is the graded dual $K = \mathrm{Lie}(V)^{\circledast}$, and $U^c(K)$ is the graded dual tensor coalgebra $T(V^{\circledast})$ which comes with the shuffle product. Thus the shuffle algebra $T(V^{\circledast})$ identifies as the *coordinate ring* of the Lie series, the completion $\widehat{\mathrm{Lie}}(V)$ of the free Lie algebra on V.

The coproduct on $T(V^{\circledast})$ is the deconcatenation coproduct. This can then be considered as an extremely simple codification of the Baker-Campbell-Hausdorff formula for Lie series in the completion $\widehat{\mathrm{Lie}}(V)$.

Proof If $X \to Y$ is a morphism of varieties and $A(Y) \xrightarrow{\phi} A(X)$ the corresponding homomorphism of coordinate rings, then the point p in X corresponding to the algebra homomorphism $A(X) \xrightarrow{p^*} k$ maps to the point q in Y corresponding to the algebra homomorphism $A(Y) \xrightarrow{q^*} k$ given by $q^* = \phi \circ p^*$.

Now given points a and b in $L = \mathrm{Hom}_k(K, k)$. They correspond to algebra homomorphisms from the coordinate ring $U^c(K) \xrightarrow{\tilde{a},\tilde{b}} k$, the unique such extending a and b, and these are $\tilde{a} = \exp(a)$ and $\tilde{b} = \exp(b)$. The pair $(a, b) \in L \times L$ corresponds to the homomorphism on coordinate rings

$$\exp(a) \otimes \exp(b) : U^c(K) \otimes U^c(K) \xrightarrow{\tilde{a}\otimes\tilde{b}} k \otimes_k k = k.$$

Now via the coproduct, which is the homomorphism of coordinate rings,

$$U^c(K) \xrightarrow{\Delta_\bullet} U^c(K) \otimes U^c(K)$$

this maps to the algebra homomorphism $\exp(a) \bullet \exp(b) : U^c(K) \to k$. This is the algebra homomorphism corresponding to the following point in L:

$$\log^\bullet(\exp(a) \bullet \exp(b)) : K \to k.$$

□

5.4 Filtered Pre- and Post-Lie Algebras with Finite Dimensional Quotients

We now assume that the filtered quotients P/P^{p+1}, which again are post-Lie algebras, are all finite dimensional. Let their duals be $Q_p = \mathrm{Hom}_k(P/P^{p+1}, k)$ and $Q = \varinjlim_p Q_p$, which is a post-Lie coalgebra. We shall assume $P = \hat{P}$ is complete with respect to this filtration. Then $P = \mathrm{Hom}(Q, k)$, and $\mathrm{Sym}(Q)$ is the

coordinate ring of P. There are two Lie algebra structures on P, given by $[,]$ and $[\![,]\!]$ of Definition 3.1. These correspond to the products $\bullet$ and $*$ on the enveloping algebra of P. We shall use the first product $\bullet$, giving the coproduct $\Delta_\bullet$ on $U^c(Q)$. For this coproduct Proposition 5.4 gives an isomorphism

$$\psi_\bullet : \mathrm{Sym}(Q) \xrightarrow{\cong} U^c(Q). \tag{20}$$

Due to the formula in Proposition 3.6 the product

$$P \times P \xrightarrow{\sharp} P$$

on each quotient P/P^i, is given by polynomial expressions. It thus corresponds to a homomorphism of coordinate rings

$$\mathrm{Sym}(Q) \xrightarrow{\Delta_\sharp} \mathrm{Sym}(Q) \otimes \mathrm{Sym}(Q). \tag{21}$$

Proposition 5.11 *Via the isomorphism $\psi_\bullet$ in (20) the coproduct $\Delta_\sharp$ above corresponds to the coproduct*

$$U^c(Q) \xrightarrow{\Delta_*} U^c(Q) \otimes U^c(Q),$$

which is the dual of the product $*$ on $U(P)$.

Remark 5.12 In order to identify the homomorphism of coordinate rings as the coproduct Δ_* it is essential that one uses the isomorphism $\psi_\bullet$ of (20). If one uses another isomorphism $\mathrm{Sym}(Q) \xrightarrow{\cong} U^c(Q)$ like the isomorphism ψ_* derived from the coproduct Δ_*, the statement is not correct. See also the end of the last remark below.

Remark 5.13 (The Connes-Kreimer Hopf algebra) For the free pre-Lie algebra T_C (see the next Sect. 6) this identifies the Connes-Kreimer Hopf algebra $\mathcal{H}_{CK}$ as the *coordinate ring* $\mathrm{Sym}(T_C^{(\circledast)})$ of the Butcher series $\hat{T}_C$ under the Butcher product.

As a variety the Butcher series $\hat{T}_C$ is endowed with the Zariski topology, and the Butcher product is continuous for this topology. In [1] another finer topology on $\hat{T}_C$ is considered when the field $k = \mathbb{R}$ or $\mathbb{C}$.

Remark 5.14 (The MKW Hopf algebra) For the free post-Lie algebra P_C (see Sect. 6) it identifies the MKW Hopf algebra $T(\mathrm{OT}_C^{(\circledast)})$ as the *coordinate ring* $\mathrm{Sym}(\mathrm{Lie}(\mathrm{OT}_C)^{(\circledast)})$ of the Lie-Butcher series $\hat{P}_C = \widehat{\mathrm{Lie}}(\mathrm{OT}_C)$. A (principal) Lie-Butcher series $\ell \in \hat{P}_C$ corresponds to an element $\mathrm{Lie}(\mathrm{OT}_C)^{\circledast} \xrightarrow{\ell} k$. This lifts via the isomorphism $\psi_\bullet$ of (20) to a character of the shuffle algebra $T(\mathrm{OT}_C^{(\circledast)}) \xrightarrow{\tilde{\ell}} k$. That the lifting from (principal) LB series to character of the MKW Hopf algebra must be done using the inclusion $\mathrm{Lie}(\mathrm{OT}_C)^{\circledast} \hookrightarrow T(\mathrm{OT}_C^{(\circledast)})$ via the Euler map of

Proposition 5.3 associated to the coproduct $\Delta_\bullet$, is a technical point which has not been made explicit previously.

Proof of Proposition 5.11 Given points $a, b \in P$. They correspond to linear maps $Q \xrightarrow{a,b} k$. Via the isomorphism $\psi_\bullet$ these extend to algebra homomorphisms $U^c(Q) \xrightarrow{\tilde{a},\tilde{b}} k$, where $\tilde{a} = \exp^\bullet(a)$ and $\tilde{b} = \exp^\bullet(b)$. The pair $(a, b) \in P \times P$ then corresponds to a homomorphism of coordinate rings

$$\exp^\bullet(a) \otimes \exp^\bullet(b) : U^c(Q) \otimes U^c(Q) \xrightarrow{\tilde{a}\otimes\tilde{b}} k \otimes_k k = k.$$

Now via the coproduct associated to $*$, which is the homomorphism of coordinate rings,

$$U^c(Q) \xrightarrow{\Delta_*} U^c(Q) \otimes U^c(Q)$$

this maps to the algebra homomorphism $\exp^\bullet(a) * \exp^\bullet(b) : U^c(Q) \to k$. This is the algebra homomorphism corresponding to the following point in P:

$$\log^\bullet(\exp^\bullet(a) * \exp^\bullet(b)) : Q \to k.$$

□

6 Free Pre- and Post-Lie Algebras

This section recalls free pre- and post-Lie algebras, and the notion of substitution in these algebras. We also briefly recall the notions of Butcher and Lie-Butcher series.

6.1 Free Post-Lie Algebras

We consider the set of rooted planar trees, or ordered trees:

$$\mathrm{OT} = \{\bullet, \ldots\},$$

and let $k\mathrm{OT}$ be the k-vector space with these trees as basis. It comes with an operation $\triangleright$, called *grafting*. For two trees t and s we define $t \triangleright s$ to be the sum of all trees obtained by attaching the root of t with a new edge onto a vertex of s, with this new edge as the leftmost branch into the vertex of s.

If C is a set, we can color the vertices of OT with the elements of C. We then get the set OT_C of labelled planar trees. The *free post-Lie algebra* on C is the free

Lie algebra $P_C = \mathrm{Lie}(\mathrm{OT}_C)$ on the set of C-labelled planar trees. The grafting operation is extended to the free Lie algebra $\mathrm{Lie}(\mathrm{OT}_C)$ by using the relations from Definition 3.1. Note that P_C has a natural grading by letting $P_{C,d}$ be the subspace generated by all bracketed expressions of trees with a total number of d vertices. In particular P_C is filtered.

The enveloping algebra of P_C identifies as the tensor algebra $T(\mathrm{OT}_C)$. It was introduced and studied in [25], see also [23] for more on the computational aspect in this algebra. Its completion identifies as

$$\hat{T}(\mathrm{OT}_C) = \prod_{d\geq 0} T(\mathrm{OT}_C)_d.$$

6.2 Free Pre-Lie Algebras

Here we consider instead (non-ordered) rooted trees

$$T = \{\bullet, \cdots\}.$$

On the vector space kT we can similarly define grafting $\triangleright$. Given a set C we get the set T_C of trees labelled by C. The free pre-Lie algebra is $A_C = kT_C$, [5]. Its enveloping algebra is the symmetric algebra $\mathrm{Sym}(T_C)$, called the Grossman-Larson algebra, and comes with the ordinary symmetric product $\cdot$ and the product $*$, [26].

6.3 Butcher and Lie-Butcher Series

Recall the pre-Lie algebra $\mathcal{X}\mathbb{R}^n$ of vector fields from Example 3.2, and the corresponding power series $\mathcal{X}\mathbb{R}^n[\![h]\!]$. Let $f \in \mathcal{X}\mathbb{R}^n$ be a vector field and $A_\bullet$ the free pre-Lie algebra on one generator $\bullet$. By sending $\bullet \mapsto f$ we get a homomorphism of pre-Lie algebras $A_\bullet \to \mathcal{X}\mathbb{R}^n$ which sends a tree τ to the associated *elementary differential* f^τ, see [15, Section III.1]. If $f \in \mathcal{X}\mathbb{R}^n[\![h]\!]$ we similarly get a homomorphism of pre-Lie algebras $A_\bullet \to \mathcal{X}\mathbb{R}^n[\![h]\!]$. The natural grading on $A_\bullet$ by number of vertices of trees $|\tau|$ of a tree τ, gives a filtration and we get a map of complete pre-Lie algebras $\hat{A}_\bullet \to \mathcal{X}\mathbb{R}^n[\![h]\!]$. If we let $\bullet \to f \cdot h$ where $f \in \mathcal{X}\mathbb{R}^n$ is a vector field, then

$$\sum_{\tau\in T} \alpha(\tau)\tau \mapsto \sum_{\tau\in T} \alpha(\tau) f^\tau h^{|\tau|}$$

and the latter is called a *Butcher series*. Often this terminology is also used about the abstract form on the left side above.

In the general setting of a Lie group G. By Example 3.3, $\mathcal{X}G$ is a post-Lie algebra, and so is also the power series $\mathcal{X}G[\![h]\!]$. Let $f \in \mathcal{X}G$ be a vector field and $P_\bullet$ the free post-Lie algebra on one generator $\bullet$. By sending $\bullet \mapsto f$ we get a homomorphism of post-Lie algebras $P_\bullet \to \mathcal{X}G$ which sends a tree τ to the associated *elementary differential* f^τ, see [18, Subsection 2.2]. We also get a map of enveloping algebras $T(OT_\bullet) \to U(\mathcal{X}G)$ which sends a forest ω to an associated differential operator f^ω. The natural grading on $P_\bullet$ by number of vertices of trees $|\tau|$ of a tree τ, gives a filtration. Sending $\bullet \mapsto f \cdot h$ we get a homomorphism of complete post-Lie algebras $\hat{P}_\bullet \to \mathcal{X}G[\![h]\!]$. The image of an element from $\hat{P}_\bullet$ is a *Lie-Butcher series* in $\mathcal{X}G[\![h]\!]$. Note that there is however not a really natural basis for $P_\bullet = \mathrm{Lie}(OT_\bullet)$. Therefore one usually consider instead the map from the completed enveloping algebra to the power series of differential operators ($F_\bullet$ below denotes ordered forests of ordered trees)

$$\hat{T}(OT_\bullet) \to U(\mathcal{X}G[\![h]\!])$$
$$\sum_{\omega \in F_\bullet} \beta(\omega)\omega \mapsto \sum_{\omega \in F_\bullet} \beta(\omega) f^\omega h^{|\omega|}$$

and the latter is a *Lie-Butcher series*. The abstract form to the left is also often called a LB-series.

6.4 Substitution

In the above setting, we get by Sect. 3.2 a commutative diagram of flow maps

$$\begin{array}{ccc} \hat{P}_{\bullet,\mathrm{field}} & \xrightarrow{\Phi_P} & \hat{P}_{\bullet,\mathrm{flow}} \\ \downarrow & & \downarrow \\ \mathcal{X}G[\![h]\!]_{\mathrm{field}} & \xrightarrow{\Phi_{\mathcal{X}G}} & \mathcal{X}G[\![h]\!]_{\mathrm{flow}}. \end{array}$$

The field f is mapped to the flow $\Phi_{\mathcal{X}G}(f)$. By perturbing the vector field $f \to f + \delta$, it is sent to a flow $\Phi_{\mathcal{X}M}(f + \delta)$. We assume the perturbation δ is expressed in terms of the elementary differentials of f, and so it comes from a perturbation $\bullet \to \bullet + \delta' = s$. Since $\mathrm{Hom}(\bullet, P_\bullet) = \mathrm{End}_{\mathrm{postLie}}(P_\bullet)$ this gives an endomorphism of the post-Lie algebra. We are now interested in the effect of this endomorphism on the flow, called *substitution* of the perturbed vector field, and we are interested in the algebraic aspects of this action. We study this for the free post-Lie algebra P_C, but most of the discussions below are of a general nature, and applies equally well to the free pre-Lie algebra, and generalises the results of [4].

7 Action of the Endomorphism Group and Substitution in Free Post-Lie Algebras

Substitution in the free pre-Lie or free post-Lie algebras on one generator gives, by dualizing, the operation of co-substitution in their coordinate rings, which are the Connes-Kreimer and the MKW Hopf algebras. In [4] they show that co-substitution on the Connes-Kreimer algebra is governed by a bialgebra $\mathcal{H}$ such that the Connes-Kreimer algebra $\mathcal{H}_{CK}$ is a comodule bialgebra over this bialgebra $\mathcal{H}$. Moreover $\mathcal{H}_{CK}$ and $\mathcal{H}$ are isomorphic as commutative algebras. This is the notion of two bialgebras in *cointeraction*, a situation further studied in [12, 20], and [11].

In this section we do the analog for the MKW Hopf algebra, and in a more general setting, since we consider free pre- and post-Lie algebras on any finite number of generators. In this case $\mathcal{H}_{CK}$ and $\mathcal{H}$ are no longer isomorphic as commutative algebras. As we shall see the situation is understood very well by using the algebraic geometric setting and considering the MKW Hopf algebra as the coordinate ring of the free post-Lie algebra. The main results of [4] also follow, and are understood better, by the approach we develop here.

7.1 A Bialgebra of Endomorphisms

Let C be a finite dimensional vector space over the field k, and P_C the free post-Lie algebra on this vector space. It is a graded vector space $P_C = \bigoplus_{d\geq 1} P_{C,d}$ graded by the number of vertices in bracketed expressions of trees, and so has finite dimensional graded pieces. It has a graded dual

$$P_C^{\circledast} = \oplus_d \mathrm{Hom}_k(P_{C,d}, k).$$

Let $\{l\}$ be a basis for P_C. It gives a dual basis $\{l^*\}$ for $P_C^{\circledast}$. The dual of $P_C^{\circledast}$ is the completion

$$\hat{P}_C = \mathrm{Hom}_k(P_C^{\circledast}, k) = \varprojlim_d P_{C,\leq d}.$$

It is naturally a post-Lie algebra and comes with a decreasing filtration $\hat{P}_C^{d+1} = \ker(\hat{P}_C \to P_{C,\leq d})$.

Due to the freeness of P_C we have:

$$\mathrm{Hom}_k(C, P_C) = \mathrm{Hom}_{\mathrm{postLie}}(P_C, P_C) = \mathrm{End}_{\mathrm{postLie}}(P_C).$$

Denote the above vector space as E_C. If we let $\{c\}$ be a basis for C, the graded dual $E_C^{\circledast} = C \otimes_k P_C^{\circledast}$ has a basis $\{a_c(l) := c \otimes l^*\}$.

The dual of $E_C^{\circledast}$ is $\hat{E}_C = \mathrm{Hom}_k(E_C^{\circledast}, k)$ which may be written as $C^* \otimes_k \hat{P}_C$. This is an affine space with coordinate ring

$$\mathcal{E}_C := \mathrm{Sym}(E_C^{\circledast}) = \mathrm{Sym}(\mathrm{Hom}_k(C, P_C)^{\circledast}) = \mathrm{Sym}(C \otimes_k P_C^{\circledast}).$$

The filtration on $\hat{P}_C$ induces also a filtration on $\hat{E}_C$.

A map of post-Lie algebras $\phi : P_C \to \hat{P}_C$ induces a map of post-Lie algebras $\hat{\phi} : \hat{P}_C \to \hat{P}_C$. We then get the inclusion

$$\hat{E}_C = \mathrm{Hom}_{\mathrm{postLie}}(P_C, \hat{P}_C) \subseteq \mathrm{Hom}_{\mathrm{postLie}}(\hat{P}_C, \hat{P}_C).$$

If $\phi, \psi \in \hat{E}_C$, we get a composition $\psi \circ \hat{\phi}$, which we by abuse of notation write as $\psi \circ \phi$. This makes $\hat{E}_C$ into a monoid of affine varieties:

$$\hat{E}_C \times \hat{E}_C \xrightarrow{\circ} \hat{E}_C.$$

It induces a homomorphism on coordinate rings:

$$\mathcal{E}_C \xrightarrow{\Delta_\circ} \mathcal{E}_C \otimes \mathcal{E}_C.$$

This coproduct is coassociative, since $\circ$ on $\hat{E}_C$ is associative. Thus $\mathcal{E}_C$ becomes a bialgebra.

Note that when $C = \langle \bullet \rangle$ is one-dimensional, then

$$\mathcal{E}_\bullet = \mathrm{Sym}(P_\bullet^{\circledast}) \cong T^c(\mathrm{OT}_\bullet^{\circledast})$$

as *algebras*, using Proposition 5.4. The coproduct $\Delta_\circ$ considered on the shuffle algebra is, however, neither deconcatenation nor the Grossman-Larson coproduct. For the free pre-Lie algebra $A_\bullet$ instead of $P_\bullet$, a description of this coproduct is given in [4, Section 4.1/4.2].

7.1.1 Hopf Algebras of Endomorphisms

The augmentation map $P_C \to C$ gives maps

$$\mathrm{Hom}_k(C, P_C) \to \mathrm{Hom}_k(C, C)$$

and dually

$$\mathrm{Hom}_k(C, C)^{\circledast} \subseteq \mathrm{Hom}_k(C, P_C)^{\circledast} \cong C \otimes_k P_C^{\circledast}.$$

Recall that $a_c(d)$ are the basis elements of $\mathrm{Hom}_k(C, C)^{\circledast}$ (the coordinate functions on $\mathrm{Hom}_k(C, C)$), where c and d range over a basis for C. We can then invert $D =$

$\det(a_c(d))$ in the coordinate ring $\mathcal{E}_C$. This gives a Hopf algebra $\mathcal{E}_C^{\times}$ which is the localized ring $(\mathcal{E}_C)_D$. Another possibility is to divide $\mathcal{E}_C$ by the ideal generated by $D-1$. This gives a Hopf algebra $\mathcal{E}_C^{\mathrm{l}} = \mathcal{E}_C/(D-1)$. A third possibility is to divide $\mathcal{E}_C$ out by the ideal generated by the $a_c(d) - \delta_{c,d}$. This gives a Hopf algebra $\mathcal{E}_C^{\mathrm{ld}}$. In the case $C = \{\bullet\}$ and $P_\bullet$ is replaced with the free pre-Lie alegbra $A_\bullet$, both the latter cases give the Hopf algebra $\mathcal{H}$ in [4, Subsection 4.1/4.2].

7.2 The Action on the Free Post-Lie Algebra

The monoid E_C acts on P_C, and $\hat{E}_C$ acts on $\hat{P}_C$. So we get a morphism of affine varieties

$$\hat{E}_C \times \hat{P}_C \xrightarrow{\star} \hat{P}_C \tag{22}$$

called *substitution*.

Let $\mathcal{H}_C = \mathrm{Sym}(P_C^{\circledast})$ be the coordinate ring of $\hat{P}_C$. We get a homomorphism of coordinate rings called *co-substitution*

$$\mathcal{H}_C \xrightarrow{\Delta_\star} \mathcal{E}_C \otimes \mathcal{H}_C. \tag{23}$$

Note that the map in (22) is linear in the second factor so the algebra homomorphism (23) comes from a linear map

$$P_C^{\circledast} \to \mathcal{E}_C \otimes P_C^{\circledast}.$$

The action $\star$ gives a commutative diagram

$$\begin{array}{ccc} \hat{E}_C \times \hat{E}_C \times \hat{P}_C & \xrightarrow{1\times\star} & \hat{E}_C \times \hat{P}_C \\ {\scriptstyle \circ\times\mathbf{1}}\downarrow & & \downarrow{\scriptstyle \star} \\ \hat{E}_C \times \hat{P}_C & \xrightarrow{\star} & \hat{P}_C \end{array}$$

which dually gives a diagram

$$\begin{array}{ccc} \mathcal{E}_C \otimes \mathcal{E}_C \otimes \mathcal{H}_C & \longleftarrow & \mathcal{E}_C \otimes \mathcal{H}_C \\ \uparrow & & \uparrow \\ \mathcal{E}_C \otimes \mathcal{H}_C & \longleftarrow & \mathcal{H}_C. \end{array}$$

This makes $\mathcal{H}_C$ into a comodule over $\mathcal{E}_C$, in fact a comodule algebra, since all maps are homomorphisms of algebras. The Butcher product $\natural$ on $\hat{P}_C$ is dual to

the coproduct $\Delta_\star : \mathcal{H}_C \to \mathcal{H}_C \otimes \mathcal{H}_C$ by Proposition 5.11. Since $\hat{E}_C$ gives an endomorphism of post-Lie algebra we have for $a \in \hat{E}_C$ and $u, v \in \hat{P}_C$:

$$a \star (u \sharp v) = (a \star u) \sharp (a \star v).$$

In diagrams

$$\begin{array}{ccccc} \hat{E}_C \times \hat{E}_C \times \hat{P}_C \times \hat{P}_C & \xrightarrow{\mathbf{1}\times\tau\times\mathbf{1}} & \hat{E}_C \times \hat{P}_C \times \hat{E}_C \times \hat{P}_C & \xrightarrow{\star\times\star} & \hat{P}_C \times \hat{P}_C \\ {\scriptstyle \mathrm{diag}\times\mathbf{1}\times\mathbf{1}}\uparrow & & & & \downarrow{\scriptstyle \sharp} \\ \hat{E}_C \times \hat{P}_C \times \hat{P}_C & \xrightarrow{\mathbf{1}\times\sharp} & \hat{E}_C \times \hat{P}_C & \xrightarrow{\star} & \hat{P}_C \end{array}$$

which dually gives a diagram

$$\begin{array}{ccccc} \mathcal{E}_C \otimes \mathcal{E}_C \otimes \mathcal{H}_C \otimes \mathcal{H}_C & \xleftarrow{\mathbf{1}\otimes\tau\otimes\mathbf{1}} & \mathcal{E}_C \otimes \mathcal{H}_C \otimes \mathcal{E}_C \otimes \mathcal{H}_C & \xleftarrow{\Delta_\star\otimes\Delta_\star} & \mathcal{H}_C \otimes \mathcal{H}_C \\ \downarrow & & & & \uparrow{\scriptstyle \Delta_*} \\ \mathcal{E}_C \otimes \mathcal{H}_C \otimes \mathcal{H}_C & \xleftarrow{\mathbf{1}\otimes\Delta_*} & \mathcal{E}_C \otimes \mathcal{H}_C & \xleftarrow{\Delta_\star} & \mathcal{H}_C. \end{array}$$

This makes $\mathcal{H}_C$ into a comodule Hopf algebra over $\mathcal{E}_C$. We also have

$$a \star (u \rhd v) = (a \star u) \rhd (a \star v)$$

giving corresponding commutative diagrams, making $\mathcal{H}_C$ into a comodule algebra over $\mathcal{E}_C$.

7.2.1 The Identification with the Tensor Algebra

The tensor algebra $T(\mathrm{OT}_C)$ is the enveloping algebra of $P_C = \mathrm{Lie}(OT_C)$. The endomorphism of post-Lie co-algebras $\mathrm{End}_{\text{postLie-co}}(P_C^{\circledast})$ identifies by Eq. (19) as $\hat{E}_C = \mathrm{Hom}_{\text{postLie}}(C, \hat{P}_C)$. It is an endomorphism submonoid of $\mathrm{End}_{\mathrm{Lie}}(P_C^{\circledast})$

By Sect. 5.2.1 the isomorphism $\mathcal{H}_C = \mathrm{Sym}(P_C^{\circledast}) \xrightarrow{\cong} T^c(\mathrm{OT}_C^{\circledast})$ is equivariant for the action of $\hat{E}_C$ and induces a commutative diagram

$$\begin{array}{ccc} \mathcal{H}_C & \xrightarrow{\Delta_\star} & \mathcal{E}_C \otimes \mathcal{H}_C \\ \cong\downarrow & & \downarrow\cong \\ T^c(\mathrm{OT}_C^{\circledast}) & \xrightarrow{\Delta_{T,\star}} & \mathcal{E}_C \otimes T^c(\mathrm{OT}_C^{\circledast}) \end{array} \tag{24}$$

Thus all the statements above in Sect. 7.2 may be phrased with $T^c(\mathrm{OT}_C^{\circledast})$ instead of $\mathcal{H}_C$ as comodule over $\mathcal{E}_C$.

7.3 The Universal Substitution

Let K be a commutative k-algebra. We then get $P^{\circledast}_{C,K} = K \otimes_k P^{\circledast}_C$, and correspondingly we get

$$E^{\circledast}_{C,K}, \quad \mathcal{H}_{C,K} = \mathrm{Sym}(P^{\circledast}_{C,K}), \quad \mathcal{E}_{C,K} = \mathrm{Sym}(E^{\circledast}_{C,K}).$$

Let the completion $\hat{P}_{C,K} = \mathrm{Hom}(P^{\circledast}_{C,K}, K)$. (Note that this is not $K \otimes_k \hat{P}_C$ but rather larger than this.) Similarly we get $\hat{E}_{C,K}$. The homomorphism of coordinate rings $\mathcal{H}_{C,K} \to \mathcal{E}_{C,K} \otimes_K \mathcal{H}_{C,K}$ corresponds to a map of affine K-varieties (see Remark 4.2)

$$\hat{E}_{C,K} \times \hat{P}_{C,K} \to \hat{P}_{C,K}. \tag{25}$$

A K-point A in the affine variety $\hat{E}_{C,K}$ then corresponds to an algebra homomorphism $\mathcal{E}_{C,K} \xrightarrow{A^*} K$, and K-points $p \in \hat{P}_{C,K}$ corresponds to algebra homomorphisms $\mathcal{H}_{C,K} \xrightarrow{p^*} K$.

In particular the map obtained from (25), using $A \in \hat{E}_{C,K}$:

$$\hat{P}_{C,K} \xrightarrow{A_\star} \hat{P}_{C,K} \tag{26}$$

corresponds to the morphism on coordinate rings

$$\mathcal{H}_{C,K} \to \mathcal{H}_{C,K} \otimes_K \mathcal{E}_{C,K} \xrightarrow{\mathbf{1}\otimes A^*} \mathcal{H}_{C,K} \otimes_K K = \mathcal{H}_{C,K} \tag{27}$$

which due to (26) being linear, comes from a K-linear map

$$P^{\circledast}_{C,K} \to P^{\circledast}_{C,K}.$$

Now we let K be the commutative algebra $\mathcal{E}_C = \mathrm{Sym}(E^{\circledast}_C)$. Then

$$\mathcal{E}_{C,K} = K \otimes_k \mathrm{Sym}(E^{\circledast}_C) = \mathrm{Sym}(E^{\circledast}_C) \otimes \mathrm{Sym}(E^{\circledast}_C).$$

There is a canonical algebra homomorphism

$$\mathcal{E}_{C,K} \xrightarrow{\mu} K \tag{28}$$

which is simply the product

$$\mathrm{Sym}(E^{\circledast}_C) \otimes_k \mathrm{Sym}(E^{\circledast}_C) \xrightarrow{\mu} \mathrm{Sym}(E^{\circledast}_C).$$

Definition 7.1 Corresponding to the algebra homomorphism μ of (28) is the point U in $\hat{E}_{C,K} = \mathrm{Hom}_k(C_K, \hat{P}_{C,K})$. This is the *universal* map (here we use the completed tensor product):

$$C \to C \otimes (P_C^{\circledast} \hat{\otimes} P_C) \tag{29}$$

sending

$$c \mapsto c \otimes \sum_{\substack{l \text{ basis} \\ \text{element of } P_C}} l^* \otimes l = \sum_l a_c(l) \otimes l$$

Using this, (26) becomes the *universal* substitution, the K-linear map

$$\hat{P}_{C,K} \xrightarrow{U_\star} \hat{P}_{C,K}.$$

Let $H = \mathrm{Hom}(C, P_C)^{\circledast}$, the degree one part of $K = \mathcal{E}_C$, and $P_{C,H} = H \otimes_k P_C$. Note that the universal map (29) is a map from C to $\hat{P}_{C,H}$.

If $a \in \hat{E}_C$ is a specific endomorphism, it corresponds to an algebra homomorphism (character)

$$K = \mathcal{E}_C \xrightarrow{\alpha} k$$
$$a_c(l) = c \otimes l^* \mapsto \alpha(c \otimes l^*).$$

Then $U_\star$ induces the substitution $\hat{P}_C \xrightarrow{a\star} \hat{P}_C$ by sending each coefficient $a_c(l) \in K$ to $\alpha(c \otimes l^*) \in k$.

The co-substitution $\mathcal{H}_C \xrightarrow{\Delta_\star} \mathcal{E}_C \otimes \mathcal{H}_C$ of (23) induces a homomorphism

$$\mathcal{E}_C \otimes \mathcal{H}_C \to \mathcal{E}_C \otimes \mathcal{E}_C \otimes \mathcal{H}_C \to \mathcal{E}_C \otimes \mathcal{H}_C$$

which is seen to coincide with the homomorphism (27) when $K = \mathcal{E}_C$. The universal substitution therefore corresponds to the map on coordinate rings which is the co-substitution map, suitably lifted.

Recall that the tensor algebra $T(\mathrm{OT}_C)$ identifies as the forests of ordered trees OF_C. We may then write $T^c(\mathrm{OT}_C^{\circledast}) = \mathrm{OF}_C^{\circledast}$. By the diagram (24) the co-substitution $\mathcal{H}_{C,K} \xrightarrow{\Delta_\star} \mathcal{H}_{C,K}$ identifies as a map $\mathrm{OF}_{C,K}^{\circledast} \xrightarrow{U_\star^T} \mathrm{OF}_{C,K}^{\circledast}$ and we get a commutative diagram and its dual

$$\begin{array}{ccc} \mathrm{OF}^{\circledast}_{C,K} & \xrightarrow{U^T_\star} & \mathrm{OF}^{\circledast}_{C,K} \\ \downarrow & & \downarrow \\ P^{\circledast}_{C,K} & \xrightarrow{U_\star} & P^{\circledast}_{C,K} \end{array}, \qquad \begin{array}{ccc} \hat{P}_{C,K} & \xrightarrow{U_\star} & \hat{P}_{C,K} \\ \downarrow & & \downarrow \\ \hat{\mathrm{OF}}_{C,K} & \xrightarrow{U_\star} & \hat{\mathrm{OF}}_{C,K} \end{array}.$$

We may restrict this to ordered trees and get

$$\mathrm{OF}^{\circledast}_{C,K} \xrightarrow{\overline{U}^T_\star} \mathrm{OT}^{\circledast}_{C,K}, \quad \hat{\mathrm{OT}}_{C,K} \xrightarrow{\overline{U}_\star} \hat{\mathrm{OF}}_{C,K}.$$

We may also restrict and get

$$C_K \to \hat{P}_{C,K} \to \hat{\mathrm{OF}}_{C,K},$$

with dual map

$$U^t : \mathrm{OF}^{\circledast}_{C,K} \to P^{\circledast}_{C,K} \to C^*_K \tag{30}$$

For use in Sect. 7.4.1, note that (29) sends C to $\hat{P}_{C,H}$ where $H = \mathrm{Hom}_k(C, P_C)^{\circledast} \subseteq K$ is the graded dual of E_C. A consequence is that $\mathrm{OF}^{\circledast}_C \subseteq \mathrm{OF}^{\circledast}_{C,K}$ is mapped to $C^*_H \subseteq C^*_K$ by U^t.

7.4 Recursion Formula

The universal substitution is described in [18], and we recall it. By attaching the trees in a forest to a root $c \in C$, there is a natural isomorphism

$$\mathrm{OT}_C \cong \mathrm{OF}_C \otimes C$$

and dually

$$\mathrm{OF}^{\circledast}_C \otimes C^* \xrightarrow{\cong} \mathrm{OT}^{\circledast}_C \tag{31}$$

Here we denote the image of $\omega \otimes \rho$ as $\omega \curvearrowright \rho$.

Proposition 7.2 ([18]) *The following gives a partial recursion formula for $\overline{U}^T_\star$, the universal co-substitution followed by the projection onto the dual ordered trees:*

$$\overline{U}^T_\star(\omega) = \sum_{\Delta_\rhd(\omega)} U^T_\star(\omega^{(1)}) \curvearrowright U^t(w^{(2)}).$$

Proof Recall the following general fact. Two maps $V \xrightarrow{\phi} W$ and $W^* \xrightarrow{\psi} V^*$ are dual iff for all $v \in V$ and $w^* \in W^*$ the pairings

$$\langle v, \psi(w^*)\rangle = \langle \phi(v), w^*\rangle.$$

We apply this to $\phi = \overline{U_\star}$ and

$$\psi : \mathrm{OF}^{\circledast}_{C,K} \xrightarrow{\Delta_\triangleright} \mathrm{OF}^{\circledast}_{C,K} \otimes \mathrm{OF}^{\circledast}_{C,K} \xrightarrow{U_\star^T \otimes U^t} \mathrm{OF}^{\circledast}_{C,K} \otimes C^*_K \xrightarrow{\curvearrowright} \mathrm{OT}^{\circledast}_{C,K}.$$

We must then show that

$$\sum_{\Delta_\triangleright(\omega)} \langle t, U_\star^T(\omega^{(1)}) \curvearrowright U^t(w^{(2)})\rangle = \langle \overline{U}_\star(t), \omega\rangle$$

So let $t = f \triangleright c$. Using first the above fact on the map (31) and its dual:

$$\begin{aligned}
\sum_{\Delta_\triangleright(\omega)} \langle t, U_\star^T(\omega^{(1)}) \curvearrowright U^t(w^{(2)})\rangle &= \sum_{\Delta_\triangleright(\omega)} \langle f \otimes c, U_\star^T(\omega^{(1)}) \otimes U^t(\omega^{(2)})\rangle \\
&= \sum_{\Delta_\triangleright(\omega)} \langle f, U_\star^T(\omega^{(1)})\rangle \cdot \langle c, U^t(\omega^{(2)})\rangle \\
&= \sum_{\Delta_\triangleright(\omega)} \langle U_\star(f), \omega^{(1)}\rangle \cdot \langle U(c), \omega^{(2)}\rangle \\
&= \langle U_\star(f) \otimes U(c), \Delta_\triangleright(\omega)\rangle \\
&= \langle U_\star(f) \triangleright U(c), \omega\rangle \\
&= \langle \overline{U}_\star(f \triangleright c), \omega\rangle = \langle \overline{U}_\star(t), \omega\rangle
\end{aligned}$$

□

We now get the general recursion formula, Theorem 3.7, in [18].

Proposition 7.3

$$U_\star^T(\omega) = \sum_{\Delta_\bullet(\omega)} U_\star^T(\omega_1) \cdot \overline{U}_\star^T(\omega_2).$$

Proof Given a forest $f \cdot t$ where t is a tree. We will show

$$\langle U_\star(ft), \omega\rangle = \sum_{\Delta_\bullet(\omega)} \langle ft, U_\star^T(\omega_1) \cdot \overline{U}_\star^T(\omega_2)\rangle.$$

We have:

$$\langle U_\star(ft), \omega\rangle = \langle U_\star(f)\cdot U_\star(t), \omega\rangle.$$

Since concatenation and deconcatenation are dual maps, this is

$$\begin{aligned}
&= \sum_{\Delta_\bullet(\omega)} \langle U_\star(f)\otimes U_\star(t), \omega_1\otimes\omega_2\rangle \\
&= \sum_{\Delta_\bullet(\omega)} \langle U_\star(f), \omega_1\rangle\cdot\langle U_\star(t), \omega_2\rangle \\
&= \sum_{\Delta_\bullet(\omega)} \langle f, U_\star^T(\omega_1)\rangle\cdot\langle t, \overline{U}_\star^T(\omega_2)\rangle.
\end{aligned}$$

Since $\overline{U}_\star^T(\omega_2)$ is a dual tree, this is:

$$= \sum_{\Delta_\bullet(\omega)} \langle ft, U_\star^T(\omega_1)\cdot\overline{U}_\star^T(\omega_2)\rangle.$$

□

7.4.1 The Case of One Free Generator

Now consider the case that $C = \langle\bullet\rangle$ is a one-dimensional vector space. Recall the isomorphism $\psi : \mathcal{E}_\bullet \cong T(\mathrm{OT}_\bullet^{\circledast})$ as algebras but the coproduct on this is different from $\mathcal{H}_\bullet \cong T(\mathrm{OT}_\bullet^{\circledast})$. To signify the difference, we denote the former by $T^\circ(\mathrm{OT}_\bullet^{\circledast})$. It is the free algebra on the alphabet $a_\bullet(t)$ where the t are ordered trees. Multiplication on $\mathcal{E}_\bullet = Sym(P_\bullet^{\circledast})$ corresponds to the shuffle product on $T^\circ(\mathrm{OT}_\bullet^{\circledast})$.

The coproduct

$$\mathcal{H}_\bullet \xrightarrow{\Delta_\star} \mathcal{E}_\bullet\otimes_k\mathcal{H}_\bullet$$

may then by Sect. 7.2.1 be written as

$$T(\mathrm{OT}_\bullet^{\circledast}) \xrightarrow{\Delta_\star} T^\circ(\mathrm{OT}_\bullet^{\circledast})\otimes_k T(\mathrm{OT}_\bullet^{\circledast}) = K\otimes_k T(\mathrm{OT}_\bullet^{\circledast}).$$

The two bialgebras $T(\mathrm{OT}_\bullet^{\circledast})$ and $T^\circ(\mathrm{OT}_\bullet^{\circledast})$ are said to be in cointeraction, a notion studied in [4, 12, 20], and [11].

The element $U^t(\omega^{(2)})$ is in $C_K^* \cong K$. By the comment following (30) it is in

$$C_H^* = \mathrm{Hom}_k(\bullet, P_\bullet)^{\circledast}\otimes_k\bullet^* \cong P_\bullet^{\circledast}.$$

Then $U^t(\omega^{(2)})$ is simply the image of $\omega^{(2)}$ by the natural projection $T(\mathrm{OT}_\bullet^\circledast) \to P_\bullet^\circledast$. We may consider $U^t(\omega^{(2)})$ as an element of $K \cong T^\circ(\mathrm{OT}_\bullet^\circledast)$ via the isomorphism ψ above. We are then using the Euler idempotent map

$$T(\mathrm{OT}_\bullet^\circledast) \xrightarrow{\pi} T(\mathrm{OT}_\bullet^\circledast) \cong T^\circ(\mathrm{OT}_\bullet^\circledast),$$

so that $U^t(\omega^{(2)}) = \pi(\omega^{(2)})$.

Let B_+ be the operation of attaching a root to a forest in order to make it a tree. By a decorated shuffle $\sqcup\!\sqcup$ below we mean taking the shuffle product of the corresponding factors in $K = T^\circ(\mathrm{OT}_\bullet^\circledast)$. By the decorated $\cdot$ product we mean concatenating the corresponding factors in $T(\mathrm{OT}_\bullet^\circledast)$. Then we may write the recursion of Proposition 7.3 as:

Proposition 7.4

$$\begin{aligned}\Delta_\star(\omega) &= \sqcup\!\sqcup_{13} \cdot_{24} \Delta_\star(\omega_1) \otimes \overline{U}_\star^T(\omega_2) \\ &= \sqcup\!\sqcup_{135} \cdot_{24} \Delta_\star(\omega_1) \otimes B^+(\Delta_\star(\omega_2^{(1)})) \otimes \pi(\omega_2^{(2)})\end{aligned}$$

Acknowledgements We would like to thank Kurusch Ebrahimi-Fard, Kristoffer Føllesdal and Frédéric Patras for discussions on the topics of this paper.

References

1. Bogfjellmo, G., Schmeding, A.: The Lie group structure of the butcher group. Found. Comput. Math. **17**(1), 127–159 (2017)
2. Butcher, J.C.: Coefficients for the study of Runge-Kutta integration processes. J. Aust. Math. Soc. **3**(2), 185–201 (1963)
3. Butcher, J.C.: An algebraic theory of integration methods. Math. Comput. **26**(117), 79–106 (1972)
4. Calaque, D., Ebrahimi-Fard, K., Manchon, D.: Two interacting Hopf algebras of trees: a Hopf-algebraic approach to composition and substitution of B-series. Adv. Appl. Math. **47**(2), 282–308 (2011)
5. Chapoton, F., Livernet, M.: Pre-Lie algebras and the rooted trees operad. Int. Math. Res. Not. **2001**(8), 395–408 (2001)
6. Chartier, P., Harirer, E., Vilmart, G.: A substitution law for B-series vector fields. Technical Report 5498, INRIA (2005)
7. Cox, D., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms, vol. 3. Springer, New York (1992)
8. Ebrahimi-Fard, K., Manchon, D.: Twisted dendriform algebras and the pre-Lie Magnus expansion. J. Pure Appl. Algebra **215**(11), 2615–2627 (2011)
9. Ebrahimi-Fard, K., Patras, F.: The pre-Lie structure of the time-ordered exponential. Lett. Math. Phys. **104**(10), 1281–1302 (2014)
10. Ebrahimi-Fard, K., Lundervold, A., Munthe-Kaas, H.Z.: On the Lie enveloping algebra of a post-Lie algebra. J. Lie Theory **25**(4), 1139–1165 (2015)
11. Foissy, L.: Chromatic polynomials and bialgebras of graphs. arXiv preprint:1611.04303 (2016)

12. Foissy, L.: Commutative and non-commutative bialgebras of quasi-posets and applications to ehrhart polynomials. arXiv preprint:1605.08310 (2016)
13. Gerstenhaber, M.: The cohomology structure of an associative ring. Ann. Math. **78**, 267–288 (1963)
14. Hairer, E.: Backward analysis of numerical integrators and symplectic methods. Ann. Numer. Math. **1**, 107–132 (1994)
15. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, vol. 31. Springer Science & Business Media, Berlin/New York (2006)
16. Hartshorne, R.: Algebraic geometry. Graduate texts in mathematics, vol. 52. Springer, New York (2013)
17. Lundervold, A., Munthe-Kaas, H.: Hopf algebras of formal diffeomorphisms and numerical integration on manifolds. Contemp. Math. **539**, 295–324 (2011)
18. Lundervold, A., Munthe-Kaas, H.: Backward error analysis and the substitution law for Lie group integrators. Found. Comput. Math. **13**(2), 161–186 (2013)
19. Manchon, D.: A short survey on pre-Lie algebras. In: Carey, A. (ed.) Noncommutative Geometry and Physics: Renormalisation, Motives, Index Theory, pp. 89–102. Switzerland European Mathematical Society Publishing House, Zuerich (2011)
20. Manchon, D.: On bialgebras and hopf algebras of oriented graphs. Confluentes Mathematici **4**(1), 1240003 (2012)
21. Munthe-Kaas, H., Krogstad, S.: On enumeration problems in Lie–Butcher theory. Futur. Gener. Comput. Syst. **19**(7), 1197–1205 (2003)
22. Munthe-Kaas, H., Stern, A., Verdier, O.: Past-Lie algebroids and Lie algebra actions. (To appear, 2018)
23. Munthe-Kaas, H.Z., Føllesdal, K.K.: Lie-Butcher series, Geometry, Algebra and Computation. arXiv preprint:1701.03654 (2017)
24. Munthe-Kaas, H.Z., Lundervold, A.: On post-Lie algebras, Lie–Butcher series and moving frames. Found. Comput. Math. **13**(4), 583–613 (2013)
25. Munthe-Kaas, H.Z., Wright, W.M.: On the Hopf algebraic structure of Lie group integrators. Found. Comput. Math. **8**(2), 227–257 (2008)
26. Oudom, J.-M., Guin, D.: On the Lie enveloping algebra of a pre-Lie algebra. J. K-theory K-theory Appl. Algebra Geom. Topol. **2**(1), 147–167 (2008)
27. Reutenauer, C.: Free Lie algebras. Handb. Algebra **3**, 887–903 (2003)
28. Vallette, B.: Homology of generalized partition posets. J. Pure Appl. Algebra **208**(2), 699–725 (2007)
29. Vinberg, È.B.: The theory of homogeneous convex cones. Trudy Moskovskogo Matematicheskogo Obshchestva **12**, 303–358 (1963)

Extension of the Product of a Post-Lie Algebra and Application to the SISO Feedback Transformation Group

Loïc Foissy

Abstract We describe both post- and pre-Lie algebra $\mathfrak{g}_{SISO}$ associated to the affine SISO feedback transformation group. We show that it is a member of a family of post-Lie algebras associated to representations of a particular solvable Lie algebra. We first construct the extension of the magmatic product of a post-Lie algebra to its enveloping algebra, which allows to describe free post-Lie algebras and is widely used to obtain the enveloping of $\mathfrak{g}_{SISO}$ and its dual.

1 Introduction

The affine SISO feedback transformation group G_{SISO} [9], which appears in Control Theory, can be seen as the character group of a Hopf algebra $\mathcal{H}_{SISO}$; let us start by a short presentation of this object (we slightly modify the notations of [9]).

1. First, let us recall some algebraic structures on noncommutative polynomials.
   a. Let x_1, x_2 be two indeterminates. We consider the algebra of noncommutative polynomials $\mathbb{K}\langle x_1, x_2\rangle$. As a vector space, it is generated by words in letters x_1, x_2; its product is the concatenation of words; its unit, the empty word, is denoted by $\emptyset$.
   b. $\mathbb{K}\langle x_1, x_2\rangle$ is a Hopf algebra with the concatenation product and the deshuffling coproduct $\Delta_{\sqcup\!\sqcup}$, defined by $\Delta_{\sqcup\!\sqcup}(x_i) = x_i \otimes \emptyset + \emptyset \otimes x_i$, for $i \in \{1, 2\}$.

L. Foissy (✉)
Fédération de Recherche Mathématique du Nord Pas de Calais FR 2956, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, Université du Littoral Côte d'Opale-Centre Universitaire de la Mi-Voix, Calais Cedex, France
e-mail: foissy@univ-littoral.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_13

c. $\mathbb{K}\langle x_1, x_2\rangle$ is also a commutative, associative algebra with the shuffle product $\sqcup\!\sqcup$: for example, if $i, j, k, l \in \{1, 2\}$,

$$x_i x_j \sqcup\!\sqcup x_k x_l = x_i x_j x_k x_l + x_i x_k x_j x_l + x_i x_k x_l x_j + x_k x_i x_j x_l + x_k x_i x_l x_j + x_k x_l x_i x_j.$$

2. The vector space $\mathbb{K}\langle x_1, x_2\rangle^2$ is generated by words $x_{i_1} \dots x_{i_k}\epsilon_j$, where $k \geq 0$, $i_1, \dots, i_k, j \in \{1, 2\}$, and (ϵ_1, ϵ_2) denotes the canonical basis of $\mathbb{K}^2$.
3. As an algebra, $\mathcal{H}_{ISO}$ is equal to the symmetric algebra $S(\mathbb{K}\langle x_1, x_2\rangle^2)$; its product is denoted by μ and its unit by 1. Two coproducts Δ_* and $\Delta_\bullet$ are defined on $\mathcal{H}_{ISO}$. For all $h \in \mathcal{H}_{ISO}$, we put $\overline{\Delta}_*(h) = \Delta_*(h) - 1 \otimes h$ and $\overline{\Delta}_\bullet(h) = \Delta_\bullet(h) - 1 \otimes h$. Then:

   - For all $i \in \{1, 2\}$, $\overline{\Delta}_*(\emptyset\epsilon_i) = \emptyset\epsilon_i \otimes 1$.
   - For all $g \in \mathbb{K}\langle x_1, x_2\rangle$, for all $i \in \{1, 2\}$:

$$\overline{\Delta}_* \circ \theta_{x_1}(g\epsilon_i) = (\theta_{x_1} \otimes Id) \circ \overline{\Delta}_*(g\epsilon_i) + (\theta_{x_2} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)\epsilon_i \otimes \epsilon_2),$$
$$\overline{\Delta}_* \circ \theta_{x_2}(g\epsilon_i) = (\theta_{x_2} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)\epsilon_i \otimes \epsilon_1),$$

     where $\theta_x(h\epsilon_i) = xh\epsilon_i$ for all $x \in \{x_1, x_2\}$, $h \in \mathbb{K}\langle x_1, x_2\rangle$, $i \in \{1, 2\}$. These are formulas of Lemma 4.1 of [9], with the notations $a_w = w\epsilon_2$, $b_w = w\epsilon_1$, $\theta_0 = \theta_{x_1}$, $\theta_1 = \theta_{x_2}$ and $\tilde{\Delta} = \overline{\Delta}_*$.
   - for all $g \in \mathbb{K}\langle x_1, x_2\rangle$:

$$\overline{\Delta}_\bullet(g\epsilon_1) = (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_1)),$$
$$\overline{\Delta}_\bullet(g\epsilon_2) = \overline{\Delta}_*(g\epsilon_2) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_2 \otimes \epsilon_1)).$$

   This coproduct $\Delta_\bullet$ makes $\mathcal{H}_{ISO}$ a Hopf algebra, and Δ_* is a right coaction on this coproduct, that is to say:

$$(\Delta_\bullet \otimes Id) \circ \Delta_\bullet = (Id \otimes \Delta_\bullet) \circ \Delta_\bullet, \quad (\Delta_* \otimes Id) \circ \Delta_* = (Id \otimes \Delta_\bullet) \circ \Delta_*.$$

4. After the identification of $\emptyset\epsilon_1$ with the unit of $\mathcal{H}_{ISO}$, we obtain a commutative, graded and connected Hopf algebra, in other words the dual of an enveloping algebra $\mathcal{U}(\mathfrak{g}_{ISO})$.

Our aim is to give a description of the underlying Lie algebra $\mathfrak{g}_{ISO}$. It turns out that it is both a pre-Lie algebra (or a Vinberg algebra [1], see [4] for a survey on these objects) and a post-Lie algebra [5, 10]: it has a Lie bracket ${}_a[-,-]$ and two nonassociative products $*$ and $\bullet$, such that for all $x, y, z \in \mathfrak{g}_{ISO}$:

$$x * {}_a[y, z] = (x * y) * z - x * (y * z) - (x * z) * y + x * (z * y),$$
$${}_a[x, y] * z = {}_a[x * z, y] + {}_a[x, y * z];$$
$$(x \bullet y) \bullet z - x \bullet (y \bullet z) = (x \bullet z) \bullet y - x \bullet (z \bullet y).$$

The Lie bracket on $\mathfrak{g}_{SISO}$ corresponding to G_{SISO} is ${}_a[-,-]_*$:

$$\forall x, y \in \mathfrak{g}_{SISO},\ {}_a[x,y]_* = {}_a[x,y] + x * y - y * x = x \bullet y - y \bullet x.$$

Let us be more precise on these structures. As a vector space, $\mathfrak{g}_{SISO} = \mathbb{K}\langle x_1, x_2\rangle^2$, and:

$$\forall f, g \in \mathbb{K}\langle x_1, x_2\rangle,\ \forall i, j \in \{1,2\},\quad {}_a[f\epsilon_i, g\epsilon_j] = \begin{cases} 0 \text{ if } i = j, \\ -f \sqcup\!\sqcup g\epsilon_2 \text{ if } i = 2 \text{ and } j = 1, \\ f \sqcup\!\sqcup g\epsilon_2 \text{ if } i = 1 \text{ and } j = 2. \end{cases}$$

The magmatic product $*$ is inductively defined. If $f, g \in \mathbb{K}\langle x_1, x_2\rangle$ and $i, j \in \{1,2\}$:

$$\begin{aligned} \emptyset\epsilon_i * g\epsilon_j &= 0, & x_2 f\epsilon_i * g\epsilon_1 &= x_2(f\epsilon_i * g\epsilon_1) + x_2(f \sqcup\!\sqcup g)\epsilon_i, \\ x_1 f\epsilon_i * g\epsilon_j &= x_1(f\epsilon_i * g\epsilon_j), & x_2 f\epsilon_i * g\epsilon_2 &= x_2(f\epsilon_i * g\epsilon_2) + x_1(f \sqcup\!\sqcup g)\epsilon_i. \end{aligned}$$

The pre-Lie product $\bullet$, first identified in [9], is given by:

$$\forall f, g \in \mathbb{K}\langle x_1, x_2\rangle,\ \forall i, j \in \{1,2\},\ f\epsilon_i \bullet g\epsilon_j = (f \sqcup\!\sqcup g)\delta_{i,1}\epsilon_j + f\epsilon_i * g\epsilon_j.$$

We shall show here that this is a special case of a family of post-Lie algebras, associated to modules over certain solvable Lie algebras.

We start with general preliminary results on post-Lie algebras. We extend the now classical Oudom-Guin construction on prelie algebras [6, 7] to the post-Lie context in the first section: this is a result of [2] (Proposition 3.1), which we prove here in a different, less direct way; our proof allows also to obtain a description of free post-Lie algebras. Recall that if $(V, *)$ is a pre-Lie algebra, the pre-Lie product $*$ can be extended to $S(V)$ in such a way that the product defined by:

$$\forall f, g \in S(V),\ f \circledast g = \sum f * g^{(1)} g^{(2)}$$

is associative, and makes $S(V)$ a Hopf algebra, isomorphic to $\mathscr{U}(V)$. For any magmatic algebra $(V, *)$, we construct in a similar way an extension of $*$ to $T(V)$ in Proposition 1. We prove in Theorem 1 that the product $\circledast$ defined by:

$$\forall f, g \in T(V),\ f \circledast g = \sum f * g^{(1)} g^{(2)}$$

makes $T(V)$ a Hopf algebra. The Lie algebra of its primitive elements, which is the free Lie algebra $\mathscr{L}ie(V)$ generated by V, is stable under $*$ and turns out to

be a post-Lie algebra (Proposition 2) satisfying a universal property (Theorem 2). In particular, if V is, as a magmatic algebra, freely generated by a subspace W, $\mathscr{L}ie(V)$ is the free post-Lie algebra generated by W (Corollary 1). Moreover, if $V = ([-,-], *)$ is a post-Lie algebra, this construction goes through the quotient defining $\mathscr{U}(V, [-,-])$, defining a new product $\circledast$ on it, making it isomorphic to the enveloping algebra of V with the Lie bracket defined by:

$$\forall x, y \in V,\ [x, y]_* = [x, y] + x * y - y * x.$$

For example, if $x_1, x_2, x_3 \in V$:

$$x_1 \circledast x_2x_3 = x_1x_2x_3 + (x_1 * x_2)x_3 + (x_1 * x_3)x_2 + (x_1 * x_2) * x_3 - x_1 * (x_2 * x_3)$$
$$x_1x_2 \circledast x_3 = x_1x_2x_3 + (x_1 * x_3)x_2 + x_1(x_2 * x_3).$$

In the particular case where $[-,-] = 0$, we recover the Oudom-Guin construction.

The second section is devoted to the study of a particular solvable Lie algebra $\mathfrak{g}_a$ associated to an element $a \in \mathbb{K}^N$. As the Lie bracket of $\mathfrak{g}_a$ comes from an associative product, the construction of the first section holds, with many simplifications: we obtain an explicit description of $\mathscr{U}(\mathfrak{g}_a)$ with the help of a product $\blacktriangleleft$ on $S(\mathfrak{g}_a)$ (Proposition 6). A short study of $\mathfrak{g}_a$-modules when $a = (1, 0, \ldots, 0)$ (which is a generic case) is done in Proposition 8, considering $\mathfrak{g}_a$ as an associative algebra, and in Proposition 9, considering it as a Lie algebra. In particular, if $\mathbb{K}$ is algebraically closed, any $\mathfrak{g}_a$ modules inherits a natural decomposition in characteristic subspaces.

Our family of post-Lie algebras is introduced in the third section; it is reminescent of the construction of [3]. Let us fix a vector space V, $(a_1, \ldots, a_N) \in \mathbb{K}^N$ and a family $F_1, \ldots, F_N$ of endomorphisms of V. We define a product $*$ on $T(V)^N$, such that for all $f, g \in T(V)$, $x \in V$, $i, j \in \{1, \ldots, N\}$:

$$\emptyset\epsilon_i * g\epsilon_j = 0,$$
$$xf\epsilon_i * g\epsilon_j = x(f\epsilon_i * g\epsilon_j) + F_j(x)(f \sqcup\!\sqcup g)\epsilon_i,$$

where $(\epsilon_1, \ldots, \epsilon_N)$ is the canonical basis of $\mathbb{K}^N$ and $\sqcup\!\sqcup$ is the shuffle product of $T(V)$. The Lie bracket of $T(V)^N$ that we shall use here is:

$$\forall f, g \in T(V),\ \forall i, j \in \{1, \ldots, N\},\ {}_a[f\epsilon_i, g\epsilon_j] = (f \sqcup\!\sqcup g)(a_i\epsilon_j - a_j\epsilon_i).$$

This Lie bracket comes from an associative product ${}_a\sqcup\!\sqcup$ defined by:

$$\forall f, g \in T(V),\ \forall i, j \in \{1, \ldots, N\},\ f\epsilon_i\ {}_a\sqcup\!\sqcup\ g\epsilon_j = a_i(f \sqcup\!\sqcup g)\epsilon_j.$$

We put $\bullet = * +\ {}_a\sqcup\!\sqcup$. We prove in Theorem 3 the equivalence of the three following conditions:

- $(T(V)^N, \bullet)$ is a pre-Lie algebra.
- $(T(V)^N,\ {}_a[-,-],*)$ is a post-Lie algebra.
- $F_1, \ldots, F_N$ defines a structure of $\mathfrak{g}_a$-module on V.

If this holds, the construction of the first section allows to obtain two descriptions of the enveloping algebra of $\mathscr{U}(T(V)^N)$, respectively coming from the post-Lie product $*$ and from the pre-Lie product $\bullet$: the extensions of $*$ and of $\bullet$ are respectively described in Propositions 13 and 14. It is shown in Proposition 15 that the two associated descriptions of $\mathscr{U}(T(V)^N)$ are equal. For $\mathfrak{g}_{SISO}$, we take $a = (1, 0)$, $V = Vect(x_1, x_2)$ and:

$$F_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

which indeed define a $\mathfrak{g}_{(1,0)}$-module. In order to relate this to the Hopf algebra $\mathscr{H}_{SISO}$ of [9], we need to consider the dual of the enveloping of $T(V)^N$. First, if $a = (1, 0, \ldots, 0)$, we observe that the decomposition of V as a $\mathfrak{g}_a$-module of the second section induces a graduation of the post-Lie algebra $T(V)^N$ (Proposition 16), unfortunately not connected: the component of degree 0 is 1-dimensional, generated by $\emptyset\epsilon_1$. Forgetting this element, that is, considering the augmentation ideal of the graded post-Lie algebra $T(V)^N$, we can dualize the product $\circledast$ of $S(T(V)^N)$ in order to obtain the coproduct of the dual Hopf algebra in an inductive way. For $\mathfrak{g}_{SISO}$, we indeed obtain the inductive formulas of $\mathscr{H}_{SISO}$, finally proving that the dual Lie algebra of this Hopf algebra, which in some sense can be exponentiated to G_{SISO}, is indeed post-Lie and pre-Lie.

Notations

1. Let $\mathbb{K}$ be a commutative field. The canonical basis of $\mathbb{K}^n$ is denoted by $(\epsilon_1, \ldots, \epsilon_n)$.
2. For all $n \geq 1$, we denote by $[n]$ the set $\{1, \ldots, n\}$.
3. We shall use Sweeder's notations: if C is a coalgebra and $x \in C$,

$$\Delta^{(1)}(x) = \Delta(x) = \sum x^{(1)} \otimes x^{(2)},$$
$$\Delta^{(2)}(x) = (\Delta \otimes Id) \circ \Delta(x) = \sum x^{(1)} \otimes x^{(2)} \otimes x^{(3)}.$$

2 Extension of a Post-Lie Product

We first generalize the Oudom-Guin extension of a pre-Lie product in a post-Lie algebraic context, as done in [2]. Let us first recall what a post-Lie algebra is.

Definition 1 A (right) *post-Lie algebra* is a family $(\mathfrak{g}, \{-,-\}, *)$, where $\mathfrak{g}$ is a vector space, $\{-,-\}$ and $*$ are bilinear products on $\mathfrak{g}$ such that:

- $(\mathfrak{g}, \{-,-\})$ is a Lie algebra.
- For all $x, y, z \in \mathfrak{g}$:

$$x * \{y,z\} = (x*y)*z - x*(y*z) - (x*z)*y + x*(z*y), \tag{1}$$

$$\{x,y\}*z = \{x*z, y\} + \{x, y*z\}. \tag{2}$$

2. If $(\mathfrak{g}, \{-,-\}, *)$ is post-Lie, we define a second Lie bracket on $\mathfrak{g}$:

$$\forall x, y \in \mathfrak{g},\ \{x,y\}_* = \{x,y\} + x*y - y*x.$$

Note that if $\{-,-\}$ is 0, then $(\mathfrak{g}, *)$ is a (right) pre-Lie algebra, that is to say:

$$\forall x, y, z \in \mathfrak{g},\ (x*y)*z - x*(y*z) = (x*z)*y - x*(z*y). \tag{3}$$

2.1 Extension of a Magmatic Product

Let V be a vector space. We here use the tensor Hopf algebra $T(V)$. In this section, we shall denote the unit of $T(V)$ by 1. Its product is the concatenation of words, and its coproduct $\Delta_{\sqcup\!\sqcup}$ is the cocommutative deshuffling coproduct. For example, if $x_1, x_2, x_3 \in V$:

$$\begin{aligned}\Delta_{\sqcup\!\sqcup}(x_1x_2x_3) = {} & x_1x_2x_3 \otimes 1 + x_1x_2 \otimes x_3 + x_1x_3 \otimes x_2 + x_2x_3 \otimes x_1 \\ & + x_1 \otimes x_2x_3 + x_2 \otimes x_1x_3 + x_3 \otimes x_1x_2 + 1 \otimes x_1x_2x_3.\end{aligned}$$

Its counit is denoted by ε: $\varepsilon(1) = 1$ and if $k \geq 1$ and $x_1, \ldots, x_k \in V$, $\varepsilon(x_1 \ldots x_k) = 0$.

Proposition 1 *Let V be a vector space and $* : V \otimes V \longrightarrow V$ be a magmatic product on V. Then $*$ can be uniquely extended as a map from $T(V) \otimes T(V)$ to $T(V)$ such that for all $f, g, h \in T(V)$, $x, y \in V$:*

- $f * 1 = f$.
- $1 * f = \varepsilon(f)1$.
- $x * (fy) = (x*f)*y - x*(f*y)$.
- $(fg)*h = \sum \left(f * h^{(1)}\right)\left(g * h^{(2)}\right)$.

Proof We first inductively extend $*$ from $V \otimes T(V)$ to V. If $n \geq 0, x, y_1, \ldots, y_n \in V$, we put:

$$x * y_1 \ldots y_n = \begin{cases} x \text{ if } n = 0, \\ x * y_1 \text{ if } n = 1, \\ \underbrace{\underbrace{(x * (y_1 \ldots y_{n-1}))}_{\in V} * \underbrace{y_n}_{\in V}}_{\in V} - \displaystyle\sum_{i=1}^{n-1} \underbrace{x * \underbrace{(y_1 \ldots (y_i * y_n) \ldots y_{n-1})}_{\in V^{\otimes(n-1)}}}_{\in V} \text{ if } n \geq 2. \end{cases}$$

This product is then extended from $T(V) \otimes T(V)$ to $T(V)$ in the following way:

- For all $f \in T(V)$, $1 * f = \varepsilon(f)1$.
- For all $n \geq 1$, for all $x_1, \ldots, x_n \in V$, $f \in T(V)$:

$$(x_1 \ldots x_n) * f = \sum (\underbrace{x_1 * f^{(1)}}_{\in V}) \ldots (\underbrace{x_n * f^{(n)}}_{\in V}) \in V^{\otimes n}.$$

This product satisfies all the required properties.

Examples If $x_1, x_2, x_3, x_4 \in V$:

$$\begin{aligned} x_1 * (x_2 x_3 x_4) = &((x_1 * x_2) * x_3) * x_4 - (x_1 * (x_2 * x_3)) * x_4 - (x_1 * (x_2 * x_4)) * x_3 \\ &+ x_1 * ((x_2 * x_4) * x_3) - (x_1 * x_2) * (x_3 * x_4) + x_1 * (x_2 * (x_3 * x_4)). \end{aligned}$$

Lemma 1

1. *For all* $k \in \mathbb{N}$, $V^{\otimes k} * T(V) \subseteq V^{\otimes k}$.
2. *For all* $f, g \in T(V)$, $\varepsilon(f * g) = \varepsilon(f)\varepsilon(g)$.
3. *For all* $f, g \in T(V)$, $\Delta_{\sqcup\!\sqcup}(f * g) = \Delta_{\sqcup\!\sqcup}(f) * \Delta_{\sqcup\!\sqcup}(g)$.
4. *For all* $f, g \in T(V)$, $y \in V$, $f * (gy) = (f * g) * y - f * (g * y)$.
5. *For all* $f, g, h \in T(V)$, $(f * g) * h = \sum f * \left(\left(g * h^{(1)}\right) h^{(2)}\right)$.

Proof 1. and 2. Immediate.

3. We prove it for $f = x_1 \ldots x_n$, by induction on n. If $n = 0$, then $f = 1$. Moreover, $\Delta_{\sqcup\!\sqcup}(1 * g) = \varepsilon(g)\Delta_{\sqcup\!\sqcup}(1) = \varepsilon(g)1 \otimes 1$, and:

$$\Delta_{\sqcup\!\sqcup}(f) * \Delta_{\sqcup\!\sqcup}(g) = \sum 1 * g^{(1)} \otimes 1 * g^{(2)} = \varepsilon\left(g^{(1)}\right)\varepsilon\left(g^{(2)}\right) 1 \otimes 1 = \varepsilon(g)1 \otimes 1.$$

If $n = 1$, then $f \in V$. In this case, from the second point, $f*g \in V$, so $\Delta_{\sqcup\!\sqcup}(f*g) = f*g \otimes 1 + 1 \otimes f*g$. Moreover:

$$\begin{aligned}\Delta_{\sqcup\!\sqcup}(f) * \Delta_{\sqcup\!\sqcup}(g) &= (f \otimes 1 + 1 \otimes f) * \Delta_{\sqcup\!\sqcup}(g) \\ &= \sum f * g^{(1)} \otimes \varepsilon\left(g^{(2)}\right) 1 + \sum \varepsilon\left(g^{(1)}\right) 1 \otimes f * g^{(2)} \\ &= f*g \otimes 1 + 1 \otimes f*g.\end{aligned}$$

If $n \geq 2$, we put $f_1 = x_1 \ldots x_{n-1}$ and $f_2 = x_n$. By the induction hypothesis applied to f_1:

$$\begin{aligned}\Delta_{\sqcup\!\sqcup}(f*g) &= \sum \Delta_{\sqcup\!\sqcup}\left(\left(f_1 * g^{(1)}\right)\left(f_2 * g^{(2)}\right)\right) \\ &= \Delta_{\sqcup\!\sqcup}\left(f_1 * g^{(1)}\right) \Delta_{\sqcup\!\sqcup}\left(f_2 * g^{(2)}\right) \\ &= \sum \left(f_1^{(1)} * (g^{(1)})^{(1)}\right)\left(f_2^{(1)} * (g^{(2)})^{(1)}\right) \otimes \left(f_1^{(2)} * (g^{(1)})^{(2)}\right)\left(f_2^{(2)} * (g^{(2)})^{(2)}\right) \\ &= \sum (f_1 f_2)^{(1)} * g^{(1)} \otimes (f_1 f_2)^{(2)} * g^{(2)} \\ &= \Delta_{\sqcup\!\sqcup}(f) * \Delta_{\sqcup\!\sqcup}(g).\end{aligned}$$

We used the cocommutativity of $\Delta_{\sqcup\!\sqcup}$ for the fourth equality.

4. We prove it for $f = x_1 \ldots x_n$, by induction on n. If $n = 0$ or 1, this is immediate. If $n \geq 2$, we put $f_1 = x_1 \ldots x_{n-1}$ and $f_2 = x_n$. The induction hypothesis holds for f_1. Moreover:

$$\begin{aligned}f * (gy) &= \sum \left(f_1 * g^{(1)}\right)\left(\left(f_2 * g^{(2)}\right) * y\right) - \sum \left(f_1 * g^{(1)}\right)\left(f_2 * \left(g^{(2)} * y\right)\right) \\ &\quad + \sum \left(\left(f_1 * g^{(1)}\right) * y\right)\left(f_2 * g^{(2)}\right) - \sum \left(f_1 * \left(g^{(1)} * y\right)\right)\left(f_2 * g^{(2)}\right), \\ (f*g) * y &= \sum \left(\left(f_1 * g^{(1)}\right) * y\right)\left(f_2 * g^{(2)}\right) + \sum \left(f_1 * g^{(1)}\right)\left(\left(f_2 * g^{(2)}\right) * y\right), \\ f * (g * y) &= \sum \left(f_1 * \left(g^{(1)} * y\right)\right)\left(f_2 * g^{(2)}\right) + \sum \left(f_1 * g^{(1)}\right)\left(f_2 * \left(g^{(2)} * y\right)\right).\end{aligned}$$

5. We prove this for $h = z_1 \ldots z_n$ and we proceed by induction on n. It is direct if $n = 0$ or 1. If $n \geq 2$, we put $h_1 = z_1 \ldots z_{n-1}$ and $h_2 = z_n$. From the fourth point:

$$\begin{aligned}(f*g) * h &= ((f*g) * h_1) * h_2 - (f*g) * (h_1 * h_2) \\ &= \sum \left(f * \left(\left(g * h_1^{(1)}\right) h_1^{(2)}\right)\right) * h_2 - \sum f * \left(\left(g * (h_1 * h_2)^{(1)}\right)(h_1 * h_2)^{(2)}\right) \\ &= \sum f * \left(\left(\left(g * h_1^{(1)}\right) h_1^{(2)}\right) * h_2\right) + \sum f * \left(\left(g * h_1^{(1)}\right) h_1^{(2)} h_2\right)\end{aligned}$$

$$
\begin{aligned}
&-\sum f*\left(\left(g*\left(h_1^{(1)}*h_2^{(1)}\right)\right)\left(h_1^{(2)}*h_2^{(2)}\right)\right)\\
&=\sum f*\left(\left(\left(g*h_1^{(1)}\right)*h_2\right)h_1^{(2)}\right)+\sum f*\left(\left(g*h_1^{(1)}\right)\left(h_1^{(2)}*h_2\right)\right)\\
&+\sum f*\left(\left(g*h_1^{(1)}\right)h_1^{(2)}h_2\right)-\sum f*\left(\left(g*\left(h_1^{(1)}*h_2\right)\right)h_1^{(2)}\right)\\
&-\sum f*\left(\left(g*h_1^{(1)}\right)\left(h_1^{(2)}*h_2\right)\right)\\
&=\sum f*\left(\left(g*\left(h_1^{(1)}*h_2\right)\right)h_1^{(2)}\right)+\sum f*\left(\left(g*\left(h_1^{(1)}h_2\right)\right)h_1^{(2)}\right)\\
&+\sum f*\left(\left(g*h_1^{(1)}\right)\left(h_1^{(2)}*h_2\right)\right)+\sum f*\left(\left(g*h_1^{(1)}\right)h_1^{(2)}h_2\right)\\
&-\sum f*\left(\left(g*\left(h_1^{(1)}*h_2\right)\right)h_1^{(2)}\right)-\sum f*\left(\left(g*h_1^{(1)}\right)\left(h_1^{(2)}*h_2\right)\right)\\
&=\sum f*\left(\left(g*\left(h_1^{(1)}h_2\right)\right)h_1^{(2)}\right)+\sum f*\left(\left(g*h_1^{(1)}\right)h_1^{(2)}h_2\right).
\end{aligned}
$$

As $\Delta_{\sqcup\!\sqcup}(h_2)=h_2\otimes 1+1\otimes h_2$, $\Delta_{\sqcup\!\sqcup}(h)=\sum h_1^{(1)}h_2\otimes h_1^{(2)}+\sum h_1^{(1)}\otimes h_1^{(2)}h_2$, so the result holds for h.

2.2 Associated Hopf Algebra and Post-Lie Algebra

Theorem 1 *Let $*$ be a magmatic product on V. This product is extended to $T(V)$ by Proposition 1. We define a product $\circledast$ on $T(V)$ by:*

$$\forall f,g\in T(V),\ f\circledast g=\sum\left(f*g^{(1)}\right)g^{(2)}.$$

Then $(T(V),\circledast,\Delta_{\sqcup\!\sqcup})$ is a Hopf algebra.

Proof For all $f\in T(V)$:

$$1\circledast f\sum\left(1*f^{(1)}\right)f^{(2)}=\sum\varepsilon\left(f^{(1)}\right)f^{(2)}=f;\qquad \circledast 1=(f*1)1=f.$$

For all $f, g, h \in T(V)$, by Lemma 1(1–5):

$$(f \circledast g) \circledast h = \sum \left(f * \left(\left(g^{(1)} * h^{(1)} \right) h^{(2)} \right) \right) \left(g^{(2)} * h^{(3)} \right) h^{(4)};$$

$$f \circledast (g \circledast h) = \sum \left(f * \left(\left(g^{(1)} * h^{(1)} \right) h^{(3)} \right) \right) \left(g^{(2)} * h^{(2)} \right) h^{(4)}.$$

As $\Delta_{\sqcup\!\sqcup}$ is cocommutative, $(f \circledast g) \circledast h = f \circledast (g \circledast h)$, so $(T(V), \circledast)$ is a unitary, associative algebra.

For all $f, g \in T(V)$, by Lemma 1(1–3):

$$\Delta_{\sqcup\!\sqcup}(f \circledast g) = \sum f^{(1)} \circledast g^{(1)} \otimes f^{(2)} \circledast g^{(2)}.$$

Hence, $(T(V), \circledast, \Delta_{\sqcup\!\sqcup})$ is a Hopf algebra.

Remark By Lemma 1:

- For all $f, g, h \in T(V)$, $(f * g) * h = f * (g \circledast h)$: $(T(V), *)$ is a right $(T(V), \circledast)$-module.
- By restriction, for all $n \geq 0$, $(V^{\otimes n}, *)$ is a right $(T(V), \circledast)$-module. Moreover, for all $n \geq 0$, $(V^{\otimes n}, *) = (V, *)^{\otimes n}$ as a right module over the Hopf algebra $(T(V), \circledast, \Delta_{\sqcup\!\sqcup})$.

Examples Let $x_1, x_2, x_3 \in V$.

$$x_1 \circledast x_2 x_3 = x_1 x_2 x_3 + (x_1 * x_2) x_3 + (x_1 * x_3) x_2 + (x_1 * x_2) * x_3 - x_1 * (x_2 * x_3)$$

$$x_1 x_2 \circledast x_3 = x_1 x_2 x_3 + (x_1 * x_3) x_2 + x_1 (x_2 * x_3).$$

The vector space of primitive elements of $(T(V), \circledast, \Delta_{\sqcup\!\sqcup})$ is $\mathcal{L}ie(V)$. Let us now describe the Lie bracket induced on $\mathcal{L}ie(V)$ by $\circledast$.

Proposition 2

1. *Let $*$ be a magmatic product on V. The Hopf algebras $(T(V), \circledast, \Delta_{\sqcup\!\sqcup})$ and $(T(V), ., \Delta_{\sqcup\!\sqcup})$ are isomorphic, via the following algebra morphism:*

$$\phi_* : \begin{cases} (T(V), ., \Delta_{\sqcup\!\sqcup}) \longrightarrow (T(V), \circledast, \Delta_{\sqcup\!\sqcup}) \\ x_1 \dots x_k \in V^{\otimes k} \longrightarrow x_1 \circledast \dots \circledast x_k. \end{cases}$$

2. *$\mathcal{L}ie(V) * T(V) \subseteq \mathcal{L}ie(V)$. Moreover, $(\mathcal{L}ie(V), [-,-], *)$ is a post-Lie algebra. The induced Lie bracket on $\mathcal{L}ie(V)$ is denoted by $\{-,-\}_*$:*

$$\forall f, g \in \mathcal{L}ie(V),\ \{f, g\}_* = [f, g] + f * g - g * f = fg - gf + f * g - g * f.$$

The Lie algebra $(\mathcal{L}ie(V), \{-,-\}_*)$ is isomorphic to $\mathcal{L}ie(V)$.

Proof

1. There exists a unique algebra morphism $\phi_* : (T(V), .) \longrightarrow (T(V), \circledast)$, sending any $x \in V$ on itself. As the elements of V are primitive in both Hopf algebras, ϕ_* is a Hopf algebra morphism. As $V^{\otimes k} * T(V) \subseteq V^{\otimes k}$ for all $k \geq 0$, we deduce that for all $x_1, \ldots, x_{k+l} \in V$:

$$x_1 \ldots x_k \circledast x_{k+1} \ldots x_{k+l} = x_1 \ldots x_{k+l} + \text{a sum of words of length } < k+l.$$

Hence, if $x_1, \ldots, x_k \in V$:

$$\phi_*(x_1 \ldots x_k) = x_1 \circledast \ldots \circledast x_k = x_1 \ldots x_k + \text{a sum of words of length } < k.$$

Consequently:

- If $k \geq 0$ and $x_1, \ldots, x_k \in V$, an induction on k proves that $x_1 \ldots x_k \in \phi_*(T(V))$, so ϕ_* is surjective.
- If f is a nonzero element of $T(V)$, let us write $f = f_0 + \ldots + f_k$, with $f_i \in V^{\otimes i}$ for all i and $f_k \neq 0$. Then:

$$\phi_*(f) = f_k + \text{terms in } \mathbb{K} \oplus \ldots \oplus V^{\otimes(k-1)},$$

so $\phi_*(f) \neq 0$: ϕ_* is injective.

Hence, ϕ_* is an isomorphism.

2. Direct computations prove that $\mathscr{L}ie(V)$ is a post-Lie algebra. Consequently, $\{-,-\}_*$ is a second Lie bracket on $\mathscr{L}ie(V)$. In $(T(V), \circledast)$, if f and g are primitive:

$$f \circledast g - g \circledast f = fg + f * g - gf - g * f = \{f, g\}_*.$$

So, by the Cartier-Quillen-Milnor-Moore's theorem, $(T(V), \circledast, \Delta_{\sqcup\!\sqcup})$ is the enveloping algebra of $(\mathscr{L}ie(V), \{-,-\}_*)$. As it is isomorphic to the enveloping algebra of $\mathscr{L}ie(V)$, namely $(T(V), ., \Delta_{\sqcup\!\sqcup})$, these two Lie algebras are isomorphic.

Let us give a combinatorial description of ϕ_*.

Proposition 3 *Let $(V, *)$ be a magmatic algebra, and $x_1, \ldots, x_k \in V$.*

- *Let $I = \{i_1, \ldots, i_p\} \subseteq [k]$, with $i_1 < \ldots < i_p$. We put:*

$$x_I^* = (\ldots((x_{i_1} * x_{i_2}) * x_{i_3}) * \ldots) * x_{i_p} \in V.$$

- *Let P be a partition of $[p]$. We denote it by $P = \{P_1, \ldots, P_p\}$, with the convention $\min(P_1) < \ldots < \min(P_p)$. We put:*

$$x_P^* = x_{P_1}^* \ldots x_{P_p}^* \in V^{\otimes p}.$$

Then:

$$\phi^*(x_1 \dots x_k) = \sum_{P \text{ partition of } [k]} x_P^*.$$

Proof By induction on k.

Examples Let $x_1, x_2, x_3 \in V$.

$$\phi_*(x_1x_2x_3) = x_1x_2x_3 + (x_1 * x_2)x_3 + (x_1 * x_3)x_2 + x_1(x_2 * x_3) + (x_1 * x_2) * x_3.$$

Theorem 2 *Let $(V, *)$ be a magmatic algebra and let $(L, \{-,-\}, \star)$ be a post-Lie algebra. Let $\phi : (V, *) \longrightarrow (L, \star)$ be a morphism of magmatic algebras. There exists a unique morphism of post-Lie algebras $\overline{\phi} : \mathcal{L}ie(V) \longrightarrow L$ extending ϕ.*

Proof Let $\psi : \mathcal{L}ie(V) \longrightarrow L$ be the unique Lie algebra morphism extending ϕ. Let us fix $h \in \mathcal{L}ie(V)$. We consider:

$$A_h = \{h \in \mathcal{L}ie(V) \mid \forall f \in \mathcal{L}ie(V),\ \psi(f * h) = \psi(f) \star \psi(h)\}.$$

If $f, g \in A_h$, then:

$$\begin{aligned}
\psi([f, g] * h) &= \psi([f * h, b] + [f, g * h]) \\
&= \{\psi(f * h), \psi(g)\} + \{\psi(f), \psi(g * h)\} \\
&= \{\psi(f) \star \psi(h), \psi(g)\} + \{\psi(f), \psi(g) \star \psi(h)\} \\
&= \{\psi(f), \psi(g)\} \star \psi(h) \\
&= \psi([f, g]) \star \psi(h).
\end{aligned}$$

So $[f, g] \in A_h$: for all $h \in \mathcal{L}ie(V)$, A_h is a Lie subalgebra of $\mathcal{L}ie(V)$. Moreover, if $h \in V$, as $\psi_{|V} = \phi$ is a morphism of magmatic algebras, $V \subseteq A_h$; as a consequence, if $h \in V$, $A_h = \mathcal{L}ie(V)$.

Let $A = \{h \in \mathcal{L}ie(V) \mid A_h = \mathcal{L}ie(V)\}$. We put $\mathcal{L}ie(V)_n = \mathcal{L}ie(V) \cap V^{\otimes n}$; let us prove inductively that $\mathcal{L}ie(V)_n \subseteq A$ for all n. We already proved that $V \subseteq A$, so this is true for $n = 1$. Let us assume the result at all rank $k < n$. Let $h \in \mathcal{L}ie(V)_n$. We can assume that $h = [h_1, h_2]$, with $h_1 \in \mathcal{L}ie(V)_k$, $h_2 \in \mathcal{L}ie(V)_{n-k}$, $1 \leq k \leq n-1$. From Lemma 1 and Proposition 2, ${}_1f * h_2 \in \mathcal{L}ie(V)_k$ and $h_2 * h_1 \in \mathcal{L}ie(V)_{n-k}$, so the induction hypothesis holds for h_1, h_2, $h_1 * h_2$ and $h_2 * h_1$. Hence, for all $f \in T(V)$:

$$\begin{aligned}
\psi(f * h) &= \psi(f * [h_1, h_2]) \\
&= \psi((f * h_1) * h_2 - f * (h_1 * h_2) - (f * h_2) * h_1 + f * (h_2 * h_1)) \\
&= (\psi(f) \star \psi(h_1)) \star \psi(h_2) - \psi(f) \star (\psi(h_1) \star \psi(h_2))
\end{aligned}$$

$$
\begin{aligned}
&- (\psi(f) \star \psi(h_2)) \star \psi(h_1) + \psi(f) \star (\psi(h_2) \star \psi(h_1)) \\
&= \psi(f) \star \{\psi(h_1), \psi(h_2)\} \\
&= \psi(f) \star \psi(h).
\end{aligned}
$$

As a consequence, $\mathscr{L}ie(V)_n \subseteq A$. Finally, $A = \mathscr{L}ie(V)$, so for all $f, g \in \mathscr{L}ie(V)$, $\psi(f * g) = \psi(f) * \psi(g)$.

Corollary 1 *Let V be a vector space. The free magmatic algebra generated by V is denoted by $\mathscr{M}ag(V)$. Then $\mathscr{L}ie(\mathscr{M}ag(V))$ is the free post-Lie algebra generated by V.*

Remark Describing the free magmatic algebra generated by V is terms of planar rooted trees with a grafting operation, we get back the construction of free post-Lie algebras of [5].

2.3 Enveloping Algebra of a Post-Lie Algebra

Let $(V, \{-,-\}, *)$ be a post-Lie algebra. We extend $*$ onto $T(V)$ as previously in Proposition 1. The usual bracket of $\mathscr{L}ie(V) \subseteq T(V)$ is denoted by $[f, g] = fg - gf$, and should not be confused with the bracket $\{-,-\}$ of the post-Lie algebra V.

Lemma 2 *Let I be the two-sided ideal of $T(V)$ generated by the elements $xy - yx - \{x, y\}$, $x, y \in V$. Then $I * T(V) \subseteq I$ and $T(V) * I = (0)$.*

Proof Let us first prove that for all $x, y \in V$, for all $h \in T(V)$:

$$\{x, y\} * h = \sum \left\{ x * h^{(1)}, y * h^{(2)} \right\}.$$

Note that the second member of this formula makes sense, as $V * T(V) \subseteq V$ by Lemma 1.

We assume that $h = z_1 \dots z_n$ and we work by induction on n. If $n = 0$, then $h = 1$ and $\{x, y\} * 1 = \{x, y\} = \{x * 1, y * 1\}$. If $n = 1$, then $h \in V$, so $\Delta_{\sqcup\!\sqcup}(h) = h \otimes 1 + 1 \otimes h$.

$$\{x, y\} * h = \{x * h, y\} + \{x, y * h\} = \{x * h, y * 1\} + \{x * 1, y * h\} = \sum \{x * h^{(1)}, y * h^{(2)}\}.$$

If $n \geq 2$, we put $h_1 = z_1 \dots z_{n-1}$ and $h_2 = z_n$. The induction hypothesis holds for h_1, h_2 and $h_1 * h_2$:

$$
\begin{aligned}
\{x, y\} * h &= (\{x, y\} * h_1) * h_2 - \{x, y\} * (h_1 * h_2) \\
&= \sum \left\{ x * h_1^{(1)}, y * h_1^{(2)} \right\} * h_2 - \sum \left\{ x * (h_1 * h_2)^{(1)}, y * (h_1 * h_2)^{(2)} \right\}
\end{aligned}
$$

$$= \sum \left\{ \left(x * h_1^{(1)} \right) * h_2 - x * \left(h_1^{(1)} * h_2 \right), y * h_1^{(2)} \right\}$$
$$+ \sum \left\{ x * h_1^{(1)}, \left(y * h_1^{(2)} \right) * h_2 - y * \left(h_1^{(2)} * h_2 \right) \right\}$$
$$= \sum \left\{ x * h^{(1)}, y * h^{(2)} \right\}.$$

Consequently, the result holds for all $h \in T(V)$.

Let $J = Vect(xy - yx - \{x, y\} \mid x, y \in V)$. For all $x, y \in V$, for all $h \in T(V)$:

$$(xy - yx - \{x, y\}) * h$$
$$= \sum \left(x * h^{(1)} \right) \left(y * h^{(2)} \right) - \left(y * h^{(1)} \right) \left(y * h^{(2)} \right) - \left\{ x * h^{(1)}, y * h^{(2)} \right\} \in J.$$

So $J * T(V) \subseteq J$. If $g \in J$, $f_1, f_2, h \in T(V)$:

$$(f_1 g f_2) * h = \sum \left(f_1 * d^{(1)} \right) \underbrace{\left(g * h^{(2)} \right)}_{\in J} \left(f_2 * h^{(3)} \right) \in I.$$

So $I * T(V) \subseteq I$. An induction on n proves that $T(V) * (T(V) J V^{\otimes n}) = (0)$ for all $n \geq 0$. So $T(V) * I = (0)$.

As a consequence, the quotient $T(V)/I$ inherits a magmatic product $*$. Moreover, I is a Hopf ideal, and this implies that it is also a two-sided ideal for $\circledast$. As $T(V)/I$ is the enveloping algebra $\mathcal{U}(V, \{-, -\})$, we obtain Proposition 3.1 of [2]:

Proposition 4 *Let $(\mathfrak{g}, \{-, -\}, *)$ be a post-Lie algebra. Its magmatic product can be uniquely extended to $\mathcal{U}(\mathfrak{g})$ such that for all $f, g, h \in \mathcal{U}(\mathfrak{g})$, $x, y \in \mathfrak{g}$:*

- $f * 1 = f$.
- $1 * f = \varepsilon(f) 1$.
- $f * (gy) = (f * g) * y - f * (g * y)$.
- $(fg) * h = \sum \left(f * h^{(1)} \right) \left(g * h^{(2)} \right)$, *where $\Delta(h) = \sum h^{(1)} \otimes h^{(2)}$ is the usual coproduct of $\mathcal{U}(\mathfrak{g})$.*

We define a product $\circledast$ on $\mathcal{U}(\mathfrak{g})$ by $f * g = \sum \left(f * g^{(1)} \right) g^{(2)}$. Then $(\mathcal{U}(\mathfrak{g}), \circledast, \Delta)$ is a Hopf algebra, isomorphic to $\mathcal{U}(\mathfrak{g}, \{-, -\}_*)$.

Proof By Cartier-Quillen-Milnor-Moore's theorem, $(\mathcal{U}(\mathfrak{g}), \circledast, \Delta)$ is an enveloping algebra; the underlying Lie algebra is $Prim(\mathcal{U}(\mathfrak{g})) = \mathfrak{g}$, with the Lie bracket defined by:

$$\{x, y\}_{\circledast} = x \circledast y - y \circledast x = xy + x * y - yx - y * x.$$

This is the bracket $\{-, -\}_*$.

Remarks

1. If $\mathfrak{g}$ is a post-Lie algebra with $\{-,-\}=0$, it is a pre-Lie algebra, and $\mathscr{U}(\mathfrak{g}) = S(\mathfrak{g})$. We obtain again the Oudom-Guin construction [6, 7].
2. By Lemma 1, $(\mathscr{U}(\mathfrak{g}), *)$ is a right $(\mathscr{U}(\mathfrak{g}), \circledast)$-module. By restriction, $(\mathfrak{g}, *)$ is also a right $(\mathscr{U}(\mathfrak{g}), \circledast)$-module.

2.4 The Particular Case of Associative Algebras

Let $(V, \triangleleft)$ be an associative algebra. The associated Lie bracket is denoted by $[-,-]_{\triangleleft}$. As $(V, 0, \triangleleft)$ is post-Lie, the construction of the enveloping algebra of $(V, [-,-]_{\triangleleft})$ can be done: we obtain a product $\triangleleft$ defined on $S(V)$ and an associative product $\blacktriangleleft$ making $(S(V), \blacktriangleleft, \Delta)$ a Hopf algebra, isomorphic to the enveloping algebra of $(V, [-,-]_{\triangleleft})$.

Lemma 3 *If $x_1, \ldots, x_k, y_1, \ldots, y_l \in V$:*

$$x_1 \ldots x_k \triangleleft y_1 \ldots y_l = \sum_{\theta:[l]\hookrightarrow[k]} \left(\prod_{i \notin Im(\theta)} x_i \right) \left(\prod_{i=1}^{k} x_{\theta(i)} \triangleleft y_i \right),$$

$$x_1 \ldots x_k \blacktriangleleft y_1 \ldots y_l = \sum_{I \subseteq [l]} \sum_{\theta:I\hookrightarrow[k]} \left(\prod_{i \notin Im(\theta)} x_i \right) \left(\prod_{j \notin I} y_j \right) \left(\prod_{i \in I} x_{\theta(i)} \triangleleft y_i \right).$$

The notation $\theta : A \hookrightarrow [k]$ means that the sum is over all injections θ from A to $[k]$, for $A = I$ or $A = [l]$.

Proof Direct computations.

Examples Let $x_1, x_2, y_2, y_2 \in V$.

$$\begin{aligned} x_1x_2 \blacktriangleleft y_1y_2 = {} & x_1x_2y_1y_2 + (x_1 \triangleleft y_1)x_2y_2 + (x_1 \triangleleft y_2)x_2y_1 + x_1(x_2 \triangleleft y_1)y_2 \\ & + x_1(x_2 \triangleleft y_2)y_1 + (x_1 \triangleleft y_1)(x_2 \triangleleft y_2) + (x_1 \triangleleft y_2)(x_2 \triangleleft y_1). \end{aligned}$$

Remark The number of terms in $x_1 \ldots x_k \triangleleft y_1 \ldots y_l$ is:

$$\sum_{i=0}^{\min(k,l)} \binom{l}{i} \binom{k}{i} i!,$$

see sequences A086885 and A176120 of [8].

3 A Family of Solvable Lie Algebras

3.1 Definition

Definition 2 Let us fix $a = (a_1, \ldots, a_N) \in \mathbb{K}^N$. We define an associative product $\triangleleft$ on $\mathbb{K}^N$:

$$\forall i, j \in [N],\ \epsilon_i \triangleleft \epsilon_j = a_j \epsilon_i.$$

The associated Lie bracket is denoted by $[-,-]_a$:

$$\forall i, j \in [N],\ [\epsilon_i, \epsilon_j]_a = a_j \epsilon_i - a_i \epsilon_j.$$

This Lie algebra is denoted by $\mathfrak{g}_a$.

Remark Let $A \in M_{N,M}(\mathbb{K})$, and $a \in \mathbb{K}^N$. The following map is a Lie algebra morphism:

$$\begin{cases} \mathfrak{g}_{a.{}^tA} \longrightarrow \mathfrak{g}_a \\ x \longrightarrow Ax. \end{cases}$$

Consequently, if $a \neq (0, \ldots, 0)$, g_a is isomorphic to $\mathfrak{g}_{(1,0,\ldots,0)}$.

Definition 3 Let $A = T(V)^N$. The elements of A will be denoted by:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} = f_1 \epsilon_1 + \ldots + f_N \epsilon_N.$$

For all $i, j \in [N]$, we define bilinear products ${}_i\sqcup\!\sqcup$ and $\sqcup\!\sqcup_j$:

$$\forall f, g \in T(V)^N, \quad f\ {}_i\sqcup\!\sqcup\ g = \begin{pmatrix} f_i \sqcup\!\sqcup g_1 \\ \vdots \\ f_i \sqcup\!\sqcup g_N \end{pmatrix}, \quad f\ \sqcup\!\sqcup_j\ g = \begin{pmatrix} f_1 \sqcup\!\sqcup g_j \\ \vdots \\ f_N \sqcup\!\sqcup g_j \end{pmatrix}.$$

In other words, if $f, g \in T(V)$, for all $k, l \in [N]$:

$$f\epsilon_k\ {}_i\sqcup\!\sqcup\ g\epsilon_l = \delta_{i,k}(f \sqcup\!\sqcup g)\epsilon_l, \qquad f\epsilon_k\ \sqcup\!\sqcup_j\ g\epsilon_l = \delta_{j,l}(f \sqcup\!\sqcup g)\epsilon_k.$$

If $a = (a_1, \ldots, a_N) \in \mathbb{K}^N$, we put ${}_a\sqcup\!\sqcup = a_1\ {}_1\sqcup\!\sqcup + \ldots + a_N\ {}_N\sqcup\!\sqcup$ and $\sqcup\!\sqcup_a = a_1\ \sqcup\!\sqcup_1 + \ldots + a_N\ \sqcup\!\sqcup_N$.

Proposition 5 *Let $f, g \in \mathbb{K}^N$. For all $f, g, h \in A$:*

$$(f \sqcup\!\sqcup_a g) \sqcup\!\sqcup_b h = f \sqcup\!\sqcup_a (g \sqcup\!\sqcup_b h), \quad (f \sqcup\!\sqcup_a g)_b \sqcup\!\sqcup h = f \sqcup\!\sqcup_a (g_b \sqcup\!\sqcup h),$$
$$(f_a \sqcup\!\sqcup g) \sqcup\!\sqcup_b h = f_a \sqcup\!\sqcup (g \sqcup\!\sqcup_b h), \quad (f_a \sqcup\!\sqcup g)_b \sqcup\!\sqcup h = f_a \sqcup\!\sqcup (g_b \sqcup\!\sqcup h),$$
$$f \sqcup\!\sqcup_a g = g_a \sqcup\!\sqcup f.$$

Proof Direct verifications, using the associativity and the commutativity of $\sqcup\!\sqcup$.

Definition 4 Let $a \in \mathbb{K}^N$. We define a Lie bracket on A:

$$\forall f, g \in A, \ {}_a[f, g] = f_a \sqcup\!\sqcup g - g_a \sqcup\!\sqcup f = g \sqcup\!\sqcup_a f - f \sqcup\!\sqcup_a g.$$

This Lie algebra is denoted by $\mathfrak{g}'_a$.

Remark If A is an associative commutative algebra and $\mathfrak{g}$ is a Lie algebra, then $A \otimes \mathfrak{g}$ is a Lie algebra, with the following Lie bracket:

$$\forall f, g \in A, \ x, y \in \mathfrak{g}, \ [f \otimes x, g \otimes y] = fg \otimes [x, y].$$

Then, as a Lie algebra, $\mathfrak{g}'_a$ is isomorphic to the tensor product of the associative commutative algebra $(T(V), \sqcup\!\sqcup)$, and of the Lie algebra $\mathfrak{g}_{-a}$. Consequently, if $a \neq (0, \ldots, 0)$, $\mathfrak{g}'_a$ is isomorphic to $\mathfrak{g}'_{(1,0,\ldots,0)}$.

3.2 Enveloping Algebra of $\mathfrak{g}_a$

Let us apply Lemma 3 to the Lie algebra $\mathfrak{g}_a$:

Proposition 6 *The symmetric algebra $S(\mathfrak{g}_a)$ is given an associative product $\blacktriangleleft$ such that for all $i_1, \ldots, i_k, j_1, \ldots, j_l \in [N]$:*

$$\epsilon_{i_1} \ldots \epsilon_{i_k} \blacktriangleleft \epsilon_{j_1} \ldots \epsilon_{j_l} = \sum_{I \subseteq [l]} k(k-1) \ldots (k - |I| + 1) \left(\prod_{q \in I} a_{j_q} \right) \left(\prod_{p \notin I} \epsilon_{j_p} \right) \epsilon_{i_1} \ldots \epsilon_{i_k}.$$

The Hopf algebra $(S(\mathfrak{g}_a), \blacktriangleleft, \Delta)$ is isomorphic to the enveloping algebra of $\mathfrak{g}_a$.

The enveloping algebra of $\mathfrak{g}_a$ has two distinguished bases, the Poincaré-Birkhoff-Witt basis and the monomial basis:

$$(\epsilon_{i_1} \blacktriangleleft \ldots \blacktriangleleft \epsilon_{i_k})_{k \geq 0, \, 1 \leq i_1 \leq \ldots \leq i_k \leq N}, \qquad (\epsilon_{i_1} \ldots \epsilon_{i_k})_{k \geq 0, \, 1 \leq i_1 \leq \ldots \leq i_k \leq N}.$$

Here is the passage between them.

Proposition 7 *Let us fix $n \geq 1$. For all $I = \{i_1 < \ldots < i_k\} \subseteq [n]$, we put:*

$$\lambda(I) = (i_1 - 1) \ldots (i_k - k), \quad \mu(I) = (-1)^k (i_1 - 1) i_2 (i_3 + 1) \ldots (i_k + k - 2).$$

We use the following notation: if $[n] \setminus I = \{q_1 < \ldots < q_l\}$, $\prod_{q \notin I}^{\blacktriangleleft} \epsilon_{i_q} = \epsilon_{i_{q_1}} \blacktriangleleft \ldots \blacktriangleleft \epsilon_{i_{q_l}}$. Then:

$$\epsilon_{i_1} \blacktriangleleft \ldots \blacktriangleleft \epsilon_{i_n} = \sum_{I \subseteq [n]} \lambda(I) \left(\prod_{p \in I} a_{i_p} \right) \left(\prod_{q \notin I} \epsilon_{i_q} \right),$$

$$\epsilon_{i_1} \ldots \epsilon_{i_n} = \sum_{I \subseteq [n]} \mu(I) \left(\prod_{p \in I} a_{i_p} \right) \left(\prod_{q \notin I}^{\blacktriangleleft} \epsilon_{i_q} \right).$$

Proof Induction on n.

3.3 Modules Over $\mathfrak{g}_{(1,0,\ldots,0)}$

Proposition 8

1. *Let V be a module over the associative (non unitary) algebra $(\mathfrak{g}_{(1,0,\ldots,0)}, \triangleleft)$. Then $V = V^{(0)} \oplus V^{(1)}$, with:*
   - *$\epsilon_1.v = v$ if $v \in V^{(1)}$ and $\epsilon_1.v = 0$ if $v \in V^{(0)}$.*
   - *For all $i \geq 2$, $\epsilon_i.v \in V^{(0)}$ if $v \in V^{(1)}$ and $\epsilon_i.v = 0$ if $i \in V^{(0)}$.*
2. *Conversely, let $V = V^{(1)} \oplus V^{(0)}$ be a vector space and let $f_i : V^{(1)} \longrightarrow V^{(0)}$ for all $2 \leq i \leq N$. One defines a structure of $(\mathfrak{g}_{(1,0,\ldots,0)}, \triangleleft)$-module over V:*

$$\epsilon_1.v = \begin{cases} v \text{ if } v \in V^{(1)}, \\ 0 \text{ if } v \in V^{(0)}; \end{cases} \qquad \text{if } i \geq 2, \ \epsilon_i.v = \begin{cases} f_i(v) \text{ if } v \in V^{(1)}, \\ 0 \text{ if } v \in V^{(0)}. \end{cases}$$

Shortly:

$$\epsilon_1 : \begin{bmatrix} 0 & 0 \\ 0 & Id \end{bmatrix}, \qquad \forall i \geq 2, \ \epsilon_i : \begin{bmatrix} 0 & f_i \\ 0 & 0 \end{bmatrix}.$$

Proof Note that in $\mathfrak{g}_{(1,0,\ldots,0)}$, $\epsilon_i \triangleleft \epsilon_j = \delta_{1,j} \epsilon_i$.

1. In particular, $\epsilon_1 \triangleleft \epsilon_1 = \epsilon_1$. If $F_1 : V \longrightarrow V$ is defined by $F_1(v) = \epsilon_1.v$, then:

$$F_1 \circ F_1(v) = \epsilon_1.(\epsilon_1.v) = (\epsilon_1 \triangleleft \epsilon_1).v = \epsilon.v = F_1(v),$$

   so F_1 is a projection, which implies the decomposition of V as $V^{(0)} \oplus V^{(1)}$. Let $x \in V^{(1)}$ and $i \geq 2$. Then $F_1(\epsilon_i.v) = \epsilon_1.(\epsilon_i.v) = (\epsilon_1 \triangleleft \epsilon_i).v = 0$, so $\epsilon_i.v \in V^{(0)}$. Let $x \in V^{(0)}$. Then $\epsilon_i.v = (\epsilon_i \triangleleft \epsilon_1).v = \epsilon_i.F_1(v) = 0$, so $\epsilon_i.v = 0$.
2. Let $i \geq 2$ and $j \in [N]$. If $v \in V^{(1)}$:

$$\epsilon_1.(\epsilon_1.v) = v = \epsilon_1.v, \quad \epsilon_i.(\epsilon_1.v) = f_i(v) = \epsilon_i.v, \quad \epsilon_j.(\epsilon_i.v) = \epsilon_j.f_i(v) = 0.v.$$

   If $v \in V^{(0)}$:

$$\epsilon_1.(\epsilon_1.v) = 0 = \epsilon_1.v, \qquad \epsilon_i.(\epsilon_1.v) = 0 = \epsilon_i.v, \qquad \epsilon_j.(\epsilon_i.v) = 0 = 0.v.$$

   So V is indeed a $(\mathfrak{g}_{(1,0,\ldots,0)}, \triangleleft)$-module.

Proposition 9 (We assume $\mathbb{K}$ algebraically closed) *Let V be an indecomposable finite-dimensional module over the Lie algebra $\mathfrak{g}_{(1,0,\ldots,0)}$. There exists a scalar λ and a decomposition:*

$$V = V^{(0)} \oplus \ldots \oplus V^{(k)}$$

such that, for all $0 \leq p \leq k$:

- $\epsilon_1 \left(V^{(p)}\right) \subseteq V^{(p)}$ *and there exists $n \geq 1$ such that* $(\epsilon_1 - (\lambda + p)Id)^n_{|V^{(p)}} = (0)$.
- *If $i \geq 2$, $\epsilon_i \left(V^{(p)}\right) \subseteq V^{(p-1)}$, with the convention $V^{(-1)} = (0)$.*

Proof First, observe that in the enveloping algebra of $\mathfrak{g}_{(1,0,\ldots,0)}$, if $i \geq 2$ and $\lambda \in \mathbb{K}$:

$$\epsilon_i \blacktriangleleft (\epsilon_1 - \lambda) = \epsilon_i\epsilon_1 + \epsilon_i - \lambda\epsilon_i = \epsilon_i\epsilon_1 + (1 - \lambda)\epsilon_i = (\epsilon_1 - \lambda + 1) \blacktriangleleft \epsilon_i.$$

Therefore, for all $i \geq 2$, for all $n \in \mathbb{N}$, for all $\lambda \in \mathbb{K}$:

$$\epsilon_i \blacktriangleleft (\epsilon_1 - \lambda)^{\blacktriangleleft n} = (\epsilon_1 - \lambda + 1)^{\blacktriangleleft n} \blacktriangleleft \epsilon_i.$$

Let V be a finite-dimensional module over the Lie algebra $\mathfrak{g}_{(1,0,\ldots,0)}$. We denote by E_λ the characteristic subspace of eigenvalue λ for the action of ϵ_1. Let us prove that for all $\lambda \in \mathbb{K}$, if $i \geq 2$, $\epsilon_i(E_\lambda) \subseteq E_{\lambda-1}$. If $x \in E_\lambda$, there exists $n \geq 1$, such that $(\epsilon_1 - \lambda Id)^{\blacktriangleleft n}.v = 0$. Hence:

$$0 = \epsilon_i.((\epsilon_1 - \lambda Id)^n.v) = (\epsilon_1 - (\lambda - 1)Id)^n.(\epsilon_i.v),$$

so $\epsilon_i \in E_{\lambda-1}$.

Let us take now V an indecomposable module, and let Λ be the spectrum of the action of ϵ_1. The group $\mathbb{Z}$ acts on $\mathbb{K}$ by translation. We consider $\Lambda' = \Lambda + \mathbb{Z}$ and let Λ'' be a system of representants of the orbits of Λ'. Then:

$$V = \bigoplus_{\lambda \in \Lambda''} \underbrace{\left(\bigoplus_{n \in \mathbb{Z}} E_{\lambda+n} \right)}_{V_\lambda}.$$

By the preceding remarks, V_λ is a module. As V is indecomposable, Λ'' is reduced to a single element. As the spectrum of ϵ_1 is finite, it is included in a set of the form $\{\lambda, \lambda+1, \ldots, \lambda+k\}$. We then take $V^{(p)} = E_{\lambda+p}$ for all p.

Definition 5 Let V be a module over the Lie algebra $\mathfrak{g}_a$. The associated algebra morphism is:

$$\phi_V : \begin{cases} \mathcal{U}(\mathfrak{g}_a) = (S(\mathfrak{g}_a), \blacktriangleleft) \longrightarrow End(V) \\ \qquad\qquad\qquad\quad \epsilon_i \longrightarrow \begin{cases} V \longrightarrow V \\ v \longrightarrow \epsilon_i.v. \end{cases} \end{cases}$$

For all $i_1, \ldots, i_k \in [N]$, we put $F_{i_1,\ldots,i_k} = \phi_V(\epsilon_{i_1} \ldots \epsilon_{i_k})$; this does not depend on the order on the indices i_p.

By Proposition 7:

Proposition 10 *For all $i_1, \ldots, i_n \in [N]$:*

$$F_{i_1} \circ \ldots \circ F_{i_n} = \sum_{\substack{I \subseteq [n], \\ I \backslash J = \{j_1 < \ldots < j_l\}}} \lambda(I) \left(\prod_{p \in I} a_{i_p} \right) F_{i_{j_1},\ldots,i_{j_l}},$$

$$F_{i_1,\ldots,i_n} = \sum_{\substack{I \subseteq [n], \\ I \backslash J = \{j_1 < \ldots < j_l\}}} \mu(I) \left(\prod_{p \in I} a_{i_p} \right) F_{i_{j_1}} \circ \ldots \circ F_{i_{j_l}}.$$

When V is a module over the associative algebra $(\mathfrak{g}_A, \triangleleft)$, these morphisms are easy to describe:

Proposition 11 *Let V be a module over the associative algebra $(\mathfrak{g}_a, \triangleleft)$; it is also a module over the Lie algebra $(\mathfrak{g}_a, [-,-]_a)$. For all $k \geq 2$, $i_1, \ldots, i_k \in [N]$, $F_{i_1,\ldots,i_k} = 0$.*

Proof As V is a module over the associative algebra $(\mathfrak{g}_a, \triangleleft)$, for any $i_1, i_2 \in [N]$:

$$F_{i_1} \circ F_{i_2} = a_{i_2} F_{i_1}.$$

We proceed by induction on k. If $k = 2$, $\epsilon_{i_1}\epsilon_{i_2} = \epsilon_{i_1} \blacktriangleleft \epsilon_{i_2} - a_{i_2}\epsilon_{i_1}$, so:

$$F_{i_1,i_2} = F_{i_1} \circ F_{i_2} - a_{i_2}F_{i_1} = a_{i_2}F_{i_1} - a_{i_2}F_{i_1} = 0.$$

Let us assume the result at rank k. Then $\epsilon_1 \dots \epsilon_{i_{k+1}} = \epsilon_{i_1} \dots \epsilon_{i_k} \blacktriangleleft \epsilon_{i_{k+1}} - ka_{i_{k+1}}\epsilon_{i_1} \dots \epsilon_{i_k}$, and $F_{i_1,\dots,i_{k+1}} = F_{i_1,\dots,i_k} \circ F_{i_{k+1}} - ka_{i_{k+1}}F_{i_1,\dots,i_k} = 0$.

4 A Family of Post-Lie Algebras

4.1 Construction

Let us fix a vector space V, a family of N endomorphisms $(F_1, \dots, F_N)$ of V and $a = (a_1, \dots, a_N) \in \mathbb{K}^N$. We define inductively a product $*$ on $T(V)^N$: for all $f, g \in T(V)^N$, $x \in V$, $i \in [N]$,

$$\emptyset\epsilon_i * g = 0, \quad xf * g = x(f * g) + F_1(x)(f \sqcup\!\sqcup_1 g) + \dots + F_N(x)(f \sqcup\!\sqcup_N g).$$

We define a second product $\bullet$ on $T(V)^N$:

$$\forall f, g \in T(V)^N, \ f \bullet g = f * g + f \ {}_a \sqcup\!\sqcup \ g.$$

Examples Let $x, y, z \in V$, $g \in T(V)$, $i, j \in [N]$. Then:

$$\begin{aligned} x\epsilon_i * g\epsilon_j &= F_j(x)g\epsilon_j, \\ xy\epsilon_i * g\epsilon_j &= (xF_j(y)g + F_j(x)(y \sqcup\!\sqcup g))\epsilon_i, \\ xyz\epsilon_i * g\epsilon_j &= (xyF_j(z)g + xF_j(y)(z \sqcup\!\sqcup g) + F_j(x)(yz \sqcup\!\sqcup g))\epsilon_i. \end{aligned}$$

Proposition 12 *Let $x_1, \dots, x_k, y_1, \dots, y_l \in V$, $i, j \in [N]$.*

$$x_1 \dots x_k\epsilon_i * y_1 \dots y_l\epsilon_j = \sum_{\sigma \in Sh(k,l)} \sum_{p=1}^{m_k(\sigma)} \left(Id^{\otimes(p_1)} \otimes F_j \otimes Id^{\otimes(k+l-p)} \right) \sigma.(x_1 \dots x_k y_1 \dots y_l)\epsilon_i.$$

Proof Induction on k.

Remark Let $*_j$ be the pre-Lie product of $T(V, F_j)$, described in [3]. For all $f, g \in T(V)$, for all $i, j \in [N]$:

$$f\epsilon_i * g\epsilon_j = (f *_j g)\epsilon_i.$$

Corollary 2 *For all $f, g, h \in T(V)^N$, for all $i \in [N]$:*

$$(f \,{}_i\sqcup\, g) * h = (f * h) \,{}_i\sqcup\, g + f \,{}_i\sqcup\, (g * h),$$
$$(f \sqcup_i g) * h = (f * h) \sqcup_i g + f \sqcup_i (g * h),$$
$$(f \sqcup g) * h = (f * h) \sqcup g + f \sqcup (g * h).$$

Proof Induction on the length of f.

Theorem 3 *The following conditions are equivalent:*

1. $(T(V)^N, \bullet)$ *is a pre-Lie algebra.*
2. $\mathfrak{g}'_a = (T(V)^N, \,{}_a[-,-], *)$ *is a post-Lie algebra.*
3. V *is a module over the Lie algebra* $\mathfrak{g}_a$*, with the action given by* $\epsilon_i.v = F_i(v)$.

Proof By Corollary 2, for all $f, g, h \in \mathfrak{g}'_a$, $\,{}_a[f, g] * h = \,{}_a[f * h, g] + \,{}_a[f, g * h]$.

1. $\Longleftrightarrow$ 2. Let $f, g, h \in \mathfrak{g}$.

$$\begin{aligned}&(f \bullet g) \bullet h - f \bullet (g \bullet h) - (f \bullet h) \bullet g + f \bullet (h \bullet g)\\ &= (f * g) * h - f * (g * h) - (f * h) * g + f * (h * g) - f * \,{}_a[g, h].\end{aligned}$$

So $(\mathfrak{g}'_a, \bullet)$ is pre-Lie if, and only if, $(\mathfrak{g}'_a, \,{}_a[-,-], *)$ is post-Lie.

2. $\Longrightarrow$ 3. Let $x, y, v \in V$ and $i, j, k \in [N]$. Then:

$$x\epsilon_i * y\epsilon_j = F_j(x)y\epsilon_i, \qquad xy\epsilon_i * z\epsilon_k = xF_k(y)z\epsilon_i + F_k(x)(y \sqcup z)\epsilon_i.$$
$$x\epsilon_i * yz\epsilon_k = F_k(x)yz\epsilon_i,$$

Hence:

$$(x\epsilon_i * y\epsilon_j) * z\epsilon_k = F_j(x)F_k(y)z\epsilon_i + F_k \circ F_j(x)y \sqcup z\epsilon_i,$$
$$x\epsilon_i * (y\epsilon_j * z\epsilon_k) = F_j(x)F_k(y)z\epsilon_i,$$
$$x\epsilon_i \,{}_a[y\epsilon_j, z\epsilon_k] = (a_j F_k(x)(y \sqcup z) - a_k F_j(x)(y \sqcup z))\epsilon_i.$$

The post-Lie relation (2) gives:

$$(a_j F_k(x) - a_k F_j(x))(y \sqcup z) = (F_j \circ F_k - F_k \circ F_j)(x)(y \sqcup z).$$

Let $y = z$ be a nonzero element of V. Then $y \sqcup z \neq 0$, and we obtain that for all $x \in V$, $a_j F_k(x) - a_k F_j(x) = (F_j \circ F_k - F_k \circ F_j)(x)$: V is a $\mathfrak{g}_a$-module.

3. $\Longrightarrow$ 2. The post-Lie relation (2) is proved by an induction on the length of f.

Example The post-Lie algebra $\mathfrak{g}_{SISO}$ is associated to $a = (1,0)$, $V = Vect(x_1, x_2)$ and:

$$F_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

As F_1 and F_2 define a module over the Lie algebra $\mathfrak{g}_{(1,0)}$, even in fact over the associative algebra $(\mathfrak{g}_{(1,0)}, \triangleleft)$, we obtain indeed a post-Lie algebra. For all $f, g \in T(V)$, for all $i, j \in \{1, 2\}$:

$$\emptyset\epsilon_i * g\epsilon_j = 0, \qquad x_2 f\epsilon_i * g\epsilon_1 = x_2(f\epsilon_i * g\epsilon_1) + x_2(f \sqcup\!\sqcup g)\epsilon_i,$$
$$x_1 f\epsilon_i * g\epsilon_j = x_1(f\epsilon_i * g\epsilon_j), \quad x_2 f\epsilon_i * g\epsilon_2 = x_2(f\epsilon_i * g\epsilon_2) + x_1(f \sqcup\!\sqcup g)\epsilon_i.$$

4.2 Extension of the Post-Lie Product

We now extend the post-Lie product of $\mathfrak{g}'_a$ to the enveloping algebra $\mathscr{U}(\mathfrak{g}'_a)$. As this Lie bracket is obtained from an associative product $\triangleleft = {}_a\sqcup\!\sqcup$, we can see $\mathscr{U}(\mathfrak{g}'_a)$ as $(S(\mathfrak{g}'_a), \blacktriangleleft, \Delta)$. The post-Lie product $*$ is extended to $\mathscr{U}(\mathfrak{g}'_a)$, and we obtain a Hopf algebra $(\mathscr{U}(\mathfrak{g}), \circledast, \Delta)$, isomorphic to $\mathscr{U}(\mathfrak{g}'_a, {}_a[-,-]_*)$, with:

$$\forall f, g \in \mathfrak{g}, \ {}_a[f,g]_* = {}_a[f,g] + f * g - g * f = f \,{}_a\sqcup\!\sqcup\, g + f * g - g \,{}_a\sqcup\!\sqcup\, f - g * f.$$

As $\bullet$ is a pre-Lie product, it can also be extended to $S(\mathfrak{g})$ and gives a product $\odot$, making $S(\mathfrak{g}'_a)$ a Hopf algebra isomorphic to $\mathscr{U}(\mathfrak{g}'_a, [-,-]_\bullet)$.

Remark Let $f, g \in \mathfrak{g}'_a$.

$$\begin{aligned} [f,g]_\bullet &= f \bullet g - g \bullet f \\ &= f \,{}_a\sqcup\!\sqcup\, g + f * g - g \,{}_a\sqcup\!\sqcup\, f - g * f \\ &= {}_a[f,g] + f * g - g * f \\ &= {}_a[f,g]_*. \end{aligned}$$

So $[-,-]_\bullet = {}_a[-,-]_*$.

The following result allows to compute $f * g_1 \dots g_k$ by induction on the length of f:

Proposition 13 *Let $x \in V$, $k \geq 1$, $f, g_1, \dots, g_k \in T(V)^N$, $i \in [N]$.*

$$\emptyset\epsilon_i * (g_1 \blacktriangleleft \dots \blacktriangleleft g_k) = 0,$$

$$xf * (g_1 \blacktriangleleft \dots \blacktriangleleft g_k) = \sum_{\substack{I=\{i_1<\dots<i_l\}\subseteq[k],\\ j_1,\dots,j_l\in[N]}} F_{j_l} \circ \dots \circ F_{j_1}(x) \left(\left(f * \prod_{i\notin I}^{\blacktriangleleft} g_i \right) \sqcup\!\sqcup_{j_1} g_{i_1} \dots \sqcup\!\sqcup_{j_l} g_{i_l} \right);$$

$$\emptyset \epsilon_i * (g_1 \dots g_k) = 0,$$

$$xf * (g_1 \dots g_k) = \sum_{\substack{I=\{i_1<\dots<i_l\}\subseteq[k],\\ j_1,\dots,j_l\in[N]}} F_{j_1,\dots,j_l}(x) \left(\left(f * \prod_{i\notin I} g_i \right) \sqcup\!\!\sqcup_{j_1} g_{i_1} \dots \sqcup\!\!\sqcup_{j_l} g_{i_l} \right).$$

Proof Induction on k.

Proposition 14 *Let $k \geq 1$, $f, g_1, \dots, g_k \in T(V)^N$. Then:*

$$f \bullet g_1 \dots g_k = f * g_1 \dots g_k + \sum_{p=1}^{k} (f * g_1 \dots g_{p-1} g_{p+1} \dots g_k)\, {}_a\!\sqcup\!\!\sqcup\, g_p.$$

Proof Induction on k.

Proposition 15 *On $S(\mathfrak{g}'_a)$, $\circledast = \odot$.*

Proof Let $f, g \in S(\mathfrak{g}'_a)$; let us prove that $f \circledast g = f \odot g$. We assume that $f = f_1 \dots f_k$, $g = g_1, \dots, g_l$, with $f_1, \dots, f_k, g_1, \dots, g_l \in \mathfrak{g}'_a$, and we proceed by induction on k. If $k = 0$, then $f = 1$ and $f \circledast g = f \odot g = g$. Let us assume the result at all ranks $< k$. We proceed by induction on l. If $l = 0$, then $g = 1$ and $f \circledast g = f \odot g = f$. Let us assume the result at all ranks $< l$. We put:

$$\Delta(f) = f \otimes 1 + 1 \otimes f + f' \otimes f'', \qquad \Delta(g) = g \otimes 1 + 1 \otimes g + g' \otimes g''.$$

The induction hypothesis on k holds for f' and f'' and the induction hypothesis on l holds for g' and g''. From:

$$\Delta(f \circledast g - f \odot g) = f^{(1)} \circledast g^{(1)} \otimes f^{(2)} \circledast g^{(2)} - f^{(1)} \odot g^{(1)} \otimes f^{(2)} \odot g^{(2)},$$

these two induction hypotheses give:

$$\Delta(f \circledast g - f \odot g) = (f \circledast g - f \odot g) \otimes 1 + 1 \otimes (f \circledast g - f \odot g).$$

So $f \circledast g - f \odot g \in Prim(S(\mathfrak{g}'_a)) = \mathfrak{g}'_a$. We obtain:

$$\pi(f \circledast g) = \pi \left(\sum_{I\subseteq[l]} \left(f * \prod_{i\in I} g_i \right) \blacktriangleleft \prod_{j\notin I} g_j \right)$$

$$= \pi \left(\sum_{[l]=I_0\sqcup\dots\sqcup I_k} \left(f_1 * \prod_{i\in I_1} g_i \right) \dots \left(f_k * \prod_{i\in I_k} g_i \right) \blacktriangleleft \prod_{i\in I_0} g_i \right)$$

$$= \pi\left(\sum_{[l]=J_1\sqcup\ldots\sqcup J_k}\prod_{p=1}^{k}\left(f_p * \prod_{i\in J_k} g_i + \sum_{j_p\in J_p}\left(f_p * \prod_{i\in J_p\setminus\{j_p\}} g_i\right) a \sqcup\!\sqcup\, g_{j_p}\right)\right)$$

$$= \pi\left(\sum_{[l]=J_1\sqcup\ldots\sqcup J_k}\left(\prod_{p=1}^{k} f_p \bullet \prod_{i\in J_p} g_i\right)\right)$$

$$= \pi\left(\left(f_1 \bullet g^{(1)}\right)\ldots\left(f_k \bullet g^{(k)}\right)\right)$$

$$= \pi(f \bullet g)$$

$$= \pi(f \odot g).$$

As $f \circledast g - f \odot g \in \mathfrak{g}'_a$, $f \circledast g = f \odot g$.

4.3 Graduation

We assume in this whole paragraph that $a = (1, 0, \ldots, 0)$ and V is finite-dimensional. We decompose the $\mathfrak{g}_a$-module V as a direct sum of indecomposables. By Proposition 9, decomposing each indecomposables, we obtain a decomposition of V of the form:

$$V = V^{(0)} \oplus \ldots \oplus V^{(k)},$$

with $F_1\left(V^{(p)}\right) \subseteq V^{(p)}$ and $F_i\left(V^{(p)}\right) \subseteq V^{(p-1)}$ for all $i \geq 2$, for all $p \in [k]$. We put $V_p = V^{(k+1-p)}$ for all $p \in [k+1]$. This defines a graduation of V, which induces a connected graduation of $T(V)$. For this graduation of V, F_1 is homogeneous of degree 0 and F_i is homogeneous of degree 1 for all $i \geq 2$. We define a graduation of $\mathfrak{g}'_a = T(V)^N$:

$$\forall n \geq 0,\ (\mathfrak{g}'_a)_n = T(V)_n\epsilon_1 \oplus \bigoplus_{i=2}^{N} T(V)_{n-1}\epsilon_i.$$

For this graduation, the product ${}_{(1,0,\ldots,0)}\sqcup\!\sqcup$ is homogeneous of degree 0. Proposition 12 implies that $*$ is homogeneous of degree 0; summing, $\bullet$ is also homogeneous of degree 0. Hence:

Proposition 16 *The decomposition of V in indecomposable $\mathfrak{g}_{(1,0,\ldots,0)}$-modules induces a graduation of the post-Lie algebra $\mathfrak{g}'_{(1,0,\ldots,0)}$.*

We put:

$$P(X) = \sum_{i=1}^{k+1} dim(V_p) X^p \in \mathbb{K}[X].$$

the formal series of $\mathfrak{g}'_{(1,0,\ldots,0)}$ is:

$$\begin{aligned} R(X) &= \sum_{p=1}^{\infty} dim((\mathfrak{g}'_{(1,0,\ldots,0)})_p) X^p \\ &= \frac{1}{1-P(X)} + (N-1)\frac{X}{1-P(X)} = \frac{1+(N-1)X}{1-P(X)}. \end{aligned}$$

Note that $R(0) = 1$: indeed, $(\mathfrak{g}'_{(1,0,\ldots,0)})_0 = Vect(\emptyset\epsilon_1)$. The augmentation ideal of $\mathfrak{g}'_{(1,0,\ldots,0)}$ is:

$$(\mathfrak{g}'_{(1,0,\ldots,0)})_+ = T(V)_+ \times T(V)^{N-1}.$$

This is a graded, connected post-Lie algebra.

Example For the SISO case, $V_1 = Vect(x_2)$ and $V_2 = Vect(x_1)$. The formal series of $\mathfrak{g}_{SISO}$ is:

$$R_{SISO}(X) = \frac{1+X}{1-X-X^2} = 1 + 2X + 3X^2 + 5X^3 + 8X^4 + 13X^5 + \ldots$$

Hence, $(dim(\mathfrak{g}_{SISO})_n)_{n\geq 0}$ is the Fibonacci sequence A000045 [8]. For example:

$$\begin{aligned} (\mathfrak{g}_{SISO})_0 &= Vect(\emptyset\epsilon_1), \\ (\mathfrak{g}_{SISO})_1 &= Vect(x_2\epsilon_1, \emptyset\epsilon_2), \\ (\mathfrak{g}_{SISO})_2 &= Vect(x_1\epsilon_1, x_2x_2\epsilon_1, x_2\epsilon_2), \\ (\mathfrak{g}_{SISO})_3 &= Vect(x_1x_2\epsilon_1, x_2x_1\epsilon_1, x_2x_2x_2\epsilon_1, x_1\epsilon_2, x_2x_2\epsilon_2). \end{aligned}$$

5 Graded Dual

We assume in this section that $a = (1, 0, \ldots, 0)$. The augmentation ideal of $\mathfrak{g}'_a$ is denoted by $(\mathfrak{g}'_a)_+$; recall that $(\mathfrak{g}'_a)_0 = Vect(\emptyset\epsilon_1)$.

- As $(\mathfrak{g}'_a)_+$ is a graded, connected Lie algebra, its enveloping algebra $\mathcal{U}((\mathfrak{g}'_a)_+)$ is a graded, connected Hopf algebra, and its graded dual also is. We denote it by $\mathcal{H}_V$.

- As an algebra, $\mathscr{H}_V$ is identified with $S((\mathfrak{g}'_a)^*)/\langle \emptyset\epsilon_1\rangle$. We identify $(\mathfrak{g}'_a)^*$ with $T(V^*)^N$ via the pairing:

$$\langle f_1 \dots f_k\epsilon_i, x_1 \dots x_l\epsilon_j\rangle = \delta_{i,j}\delta_{k,l} f_1(x_1)\dots f_k(x_k).$$

- The coproduct dual of $\odot = \circledast$ is denoted by $\Delta_\bullet$.
- The dual of the product $\sqcup\!\sqcup_j$ defined on $\mathfrak{g}'_a$ is denoted by $\Delta_{\sqcup\!\sqcup_j}$, defined on $(\mathfrak{g}'_a)^* = T(V^*)^N$.
- We define a coproduct Δ_* on $S((\mathfrak{g}'_a)^*_+)$, dual of the right action $*$. Therefore, this is right coaction of $(\mathscr{H}_V, \Delta_\bullet)$ on itself:

$$(\Delta_* \otimes Id)\circ\Delta_* = (Id\otimes\Delta_\bullet)\circ\Delta_*.$$

Notations

1. For all $y \in V^*$, we define $\theta_y : (\mathfrak{g}'_a)^* \longrightarrow (\mathfrak{g}'_a)^*$ by $\theta_y(f) = yf$.
2. For all $x \in (\mathscr{H}_V)_+$, we put $\overline{\Delta}_\bullet(x) = \Delta_\bullet(x) - 1\otimes x$ and $\overline{\Delta}_*(x) = \Delta_*(x) - 1\otimes x$. For all $g, f, f_1, \dots, f_k \in (\mathfrak{g}'_a)^*_+$:

$$\langle \overline{\Delta}_*(g), f\otimes f_1\dots f_k\rangle = \langle g, f * f_1\dots f_k\rangle.$$

5.1 Deshuffling Coproducts

Proposition 17 *For all $g \in T(V)$, for all $i \in [N]$, $\Delta_{\sqcup\!\sqcup_j}(g\epsilon_k) = \Delta_{\sqcup\!\sqcup}(g)(\epsilon_k\otimes\epsilon_j)$.*

Proof Let $f_1, f_2 \in T(V)$, $i_1, i_2 \in [N]$.

$$\begin{aligned}\langle \Delta_{\sqcup\!\sqcup_j}(g\epsilon_k), f_1\epsilon_{i_1}\otimes f_2\epsilon_{i_2}\rangle &= \langle g\epsilon_k, f_1\epsilon_{i_1} \sqcup\!\sqcup_j f_2\epsilon_{i_2}\rangle\\ &= \delta_{i_2,j}\langle g\epsilon_k, f_1\sqcup\!\sqcup f_2\epsilon_{i_1}\rangle\\ &= \delta_{i_2,j}\delta_{i_1,k}\langle g, f_1\sqcup\!\sqcup f_2\rangle\\ &= \delta_{i_2,j}\delta_{i_1,k}\langle \Delta_{\sqcup\!\sqcup}(g), f_1\otimes f_2\rangle\\ &= \langle \Delta_{\sqcup\!\sqcup}(g)(\epsilon_k\otimes\epsilon_j), f_1\epsilon_{i_1}\otimes f_2\epsilon_{i_2}\rangle.\end{aligned}$$

As the pairing is nondegenerate, we obtain the result.

Notations We define inductively, for $l \geq 0$, $j_1, \dots, j_l \in [N]$:

$$\begin{cases}\Delta_{\sqcup\!\sqcup_\emptyset} = Id,\\ \Delta_{\sqcup\!\sqcup_{j_1,\dots,j_l}} = \left(\Delta_{\sqcup\!\sqcup_{j_1}}\otimes Id^{\otimes(l-1)}\right)\circ\Delta_{\sqcup\!\sqcup_{j_2,\dots,j_l}}.\end{cases}$$

For all $g \in T(V^*)$, for all $i \in [N]$:

$$\Delta_{\sqcup_{j_1,\dots,j_l}}(g\epsilon_k) = \Delta^{(l)}_{\sqcup}(g)(\epsilon_k \otimes \epsilon_{j_1} \otimes \dots \otimes \epsilon_{j_l});$$

for all $f_1, \dots, f_l \in T(V)$:

$$\langle \Delta_{\sqcup_{j_1,\dots,j_l}}(g), f_1 \otimes \dots \otimes f_{l+1} \rangle = \langle g, f_1 \sqcup_{j_1} \dots \sqcup_{j_l} f_{l+1} \rangle.$$

5.2 Dual of the Post-Lie Product

Dualizing:

Proposition 18 *In $\mathscr{H}_V = S((\mathfrak{g}'_a)^*)/\langle \emptyset\epsilon_1 \rangle$:*

- *For all $i \in [N]$, $\Delta_*(\emptyset\epsilon_i) = \emptyset\epsilon_i \otimes 1 + 1 \otimes \emptyset\epsilon_i$.*
- *For all $y \in V^*$, $g \in (\mathfrak{g}'_a)^*$:*

$$\overline{\Delta}_* \circ \theta_y(g) = \sum_{l \geq 0} \sum_{j_1,\dots,j_l \in [N]} (\theta_{F^*_{j_1,\dots,j_l}(y)} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id) \circ \Delta_{\sqcup_{j_1,\dots,j_l}}(g),$$

where we denote by μ the sum of the iterated products of $\mathscr{H}_V$:

$$\mu : \begin{cases} T(\mathscr{H}_V) \longrightarrow \mathscr{H}_V \\ g_1 \otimes \dots \otimes g_k \longrightarrow g_1 \dots g_k. \end{cases}$$

In order to obtain a better description of the coproduct $\overline{\Delta}_*$, we are going to identify the following three objects:

$$\begin{array}{ccc} & S((\mathfrak{g}'_a)^*_+) & \\ \swarrow^{\sim} & & \searrow^{\sim} \\ S((\mathfrak{g}'_a)^*)/\langle \emptyset\epsilon_1 \rangle & & S((\mathfrak{g}'_a)^*)/\langle \emptyset\epsilon_1 - 1 \rangle \end{array}$$

Both identification sends $x \in (\mathfrak{g}'_a)^*_+$ to its class. Let us reformulate Proposition 18 in the vector space $S((\mathfrak{g}'_a)^*)/\langle \emptyset\epsilon_1 - 1 \rangle$:

$$\begin{aligned} \overline{\Delta}_* \circ \theta_y(g\epsilon_k) = & \sum_{l \geq 0} \sum_{j_1,\dots,j_l \in [N]} (\theta_{F^*_{j_1,\dots,j_l}(y)} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta^{(l)}_{\sqcup}(g)\epsilon_k \otimes \epsilon_{j_1} \otimes \dots \otimes \epsilon_{j_l}) \\ & - \left(\sum_{l \geq 0} \sum_{j_1,\dots,j_l \in [N]} (\theta_{F^*_{j_1,\dots,j_l,1}(y)} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta^{(l)}_{\sqcup}(g)\epsilon_k \otimes \epsilon_{j_1} \otimes \dots \otimes \epsilon_{j_l}) \right) (1 \otimes \emptyset\epsilon_1). \end{aligned}$$

Finally, identifying in $S((\mathfrak{g}'_a)^*_+)$:

Proposition 19 *For all $j_1, \dots, j_l \in [N]$, we put:*

$$G_{j_1,\dots,j_l} = F_{j_1,\dots,j_l} - F_{j_1,\dots,j_l,1}.$$

In $S((\mathfrak{g}'_a)^*_+)/\langle \emptyset\epsilon_1 - 1\rangle$:

- *For all $i \in [N]$, $\overline{\Delta}_*(\emptyset\epsilon_i) = \emptyset\epsilon_i \otimes 1$.*
- *For all $y \in V^*$, for all $g \in (\mathfrak{g}'_a)^*_+$:*

$$\overline{\Delta}_* \circ \theta_y(g) = \sum_{l\geq 0} \sum_{j_1,\dots,j_l \in [N]} (\theta_{G^*_{j_1,\dots,j_l}(y)} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id) \circ \Delta_{\sqcup\!\sqcup_{j_1,\dots,j_l}}(g).$$

Example For $\mathfrak{g}_{SISO}$, as V is a module over the associative algebra $(\mathfrak{g}_{(1,0)}, \triangleleft)$, if $l \geq 2$, $F_{j_1,\dots,j_l} = 0$ by Proposition 11, so $G_{j_1,\dots,j_l} = 0$. Moreover:

$$F_\emptyset = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad F_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

$$G_\emptyset = F_\emptyset - F_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad G_1 = F_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad G_2 = F_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

$$G^*_\emptyset = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad G^*_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad G^*_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

The coproduct $\overline{\Delta}_*$ on $S((\mathfrak{g}_{SISO})^*_+)$ is given by:

- For all $i \in [2]$, $\overline{\Delta}_*(\emptyset\epsilon_i) = \emptyset\epsilon_i \otimes 1$.
- For all $g \in \mathbb{K}\langle x_1, x_2\rangle$, for all $i \in [2]$:

$$\overline{\Delta}_* \circ \theta_{x_1}(g\epsilon_i) = (\theta_{x_1} \otimes Id) \circ \overline{\Delta}_*(g\epsilon_i) + (\theta_{x_2} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)\epsilon_i \otimes \epsilon_2),$$

$$\overline{\Delta}_* \circ \theta_{x_2}(g\epsilon_i) = (\theta_{x_2} \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)\epsilon_i \otimes \epsilon_1).$$

These are formulas of Lemma 4.1 of [9], where $a_w = w\epsilon_2$, $b_w = w\epsilon_1$, $\theta_0 = \theta_{x_1}$, $\theta_1 = \theta_{x_2}$ and $\tilde{\Delta} = \overline{\Delta}_*$.

5.3 Dual of the Pre-Lie Product

Notations We denote by $\Delta_{1\sqcup\!\sqcup}$ the coproduct on $T_+(V^*) \otimes (V)^{N-1}$ dual to the product ${}_1\sqcup\!\sqcup$. As ${}_1\sqcup\!\sqcup = \sqcup\!\sqcup_1{}^{op}$, $\Delta_{1\sqcup\!\sqcup} = \Delta^{cop}_{\sqcup\!\sqcup_1}$, and for all $g \in T(V)$, for all

$i \in [N]$:

$$\Delta_{1 \sqcup\!\sqcup}(g\epsilon_i) = \Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_k).$$

Dualizing:

Proposition 20 *In $S((\mathfrak{g}'_a)^*_+)/\langle \emptyset\epsilon_1 \rangle$, for all $g \in (\mathfrak{g}'_a)^*_+$:*

$$\overline{\Delta}_\bullet(g) = \overline{\Delta}_*(g) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id) \circ \Delta_{1 \sqcup\!\sqcup}(g).$$

Rewriting this formula in $S((\mathfrak{g}'_a)^*_+)/\langle \emptyset\epsilon_1 - 1 \rangle$:

$$\begin{aligned}
\overline{\Delta}_\bullet(g\epsilon_1) &= \overline{\Delta}_*(g\epsilon_1) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_1)) \\
&= \overline{\Delta}_*(g\epsilon_1) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)((\Delta_{\sqcup\!\sqcup}(g) - g \otimes \emptyset)(\epsilon_1 \otimes \epsilon_1)) \\
&= \overline{\Delta}_*(g\epsilon_1)(1 \otimes (1 - \emptyset\epsilon_1)) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_1)) \\
&= (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_1)).
\end{aligned}$$

Identifying in $S((\mathfrak{g}'_a)^*_+)$:

Proposition 21 *In $S((\mathfrak{g}'_a)^*_+)/\langle \emptyset\epsilon_1 - 1 \rangle$, if $g \in T(V^*)$:*

$$\overline{\Delta}_\bullet(g\epsilon_1) = (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_1 \otimes \epsilon_1)),$$

$$\text{if } i \geq 2,\ \overline{\Delta}_\bullet(g\epsilon_i) = \overline{\Delta}_*(g\epsilon_i) + (Id \otimes \mu) \circ (\overline{\Delta}_* \otimes Id)(\Delta_{\sqcup\!\sqcup}(g)(\epsilon_i \otimes \epsilon_1)),$$

with the convention $\emptyset\epsilon_1 = 1$. We put $\Delta_\bullet(g) = \overline{\Delta}_\bullet(g) + 1 \otimes g$ for all $g \in (\mathfrak{g}'_a)^*_+$ and extend $\Delta_\bullet$ to $S((\mathfrak{g}'_a)^*_+)$ as an algebra morphism. This coproduct makes $S((\mathfrak{g}'_a)^*_+)$ a Hopf algebra, isomorphic to the graded dual of the enveloping algebra of $((\mathfrak{g}'_a)_+, [-,-]_*)$.

Remark These are *mutatis mutandis* the formulas of Lemma 4.3 in [9].

Acknowledgements The research leading these results was partially supported by the French National Research Agency under the reference ANR-12-BS01-0017.

References

1. Cartier, P.: Vinberg algebras, Lie groups and combinatorics. In: Quanta of Maths, Clay Mathematics Proceedings, vol. 11, pp. 107–126. American Mathematical Society, Providence (2010)
2. Ebrahimi-Fard, K., Lundervold, A., Munthe-Kaas, H.Z.: On the Lie enveloping algebra of a post-Lie algebra. J. Lie Theory **25**(4), 1139–1165 (2015)
3. Foissy, L.: A pre-Lie algebra associated to a linear endomorphism and related algebraic structures. Eur. J. Math. **1**(1), 78–121 (2015). arXiv:1309.5318

4. Manchon, D.: A short survey on pre-Lie algebras. In: Carey, A.L. (ed.) Noncommutative Geometry and Physics: Renormalisation, Motives, Index Theory. Lectures in Mathematics and Physics, pp. 89–102. European Mathematical Society, Zürich (2011)
5. Munthe-Kaas, H.Z., Lundervold, A.: On post-Lie algebras, Lie-Butcher series and moving frames. Found. Comput. Math. **13**(4), 583–613 (2013). arXiv:1203.4738
6. Oudom, J.-M., Guin, D.: Sur l'algèbre enveloppante d'une algèbre pré-Lie. C. R. Math. Acad. Sci. Paris **340**(5), 331–336 (2005)
7. Oudom, J.-M., Guin, D.: On the Lie enveloping algebra of a pre-Lie algebra. J. K-Theory **2**(1), 147–167 (2008). arXiv:math/0404457
8. Sloane, N.J.A.: On-line encyclopedia of integer sequences. http://oeis.org/
9. Steven Gray, W., Ebrahimi-Fard, K.: SISO output affine feedback transformation group and its Faà di Bruno Hopf algebra. SIAM J. Control Optim. **55**(2), 885–912 (2017)
10. Vallette, B.: Homology of generalized partition posets. J. Pure Appl. Algebra **208**(2), 699–725 (2007). arXiv:math/0405312

Infinite Dimensional Rough Dynamics

Massimiliano Gubinelli

Abstract We review recent results about the analysis of controlled or stochastic differential systems via local expansions in the time variable. This point of view has its origin in Lyons' theory of rough paths and has been vastly generalised in Hairer's theory of regularity structures. Here our concern is to understand this local expansions when they feature genuinely infinite dimensional objects like distributions in the space variable. Our analysis starts reviewing the simple situation of linear controlled rough equations in finite dimensions, then we introduce unbounded operators in such linear equations by looking at linear rough transport equations. Loss of derivatives in the estimates requires the introduction of new ideas, specific to this infinite dimensional setting. Subsequently we discuss how the analysis can be extended to systems which are not intrinsically rough but for which local expansion allows to highlight other phenomena: in our case, regularisation by noise in linear transport. Finally we comment about other application of these ideas to fully-nonlinear conservations laws and other PDEs.

1 Introduction

In this short note we want to review recent results in the analysis of the rough dynamics of certain partial differential equations (PDEs). The adjective *rough* refers to the fact that the description of such dynamics does not rely on differential equations but on local expansion in the time variable. This shift of point of view has originated in the seminal work of Lyons on *rough paths* [16] and subsequent developments [7, 12, 13]. Rough path theory deals with the study of controlled differential systems under the action of non–smooth inputs. The low regularity of the input signals does not allow a differential description of the change in time of the system. Instead, in order to describe such systems, one has to rely on truncated

M. Gubinelli (✉)
IAM and Hausdorff Center for Mathematics, Bonn, Germany
e-mail: gubinelli@iam.uni-bonn.de

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_14

series expansions of the solutions. The short time description allow weaker norms to be used to control analytically the problem. Sometimes they also encode essential informations on the dynamics which cannot be recovered from the differential description. Consistency conditions links the various terms of these expansions and make rough paths a theory where analysis has to be supplemented by algebraic and combinatoric considerations. Recently Hairer [15] has given a vast generalization of *rough paths* by showing how to provide a local (space–time) description of the relations encoded in PDEs via the notion of *regularity structures*.

In order to keep the analysis at a relatively simple level we will not discuss regularity structures and we will stick to rough dynamics of PDEs. That is we will focus on dynamics in which we will need only a local expansion in time to give a detailed description of the system. The spatial dependence of the system will be described in a classical (infinite dimensional) setting, e.g. by means of Banach spaces of functions or distributions.

2 Linear Rough Equations

Consider the following controlled linear differential equation in $\mathbb{R}^N$

$$\mathrm{d}y(t) = A_\alpha y(t)\mathrm{d}x^\alpha(t), \qquad y(0) = y_0 \tag{1}$$

where $y \in C([0,1];\mathbb{R}^N)$ is the unknown, $x \in C^1([0,1];\mathbb{R}^M)$ is the control and $(A_\alpha)_{\alpha=1,\dots,M}$ a family of linear transformations of $\mathbb{R}^N$ (summation over repeated indexes is implied). In order for this formulation to make sense we need that x is differentiable (or at least of bounded variation). This formulation is useless if we want to study the behaviour of y when we feed as input x an approximation B^ε of a sample path of a Brownian motion B (for example) and try to remove the approximation by taking the limit $B^\varepsilon \to B$. In order to gain a description which can be taken along in this limit we resort to a series expansion of the solution

$$y(t) = y_0 + A_\alpha y_0 \int_0^t \mathrm{d}x^\alpha(s) + A_\alpha A_\beta y_0 \int_0^t \int_0^s \mathrm{d}x^\beta(u)\mathrm{d}x^\alpha(s) + \cdots \tag{2}$$

where the $\cdots$ stays for terms featuring higher order *iterated integrals* of the control x. Iterated integrals $\mathbb{X}^{\alpha_1\cdots\alpha_n}(s,t)$ are defined recursively by

$$\mathbb{X}^{\alpha_1}(s,t) = \int_s^t \mathrm{d}x^{\alpha_1}(r), \qquad \mathbb{X}^{\alpha_1\cdots\alpha_{n+1}}(s,t) = \int_s^t \mathbb{X}^{\alpha_1\cdots\alpha_n}(s,r)\mathrm{d}x^{\alpha_{n+1}}(r).$$

In the series expansion the flow property of Eq. (1) is encoded by *Chen's relations* among the iterated integrals [3]:

$$\mathbb{X}^{\alpha_1\cdots\alpha_n}(s,t) = \sum_{k=0}^{n} \mathbb{X}^{\alpha_1\cdots\alpha_k}(s,u)\mathbb{X}^{\alpha_{k+1}\cdots\alpha_n}(u,t) \tag{3}$$

where we let $\mathbb{X}^{\alpha_i \cdots \alpha_j}(s,u) = 1$ if $i \geqslant j$. One key insight of rough path theory is the fact that the series expansion can be truncated at some level and still provide enough information to reconstruct the function y if we can guarantee that the remainder is small enough. Namely if we assume that there exists $\gamma > 0$ such that

$$|\mathbb{X}^{\alpha_1 \cdots \alpha_k}(s,t)| \leqslant C|t-s|^{k\gamma}, \tag{4}$$

for all $k = 0, \ldots, n$ where n is such that $(n+1)\gamma > 1$ then there exists only one continuous function y subject to the initial condition $y(0) = y_0$ and such that

$$y(t) - y(s) = (A_\alpha \mathbb{X}^\alpha(s,t) + \cdots + A_{\alpha_n} \cdots A_{\alpha_1} \mathbb{X}^{\alpha_1 \cdots \alpha_n}(s,t))y(s) + O(|t-s|^{(n+1)\gamma}) \tag{5}$$

for all $0 \leqslant s \leqslant t$. This formulation shows that the input x affects the solution y only via the iterated integrals $\mathbb{X}$. The given of a family of maps $(\mathbb{X}^{\alpha_1 \cdots \alpha_n}(s,t))_{s,t,(\alpha_i)}$ satisfying (3) and (4) up to a certain level n defines a γ-Hölder *rough path*. In many situations this allows to plug into the equation very general inputs x for which a suitable rough path $\mathbb{X}$ can be identified. For example, quite general approximations of the Brownian motions B^ε give rise to iterated integrals $\mathbb{B}^\varepsilon$ which converge in the appropriate Hölder–like topology to the step-2 (Stratonovich) Brownian rough path $\mathbb{B}$ above the Brownian motion B (i.e. such that its first component $\mathbb{B}^\alpha = B^\alpha$ for $\alpha = 1, \ldots, M$) [12].

Equation (5) has the form

$$\delta y(s,t) = G(s,t) + y^\natural(s,t), \qquad |y^\natural(s,t)| \lesssim |t-s|^z \tag{6}$$

where $\delta y(s,t) := y(t) - y(s)$ for some $z > 1$. The key argument in the analysis is the observation that Eq. (6) is *rigid*, in the sense that bounds on G determines both bounds on δy and $y^\natural$ in a unique way. Indeed given G it can exists at most one pair $(y, y^\natural)$ solving this equation. Explitic bounds on $y^\natural$ depends on the *coherence* δG of G, namely on the combination

$$\delta G(s,u,t) = G(s,t) - G(s,u) - G(u,t), \qquad s \leqslant u \leqslant t.$$

In particular we have a *sewing lemma*:

Lemma 1 (Sewing Lemma) *Assume that there exists a constant L such that*

$$|\delta G(s,u,t)| \leqslant L|t-s|^z, \qquad s \leqslant u \leqslant t, \tag{7}$$

for some $z > 1$. Then there exists a unique y such that Eq. (6) holds and moreover there exists a universal constant C_z such that

$$|y^\natural(s,t)| \leqslant C_z L|t-s|^z, \qquad s \leqslant t.$$

Remark 1 Similar results hold for general regular *controls* $\omega(s,t)$ replacing $|t-s|$ in the above estimates, namely functions for which

$$\omega(s,u)+\omega(u,t)\leqslant\omega(s,t),\qquad s\leqslant u\leqslant t,$$

$$|\omega(s,t)|\to 0,\qquad \text{as } |t-s|\to 0.$$

If we call *coherent* a germ G which satisfy Eq. (7) then the sewing lemma essentially states that coherent germs can be uniquely *integrated*, that is there exists a unique function y (up to constants) for which the germ gives the local expansion up to an error of size $|t-s|^{1+}$.

In the case of Eq. (5) the germ depends itself on y and reads

$$G(s,t)=(A_\alpha\mathbb{X}^\alpha(s,t)+\cdots+A_{\alpha_n}\cdots A_{\alpha_1}\mathbb{X}^{\alpha_1\cdots\alpha_n}(s,t))y(s),\qquad s<t.$$

Its coherence δG can be computed via simple algebraic manipulations and Chen's relations (3) for the iterated integrals $\mathbb{X}$. Using the regularity hypothesis (4) to control the size of the iterated integrals and the Eq. (5) to replace the instances of δy we can reduce the estimate of the coherence to a control of the size of y and of $y^\natural$ where $y^\natural(s,t)$ denotes the $O(|t-s|^{(n+1)\gamma})$ term in the r.h.s. of Eq. (5). Namely,

$$\sup_{s<u<t}\frac{|\delta G(s,u,t)|}{|s-t|^z}\leqslant C_{\mathbb{X},A}\left[\sup_t|y(t)|+\sup_{s<t}\frac{|y^\natural(s,t)|}{|t-s|^z}\right],$$

where $z=(n+1)\gamma>1$. An application of the sewing lemma gives

$$\sup_{s<t}\frac{|y^\natural(s,t)|}{|t-s|^z}\leqslant C_{\mathbb{X},A,z}\left[\sup_t|y(t)|+\sup_{s<t}\frac{|y^\natural(s,t)|}{|t-s|^z}\right].$$

In order to be able to conclude a bound on $y^\natural$ from this relation we need a small constant in front of the $y^\natural$ contribution in the r.h.s. This can be accomplished in many ways, one possibility is to localize the above estimates over intervals $|s-t|\leqslant\tau$ where $\tau>0$ is a small constant. Careful bookkeping shows that this gains at least a power of τ^γ:

$$\sup_{s,t:|t-s|\leqslant\tau}\frac{|y^\natural(s,t)|}{|t-s|^z}\leqslant C_{\mathbb{X},A,z}\left[\sup_t|y(t)|+\tau^\gamma\sup_{s,t:|t-s|\leqslant\tau}\frac{|y^\natural(s,t)|}{|t-s|^z}\right]$$

and choosing τ small enough (depending on $\mathbb{X}$, A, z) we obtain the key bound

$$\sup_{s,t:|t-s|\leqslant\tau}\frac{|y^\natural(s,t)|}{|t-s|^z}\leqslant 2C_{\mathbb{X},A,z}\sup_t|y(t)|, \tag{8}$$

where $C_{\mathbb{X},A,z}$ is a constant which depends on the norm of A as a bounded operator and on that of $\mathbb{X}$ as a γ-Hölder rough path. From a this bound and an approximation argument via ODEs existence and uniqueness of solutions to (5) easily follows.

3 Unbounded Drivers

There is no substantial difficulty in generalising the rough equation (5) to an infinite dimensional setting. We can take y as a path in a Banach space V and x in another space W and $(A_\alpha)_\alpha$ as a suitable bounded linear maps. Care must be exercised in the construction of the iterated integral taking values in appropriate completion of the algebraic tensor products $W^{\otimes n}$ in order for the pairings with the powers of A_α to make sense in (5) (see e.g. [12]). In a different direction, we can consider the situation where the operators $(A_\alpha)_\alpha$ themselves are not bounded. For simplicity we will assume that x still takes values in a finite–dimensional space $\mathbb{R}^M$ and to be definite we concentrate on the case where y solves the linear transport equation

$$\begin{cases} \mathrm{d}y(t,\xi) = V_\alpha(\xi)\nabla y(t,\xi)\mathrm{d}x^\alpha(t), \\ y(0,\xi) = y_0(\xi), \end{cases} \qquad \xi \in \mathbb{R}^d, t \geqslant 0 \tag{9}$$

where $y \in C([0,1]; L^2(\mathbb{R}^d))$, $y_0 \in L^2(\mathbb{R}^d)$, $(V_\alpha\nabla)_{\alpha=1,\dots,M}$ is a family of bounded vector fields on $\mathbb{R}^d$. Here we do not want to assume smoothness of the solution (which in general will not hold) and the action of the vectorfields $V_\alpha\nabla$ is understood in the weak sense by integrating this relation agains smooth functions of the space variable ξ. When x is smooth this equation can be understood as a standard transport type PDE or in the L^2 context as a differential equation involving the unbounded, type dependent, family of operators $H(t) = \dot{x}^\alpha(t)V_\alpha\nabla$ for $t \in [0,1]$. Uniqueness of solutions holds under a Lipshitz condition on V (for example via the method of caracteristics).

It is not at all obvious how to describe solutions y in such a way that x appears only via iterated integrals $\mathbb{X}$ as in the finite dimensional setting. Following formally the above series expansion strategy we can still obtain the Eq. (5). We will consider only the case when $\gamma > 1/3$ in order to simplify some formulas. Our discussion will retain the basic features of the general problem. The equation for y has to be understood as a distributional equality:

$$\delta y(s,t)(\varphi) = y(s)((\mathbb{A}^{1,*}(s,t) + \mathbb{A}^{2,*}(s,t))\varphi) + y^\natural(s,t)(\varphi) \tag{10}$$

where $\varphi \in C^\infty(\mathbb{R}^d)$ is a compactly supported test function, $\delta y(s,t) = y(t) - y(s)$, $y(t)(\varphi)$ denotes the pairing of φ with the L^2 function $y(t)$ given by the L^2 scalar product and $\mathbb{A}^n(s,t) = X^{\alpha_1\cdots\alpha_n}(V_{\alpha_1\cdots\alpha_n}\nabla)$ is a family of linear operators indexed by s,t and $\mathbb{A}^{n,*}(s,t)$ denotes its adjoint with respect to the above pairing. In Eq. (10)

we assume that $|y^\natural(s,t)(\varphi)| \leqslant C_\varphi |t-s|^{3\gamma}$ for any $\varphi \in C^\infty(\mathbb{R}^d)$ where the constant can depends on φ. Note that $3\gamma > 1$ by assumption.

Equation (10) describes how y varies with t as a distribution modulo a remainder term $|y^\natural(s,t)(\varphi)|$. From this information is not clear if and how it is possible to recover $y \in C([0,1]; L^2(\mathbb{R}^d))$ given an initial condition y_0. We call the family of (unbounded) operators $(\mathbb{A}^n)_{n=1,2}$ an *unbounded rough driver* [1]. It satisfies operator Chen's relations

$$\mathbb{A}^1(s,t) = \mathbb{A}^1(s,u) + \mathbb{A}^1(u,t), \qquad \mathbb{A}^2(s,t) = \mathbb{A}^2(s,u) + \mathbb{A}^2(u,t) + \mathbb{A}^1(s,u)\mathbb{A}^1(u,t).$$

In a recent joint work with I. Bailleul [1] we show that Eq. (10) uniquely determines a function $y \in C([0,1]; L^2(\mathbb{R}^d))$ assuming C^3 regularity of the vector fields V (but it should be possible to drop the regularity to C^ρ for any ρ such that $\rho\gamma > 1$). One main technical tool is the generalization of (8) to an a priori bound in spaces of distributions with given regularity. We were able to show that

$$\sup_{s,t:|t-s|\leqslant\tau} \frac{|y^\natural(s,t)(\varphi)|}{|t-s|^{3\gamma}} \leqslant C_{\gamma,\tau}(\mathbb{X}, A)[\sup_t \|y(t)\|_{L^2}]\|\varphi\|_{W^{3,2}} \tag{11}$$

where $(W^{k,2})_{k\in\mathbb{R}}$ denotes the scale of Sobolev spaces on $\mathbb{R}^d$. This does not follows immediately from the sewing lemma due to a loss in derivatives. Let us explain this in more detail. We will assume to have an a priori bound for y in L^2: $\sup_t \|y(t)\|_{L^2} \leqslant 1$. The germ of Eq. (10) is given by

$$G(s,t) = y(s)((\mathbb{A}^{1,*}(s,t) + \mathbb{A}^{2,*}(s,t))\varphi). \tag{12}$$

Its coherence δG can be computed as

$$\delta G(s,u,t) = -(\delta y(s,u) - y(s)\mathbb{A}^{1,*}(s,u))(\mathbb{A}^{1,*}(u,t)\varphi) - \delta u(s,u)(\mathbb{A}^{2,*}(u,t)\varphi).$$

Using Eq. (10) we rewrite this as

$$\delta G(s,u,t) = -y^\natural(s,u)((\mathbb{A}^{1,*}(u,t) + \mathbb{A}^{2,*}(u,t))\varphi)$$
$$-y(s)((\mathbb{A}^{2,*}(s,u)\mathbb{A}^{1,*}(u,t) + \mathbb{A}^{1,*}(s,u)\mathbb{A}^{2,*}(u,t) + \mathbb{A}^{2,*}(s,u)\mathbb{A}^{2,*}(u,t))\varphi)$$

which gives, using the regularity of the unbounded driver and the a priori bound for $\|y(t)\|_{L^2}$,

$$\|\delta G(s,u,t)\|_{(W^{3,2})^*} \lesssim_{\mathbb{A}} (1 + \tau^\gamma \|y^\natural\|_{2\gamma,(W^{2,2})^*} + \tau^\gamma \|y^\natural\|_{\gamma,(W^{1,2})^*})|t-s|^{3\gamma}, \quad s < u < t \tag{13}$$

where

$$\|y^\natural\|_{\alpha,V} = \sup_{s<t:|t-s|\leqslant\tau} \frac{\|y^\natural(s,t)\|_V}{|t-s|^\alpha}.$$

An application of the sewing lemma allows to conclude that

$$\|y^\natural\|_{3\gamma,(W^{3,2})^*} \lesssim_{\mathbb{A}} [1+\tau^\gamma \|y^\natural\|_{2\gamma,(W^{2,2})^*} + \tau^\gamma \|y^\natural\|_{\gamma,(W^{1,2})^*}], \tag{14}$$

which is an estimate which cannot be closed due to loss of derivatives in the l.h.s. (we need one derivative more on test functions to estimate the action of $y^\natural$ in l.h.s. w.r.t. the r.h.s. of the equation). This loss of derivatives has to be compensated thanks to a gain of time regularity via an interpolation argument and the last estimate becomes

$$\|y^\natural\|_{3\gamma,(W^{3,2})^*} \lesssim_{\mathbb{A}} [1+\tau^\gamma \|y^\natural\|_{3\gamma,(W^{3,2})^*}] \tag{15}$$

which, after a standard reasoning, can be converted into uniform a priori estimates for $y^\natural$ of the form

$$\|y^\natural\|_{\gamma,(W^{1,2})^*} + \|y^\natural\|_{2\gamma,(W^{2,2})^*} + \|y^\natural\|_{3\gamma,(W^{3,2})^*} \lesssim_{\mathbb{A}} 1. \tag{16}$$

I will sketch now the idea behind the interpolation leading to (15). The loss of regularity preventing to close the estimate (14) is mainly due to the use of Eq. (10) in order to simplify some terms in δG as we seen above. On the other hand the reader can easily check that the terms in δG with higher loss of regularity come also with better time regularity (meaning that they feature larger powers of the time increment $|t-s|$). All this time regularity is then wasted in the estimates. In order to prevent the loss of space regularity we split the test function φ as $\varphi = J_\varepsilon\varphi + \bar{J}_\varepsilon\varphi$ where J_ε is a regularisation operator (for example by convolution at scale ε) and $\bar{J}_\varepsilon = 1 - J_\varepsilon$ a remainder term. Then δG applied to φ is decomposed as

$$\delta G(s,u,t)(\varphi) = G(s,t)\bar{J}_\varepsilon\varphi - G(s,u)\bar{J}_\varepsilon\varphi - G(u,t)\bar{J}_\varepsilon\varphi + \delta G(s,u,t)J_\varepsilon\varphi.$$

The first three terms are estimated directly from the definition of the germ G given in Eq. (12), where there is no loss of derivatives at the price of insufficient time regularity since these terms are much bigger than $|t-s|^{3\gamma}$. However this can be compensated by the convergence of the approximations assuming the natural estimate $\|\bar{J}_\varepsilon\varphi\|_{W^{k,2}} \lesssim \varepsilon^{3-k}\|\varphi\|_{W^{3,2}}$. The remaining contribution $\delta G(s,u,t)J_\varepsilon\varphi$ is estimated as in Eq. (13) where the loss of derivatives is compensated by the regularization producing diverging factors of ε^{-1}. Here however the better time regularity can be used to compensate for this divergence. Overall one chooses $\varepsilon = |t-s|^\gamma$ and check that this results in the estimate (15).

These a priori estimates on the solutions are a key step in the analysis. At variance with the Banach space setting here we cannot rely on a fixpoint argument

to prove existence and uniqueness of solutions to (10) due to the same loss of derivatives in the equation. However the a priori bound (16) can be used together with an approximation procedure to prove existence of solutions via a compactness argument. Uniqueness for distributional solutions to (10) derives from a study of the dynamics of $y^2(t,\xi) = y(t,\xi)^2$. Let us sketch this classical argument. Assume u is an L^2 solution to the transport equation

$$\partial_t u = V \cdot \nabla u .$$

With a formal computation we deduce that the function u^2 satisfies the same transport equation

$$\partial_t u^2 = 2u\partial_t u = 2uV \cdot \nabla u = V \cdot \nabla u^2 .$$

Integrating over space we get

$$\partial_t \int u_t^2 = \int (V \cdot \nabla u_t^2) = -\int (\operatorname{div} V) u_t^2 ,$$

and Gronwall lemma allows to conclude that

$$\int u_t^2 \leqslant \left(\int u_0^2 \right) \exp(t \| \operatorname{div} V \|_{L^\infty}),$$

which in particular implies uniqueness for L^2 solutions since the equation is linear. This proof has two key elements: the possibility to obtain the dynamics of u^2 and the Gronwall estimate. Even in this classical setting the weak formulation however cannot be directly used to compute the dynamics of u^2. This is a classical problem for weak solutions and a standard approach is to use a convolutional smoothing u_ε of u in order to be able to use it as a test function. The convergence as $\varepsilon \to 0$ depends on a *commutator lemma* and ultimately on sufficient regularity for the vectorfield V which dictates sufficient conditions for uniqueness.

To reproduce this line of proof for the rough dynamics (10) we need to redo the commutator argument and find a replacement for the Gronwall lemma. Commutator arguments depends on an approximation procedure. Essentially in the same spirit, but maybe conceptually clearer, we can proceed instead to "splitting" the product $y(t,\xi)^2$ and analyse the dynamics of the tensorized quantity $Y(t,\xi,\xi') = y(t,\xi)y(t,\xi')$. This function solves another rough PDE:

$$\delta Y(s,t) = (\Gamma^1_{\mathbb{A}}(s,t) + \Gamma^2_{\mathbb{A}}(s,t))Y(s) + Y^\natural(s,t) \tag{17}$$

where the tensorized driver $\Gamma_{\mathbb{A}}$ is defined in terms of $\mathbb{A}$ as

$$\Gamma^1_{\mathbb{A}}(s,t) = \mathbb{I} \otimes \mathbb{A}^1(s,t) + \mathbb{A}^1 \otimes \mathbb{I}(s,t)$$

$$\Gamma^2_{\mathbb{A}}(s,t) = \mathbb{I} \otimes \mathbb{A}^2(s,t) + \mathbb{A}^2 \otimes \mathbb{I}(s,t) + \mathbb{A}^1(s,t) \otimes \mathbb{A}^1(s,t)$$

and where we understand that factor in the left and in the right of the tensor product acts respectively on the ξ or ξ' variable. In order to recover informations on y^2 from Y we need to test with functions Φ_ε of the form $\Phi_\varepsilon(\xi,\xi') = \varphi(\xi_+)\psi(\xi_-/\varepsilon)\varepsilon^{-d}$ where here $\xi_\pm = (\xi \pm \xi')/2$ are coordinated parallel $(+)$ and transverse $(-)$ to the diagonal $\xi = \xi'$ in the space $(\xi,\xi') \in \mathbb{R}^2$. With this choice of test function we have

$$y(t)^2(\varphi) = \int y(t)^2(x)\varphi(x)\mathrm{d}x = \lim_{\varepsilon\to 0}\iint y(t)(x)y(t)(y)\varepsilon^{-d}\psi\left(\frac{x-y}{\varepsilon}\right)\varphi\left(\frac{x+y}{2}\right)\mathrm{d}x\mathrm{d}y$$

guaranteeing that $\lim_{\varepsilon\to 0} Y(t)(\Phi_\varepsilon) = y(t)^2(\varphi)$ for all $t \geqslant 0$. In this limit however the functions Φ_ε become singular on the diagonal and suitable estimates are needed to control the remainder $Y^\natural(s,t)(\Phi_\varepsilon)$. A careful analysis of the argument needed to get the a priori estimates (11) shows that a key property in order to be able to pass to the limit is that quantities of the form $\Gamma^{1,*}_{\mathbb{A}}(s,t)\Phi_\varepsilon$ (and other similar objects) stays bounded as $\varepsilon \to 0$. We called this property *renormalizability* (inspired by a similar construction in the work of DiPerna and Lions [10] on renormalised solutions for transport equations). Derivatives in the ξ_+ directions do not pose any problem. The derivatives in the ξ_- direction can be source of problems when ξ_- is very small. Fortunately, as some computations shows, these derivatives are always paired with factors of the form $V(\xi)-V(\xi')$ (or with derivatives of V) and which go to zero with ξ_- if V is sufficiently smooth compensating the divergences coming from transverse derivatives of Φ_ε. This allows to pass to the limit in Eq. (17) applied to Φ_ε and prove the convergence of germs

$$Y(s)(\Gamma^{1,*}_{\mathbb{A}}(s,t)\Phi_\varepsilon + \Gamma^{2,*}_{\mathbb{A}}(s,t)\Phi_\varepsilon) \to y(s)^2(\mathbb{A}^{1,*}(s,t)\varphi + \mathbb{A}^{2,*}(s,t)\varphi)$$

showing that $y^2 \in C([0,1];L^1(\mathbb{R}^d))$ satisfies again Eq. (10) but in L^1 instead of L^2. Namely

$$\delta y^2(s,t)(\varphi) = y(s)^2(\mathbb{A}^{1,*}(s,t)\varphi + \mathbb{A}^{2,*}(s,t)\varphi) + O(|t-s|^{3\gamma})\|\varphi\|_{W^{3,\infty}}.$$

If the vectorfields V_α are divergence free, testing now with the function $\varphi(\xi) = 1$ gives $y(s)^2(\mathbb{A}^{1,*}(s,t)\varphi + \mathbb{A}^{2,*}(s,t)\varphi) = 0$ and therefore $\delta y^2(s,t)(\varphi) = O(|t-s|^{3\gamma})$ from which we deduce that $y_t^2(1)$ is constant in t and since $y_0^2(1) = 0$, $y_t = 0$ for all $t \geq 0$. Consider now the case where the vectorfields are not divergence free but $div\, V_\alpha \in L^\infty$. In this case we have, always with $\varphi(\xi) = 1$,

$$|(V_\alpha\cdot\nabla)^*\varphi(\xi)| \lesssim \varphi(\xi), \quad |(V_\alpha\cdot\nabla)^*(V_\beta\cdot\nabla)^*\varphi(\xi)| \lesssim_V \varphi(\xi).$$

Then from

$$\delta y^2(s,t)(\varphi) = y(s)^2(\mathbb{A}^{1,*}(s,t)\varphi + \mathbb{A}^{2,*}(s,t)\varphi) + O(|t-s|^{3\gamma})$$

we get

$$h(t) \leqslant h(s) + h(s)|t-s|^{\gamma} + C(\sup_{r \leqslant t} h(r))|t-s|^{3\gamma},$$

where $h(t) := y(t)^2(\varphi)$. This inequality is powerful enough (a *rough Gronwall* lemma) to conclude that

$$\sup_{t \leqslant T} h(t) \lesssim_T h(0),$$

and from there conclude uniqueness.

4 Regularization by Noise in Transport Equations

Rough dynamics can appear also in the study of interactions of two more regular dynamics. In this section we illustrate the case of the regularisation by noise of transport equation. Let V be a vectorfield and x the sample path of a d-dimensional Brownian motion. The equation for $y \in C([0,1], L^{\infty}(\mathbb{R}^d))$:

$$\mathrm{d}y(t,\xi) = V(\xi)\nabla y(t,\xi)\mathrm{d}t + \nabla y(t,\xi) \cdot \mathrm{d}x(t), \qquad y(0,\xi) = y_0(\xi) \tag{18}$$

can be easily understood as a Stratonovich SPDE or, using a rough path $\mathbb{X}$ over x as a rough transport equation as in the previous section. Here we are interested in the problem of uniqueness of solutions when we do not require V to be Lipshitz. In this case it is known that even when $x = 0$ this PDE can have many L^{∞} weak solutions corresponding to non–uniqueness of solutions to the ODE for the characteristics. In collaboration with F. Flandoli and E. Priola [11], we showed that if x is a Brownian motion then any V which is of some, arbitrarily small, Hölder regularity, give rise to a unique solution of Eq. (18) when interpreted as a Stratonovich SPDE. The proof proceeds via a detailed study of the flow of stochastic characteristics. A more intrinsic approach is provided by an appropriate rough dynamical point of view on the same equation. In particular we make the change of variables $z(t,\xi) = y(t, \xi - x(t))$ which can be interpreted as an exact integration of the stochastic vector field $\dot{x}(t) \cdot \nabla$. The equation for z is now a standard PDE

$$\mathrm{d}z(t,\xi) = V(\xi - x(t))\nabla z(t,\xi)\mathrm{d}t, \qquad z(0,\xi) = y_0(\xi - x(0)). \tag{19}$$

This PDE however is not well posed in L^{∞} since V is not Lipshitz. The stochastic shift however should provide some sort of regularization similar to a convolution with a smooth kernel. In order to highlight this stochastic effect we can introduce an unbounded driver $\mathbb{A}^1(s,t)$ as the time average of $V(\xi - x(t))\nabla$:

$$\mathbb{A}^1(s,t) = \int_s^t V(\xi - x(r))\nabla \mathrm{d}r.$$

We can reformulate the ODE as a rough distributional equation

$$\delta z(s,t) = \mathbb{A}^1(s,t)z(s) + z^\sharp(s,t). \tag{20}$$

where we truncate the expansion at level 1 and require that the remainder satisfies $|z^\sharp(s,t)(\varphi)| \lesssim |t-s|^{2\gamma}\|\varphi\|_{W^{2,\infty}}$ and where we choose $\gamma > 1/2$. The rough dynamics (20) encodes the behaviour of z only in terms of the rough driver $\mathbb{A}$ provided we can show that $\mathbb{A}^1(s,t) \lesssim |t-s|^\gamma$. Note that this driver is quite regular in the time variable, less so in the space one. A direct stochastic computation shows, quite surprisingly, that if V is α–Hölder in ξ then for almost every sample path of a Brownian motion B the random field

$$F(t,\xi) = \int_0^t V(\xi - B(r))\mathrm{d}r,$$

belongs locally to $C_t^\gamma C_x^{\alpha+1-\varepsilon}$ for some small ε. Here we see at work a *regularisation by noise* phenomenon where, due to the averaging in time along the trajectories of the Brownian motion, the random field F gains almost one degree of regularity in Hölder spaces with respect to V. The price to pay for this effect is in the time regularity: F is just γ Hölder in time if we consider it as a $(\alpha+1-\varepsilon)$–Hölder functions in the space variable. This effect has been noted first by Davie [8], for further results and applications see [2].

Going back to the rough dynamics (20) it is possible to show that if $\alpha > 1/2$ the rough driver $\mathbb{A}^1$ is renormalizable and from that deduce uniqueness via a tensorization argument in L^∞ (a bit different from the previous one which was set up in L^2). This result does not recover the conclusions of our work with Flandoli and Priola (where uniqueness was proven for any $\alpha > 0$) but being a completely deterministic argument it has other good features (e.g. contraction and stability of the flow for the SPDE) which are not know in the stochastic setting and usually nontrivial to obtain.

In [4, 6], in collaboration with K. Chouk, we used rough dynamics to study randomly modulated non–linear dispersive equations with ideas similar to those exposed in this section.

5 Other Rough Dynamics

In the recent paper [9], in collaboration with A. Deya, M. Hofmanová and S. Tindel, we introduce the rough dynamical point of view in the study of fully nonlinear scalar stochastic conservation laws of the form

$$\mathrm{d}u(t,\xi) = \mathrm{div}(A_\alpha(t,x,u(t,\xi)))\mathrm{d}x^\alpha.$$

By introducing the *kinetic function* $f(t,\xi,v) = \mathbb{I}_{v<u(t,\xi)}$ this equation can be transformed in a rough linear kinetic equation

$$\mathrm{d}f(t,\xi,v) = V_\alpha(\xi,v)\nabla_{\xi,v} f(t,\xi,v)\mathrm{d}x^\alpha + \partial_v \mathrm{d}_t m \tag{21}$$

where V is a family of vector fields which depends on the original non–linearity A and m is a *kinetic measure* which comes as the price to pay to have linearized the equation and which essentially ensures that solutions to this linear equation have the form $\mathbb{I}_{v<u(t,\xi)}$. A solution of the kinetic equation is a pair (f,m) satisfying the finite–increment version of Eq. (21). Uniqueness for kinetic solution relies on fine properties of the measure m but the general scheme of proof goes via tensorization and passage to the diagonal much like in the linear transport case. The lack of precise informations about the kinetic measure m is source of some complications, much like in the classical setting. Test functions in the tensorization argument has to be chosen properly in order to be able to estimate the action of the kinetic measure, for details refer to the original paper.

Other partial results deal with the description of the rough dynamics for viscosity solutions of fully non–linear SPDEs [14] or with an attempt to the rough analysis of stochastic wave equations with multiplicative noise [5].

References

1. Bailleul, I., Gubinelli, M.: Unbounded rough drivers. Annales de la facultè des sciences Mathématiques de Toulouse **26**(4), 795–830 (2017). https://doi.org/10.5802/afst.1553
2. Catellier, R., Gubinelli, M.: Averaging along irregular curves and regularisation of ODEs. Stoch. Process. Appl. **126**(8), 2323–2366 (2016). https://doi.org/10.1016/j.spa.2016.02.002
3. Chen, K.T.: Iterated path integrals. Bull. Am. Math. Soc. **83**(5), 831–879 (1977). https://doi.org/10.1090/S0002-9904-1977-14320-6
4. Chouk, K., Gubinelli, M.: Nonlinear PDEs with modulated dispersion II: Korteweg–de Vries equation (2014). http://arxiv.org/abs/1406.7675. arXiv:1406.7675
5. Chouk, K., Gubinelli, M.: Rough sheets (2014). arXiv:1406.7748
6. Chouk, K., Gubinelli, M.: Nonlinear PDEs with modulated dispersion I: nonlinear Schrödinger equations. Commun. Partial Differ. Equ. **40**(11), 2047–2081 (2015)
7. Davie, A.M.: Differential equations driven by rough paths: an approach via discrete approximation. Appl. Math. Res. Express. AMRX (2) **40**, Art. ID abm009 (2007)
8. Davie, A.M.: Uniqueness of solutions of stochastic differential equations. Int. Math. Res. Not. IMRN (24) **26**, Art. ID rnm124 (2007). https://doi.org/10.1093/imrn/rnm124
9. Deya, A., Gubinelli, M., Hofmanová, M., Tindel, S.: A priori estimates for rough PDEs with application to rough conservation laws. arXiv:1604.00437 [math] (2016). arXiv:1604.00437
10. DiPerna, R.J., Lions, P.L.: Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**(3), 511–547 (1989). https://doi.org/10.1007/BF01393835
11. Flandoli, F., Gubinelli, M., Priola, E.: Well-posedness of the transport equation by stochastic perturbation. Invent. Math. **180**(1), 1–53 (2010). https://doi.org/10.1007/s00222-009-0224-4
12. Friz, P.K., Hairer, M.: A Course on Rough Paths: with an Introduction to Regularity Structures. Universitext. Springer, Cham (2014)
13. Gubinelli, M.: Controlling rough paths. J. Funct. Anal. **216**(1), 86–140 (2004). https://doi.org/10.1016/j.jfa.2004.01.002

14. Gubinelli, M., Tindel, S., Torrecilla, I.: Controlled viscosity solutions of fully nonlinear rough PDEs (2014). arXiv:1403.2832
15. Hairer, M.: A theory of regularity structures. Invent. Math. **198**(2), 269–504 (2014). https://doi.org/10.1007/s00222-014-0505-4
16. Lyons, T.J.: Differential equations driven by rough signals. Rev. Mat. Iberoamericana **14**(2), 215–310 (1998). https://doi.org/10.4171/RMI/240

Heavy Tailed Random Matrices: How They Differ from the GOE, and Open Problems

Alice Guionnet

Abstract Since the pioneering works of Wishart and Wigner on random matrices, matrices with independent entries with finite moments have been intensively studied. Not only it was shown that their spectral measure converges to the semi-circle law, but fluctuations both global and local were analyzed in fine details. More recently, the domain of universality of these results was investigated, in particular by Erdos-Yau et al and Tao-Vu et al. This survey article takes the opposite point of view by considering matrices which are not in the domain of universality of Wigner matrices: they have independent entries but with heavy tails. We discuss the properties of these matrices. They are very different from Wigner matrices: the limit law of the spectral measure is not the semi-circle distribution anymore, the global fluctuations are stronger and the local fluctuations may undergo a transition and remain rather mysterious.

1 Introduction

Random matrices were introduced by Wishart [36] in the twenties to study large arrays of data and then in the 1950s by Wigner [35] to model Hamiltonians of quantum systems. In both cases, it appeared natural to assume the dimension of the matrices to be large. Moreover, it is natural to take the entries as independent as possible within the known constraints of the model. A typical model for such a matrix is to take a symmetric matrix filled with independent equidistributed Gaussian random variables: the so-called Gaussian orthogonal ensemble (GOE). To fix the ideas, the matrix will be $N \times N$ with independent centered Gaussian entries with variance $1/N$ (and variance $2/N$ on the diagonal). The properties of

A. Guionnet (✉)
Université de Lyon, CNRS, ENSL, Lyon, France
e-mail: aguionne@ens-lyon.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_15

the spectrum and the eigenvectors of such matrices were studied in details, thanks to the fact that the law of such matrices is invariant under multiplication by orthogonal matrices and that the law of the eigenvalues has a simple expression as particles in Coulomb-gas interaction. Understanding how the details of the model could influence the spectral properties of the random matrices then became a central question. Assuming the entries to be still independent, it was shown that if the entries have sufficiently light tails, the fluctuations of the extreme eigenvalues are similar to that of the GOE [30] and in the bulk [21]. A series of remarkable works then focussed on obtaining optimal assumptions on the tails, which are a finite fourth (respectively the second) moment to observe the same fluctuations [18, 20, 25, 33]. However, there are matrices of interest which do not belong to this domain of universality. Typically, these matrices will have most entries which are very tiny, but a finite number of entries per row or column will be of order one. This is in contrast with light tails matrices where all entries are of order $1/\sqrt{N}$. An example is given by the adjacency matrix of an Erdös-Rényi graph whose entries are independent Bernoulli variables which are equal to one with probability c/N for some finite constant c. Such matrices are much less known. We shall in this note outline the main results and open problems related to such matrices. Roughly, the convergence of the empirical measure of the eigenvalues can be derived under rather general assumptions, but the limiting measure is not anymore the semi-circle law and is not compactly supported [6, 7, 37]. Under slightly more demanding hypotheses, the central limit theorem around this convergence can be derived: fluctuations occur in larger scale than for light tail matrices, in fact the usual central limit rescaling by a square root of the dimension is needed as soon as the moment of order two of the entries is infinite [9]. Local law could be derived only for α-stable entries [12, 13]. It shows a transition in the regime where $\alpha < 1$: for small eigenvalues the eigenvectors are delocalized whereas for large eigenvalues they are localized, again a phenomenon which does not occur for light tail matrices. Even words in independent heavy tail matrices behave differently: they are not asymptotically free in general and one need a new non-commutative framework, namely traffic distributions, to analyze them.

In the sequel, a Wigner matrix will be a symmetric matrix with centered independent equidistributed entries. The case of Hermitian matrices with complex entries is similar but will not be treated here for simplicity. We will denote $\mathbf{X}$ a Wigner matrix with light tails and $\mathbf{A}$ a Wigner matrix with heavy tails.

2 Macroscopic Limit

Going back to Wigner [35], it was shown that the spectral measure of random matrices with light tail entries converges towards the same asymptotic law: the semi-circle law. In this section, we discuss the convergence of the spectral measure of random matrices with heavy tails matrices and show that it converges towards

different limiting measures. Let us be more precise. Let $\mathbf{X}^N$ be a symmetric matrix so that $(X_{ij}^N)_{i\leq j}$ are independent and such that

$$\mathbb{E}[\mathbf{X}_{ij}^N] = 0, \ \lim_{N\to\infty} \max_{1\leq i,j\leq N} |N\mathbb{E}[|\mathbf{X}_{ij}^N|^2] - 1| = 0. \tag{1}$$

Assume moreover that for all $k \in \mathbb{N}$ we have

$$B_k := \sup_{N\in\mathbb{N}} \sup_{(i,j)\in\{1,\cdots,N\}^2} \mathbb{E}[|\sqrt{N}\mathbf{X}_{ij}^N|^k] < \infty. \tag{2}$$

Then, Wigner proved the almost sure convergence

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{Tr}\left((\mathbf{X}^N)^k\right) = \begin{cases} 0 & \text{if k is odd,} \\ C_{\frac{k}{2}} & \text{otherwise,} \end{cases} \tag{3}$$

where $C_{k/2} = \frac{\binom{k}{\frac{k}{2}}}{\frac{k}{2}+1}$ are the Catalan numbers. The proof is based on an expansion of the trace of moments of matrices in terms of the entries, together with the observation that the indices which will contribute to the first order of this expansion can be described by rooted trees. Based on the fact that the Catalan numbers are the moments of the semi-circle distribution

$$\sigma(dx) = \frac{1}{2\pi}\sqrt{4-x^2}1_{|x|\leq 2}dx. \tag{4}$$

one can use density arguments (see e.g. [4]) to show that as soon as B_3 (in fact "$B_{2+\varepsilon}$") is finite, the eigenvalues $(\lambda_1, \cdots, \lambda_N)$ of $\mathbf{X}^N$ satisfy the almost sure convergence

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(\lambda_i) = \int f(x)d\sigma(x) \tag{5}$$

where f is a bounded continuous function.

In contrast, the limit my be different as soon as $B_{2+\varepsilon}$ is infinite. The new hypothesis is that all moments are of order $1/N$: Assume that $\mathbb{E}[\mathbf{A}_{ij}^N] = 0$ and

$$\lim_{N\to\infty} N\mathbb{E}[(\mathbf{A}_{ij}^N)^{2k}] = M_k, \qquad \forall k \in \mathbb{N}. \tag{6}$$

Note that this includes the case of the adjacency matrix of a Erdös-Rényi graph with $M_k = c$ for all k. Then, Zakharevich [37] showed that $N^{-1}\,\mathrm{Tr}((\mathbf{A}^N)^p)$ goes to zero if p is odd and

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\mathrm{Tr}((\mathbf{A}^N)^{2k})\right] = \sum_{G=(V,E)\in\mathcal{T}_k} \sum_{P\in P_k(G)} \prod_{e\in E} M_{m(P,e)/2}\,. \tag{7}$$

where $\mathcal{T}_k$ is the set of rooted trees with at most k edges, $P_k(G)$ the set of closed paths on G with $2k$ steps, going through all edges of G, starting from the root and $m(P,e)$ the (even) number of times that the path goes through the edge e. The probability measure with the above moments is very different from the semi-circle law: in general it has unbounded support. One can generalize this result to the case where the entries have no moments at all by using convergence of the Stieltjes transform $G_\mu(z) = \int (z-x)^{-1} d\mu(x)$. Assume that the law μ_N of A_{ij}^N satisfies

$$\lim_{N\to\infty} N\left(\int (e^{-iux^2} - 1) d\mu_N(x)\right) = \Phi(u) \tag{8}$$

with Φ such that there exists g on $\mathbb{R}^+$, with $g(y)$ bounded by Cy^κ for some $\kappa > -1$, such that for $u \in \mathbb{C}^-$,

$$\Phi(u) = \int_0^\infty g(y) e^{\frac{iy}{u}} dy. \tag{9}$$

An example is given by α stable laws with $\Phi(u) = c(iu)^{\alpha/2}$ and $g(y) = Cy^{\alpha/2-1}$ for some constants c, C. Another example is provided by the adjacency matrix of Erdös-Rényi graph with $\Phi(u) = c(e^{iu} - 1)$ and g a Bessel function [9]. Then, it was shown in [7, 9] that $G_N(z) = \frac{1}{N}\mathrm{Tr}(z - \mathbf{A}^N)^{-1}$ converges almost surely towards G given by, for $z \in \mathbb{C}^+$

$$G(z) = i \int e^{itz} e^{\rho_z(t)} dt \tag{10}$$

where $\rho_z : \mathbb{R}^+ \to \{x + iy; x \le 0\}$ is the unique solution analytic in $z \in \mathbb{C}^+$ of the equation

$$\rho_z(t) = \int_0^\infty g(y) e^{\frac{iy}{t} z + \rho_z(\frac{y}{t})} dy \tag{11}$$

This entails the convergence of the spectral measure of $\mathbf{A}^N$ as in (5), with σ replaced by a probability measure with Stieltjes transform given by (10). To give some heuristics of the proof of such convergence, let us take $z \in \mathbb{C}\backslash\mathbb{R}$. Then, Schur complement formula reads

$$(z - \mathbf{A})^{-1}_{ii} = \frac{1}{z - A_{ii} - \langle A_i, (z - \mathbf{A}^{(i)})^{-1} A_i\rangle}$$

where $A_i = (A_{ij})_{j\neq i}$ and $\mathbf{A}^{(i)}$ is the $(N-1)\times(N-1)$ matrix obtained from $\mathbf{A}$ by removing the ith row and column. A_{ii} goes to zero with N, but

$$\langle A_i, (z-\mathbf{A}^{(i)})^{-1}A_i\rangle \simeq \sum_{j\neq i} A_{ij}^2 (z-\mathbf{A}^{(i)})^{-1}_{jj}$$

is a non trivial random variable in the heavy tail case. We can compute its Fourier transform thanks to our hypothesis (8), and deduce fractional moments of the resolvent as follows. Observe that for $\beta > 0$, there exists a constant C_β such that for all $z = a + ib, b < 0$

$$\frac{1}{z^\beta} = C_\beta \int_0^\infty dx x^{\beta-1} e^{-ixz}\,.$$

As a consequence, we can guess thanks to (8) that

$$\begin{aligned}\mathbb{E}[((z-\mathbf{A})^{-1}_{ii})^\beta] &\simeq C_\beta \int_0^\infty dx x^{\beta-1} e^{-ixz}\mathbb{E}[e^{-ix\sum_{j\neq i} A_{ij}^2 (z-\mathbf{A}^{(i)})^{-1}_{jj}}]\\ &\simeq C_\beta \int_0^\infty dx x^{\beta-1} e^{-ixz} e^{\frac{1}{N}\sum \Phi(x(z-\mathbf{A}^{(i)})^{-1}_{jj})} \simeq C_\beta \int_0^\infty dx x^{\beta-1} e^{-ixz} e^{\rho_z^N(x)}\end{aligned}$$

with the order parameter $\rho_z^N(x) := \mathbb{E}[\frac{1}{N}\sum \Phi(x(z-\mathbf{A}^{(i)})^{-1}_{jj})]$. Here, we used self-averaging of the order parameter. We next can get an equation for the order parameter thanks to hypothesis (9) which implies, again by Schur complement formula, that

$$\rho_z^N(x) \simeq \int_0^\infty g(y) e^{\frac{iy}{\tau}z + \rho_z^N(\frac{y}{\tau})} dy\,.$$

Hence, if the above heuristics are true, we get convergence of fractional moments of the resolvent as soon as the equation for ρ_z as a unique solution, to which the order parameter ρ_z^N converges. In particular, the Stieltjes transform

$$G_N(z) = \frac{1}{N}\sum_{i=1}^N \frac{1}{z-\lambda_i} = \frac{1}{N}\sum_{i=1}^N (z-\mathbf{A})^{-1}_{ii}$$

converges towards $C_1\int_0^\infty dx x^{\beta-1}e^{-ixz}e^{\rho_z(x)}$. The above arguments were made rigorous in [7–9].

It is quite difficult to study the limiting probability measure whose Stieljes transform is given by the intricate fixed point Eq. (10). It is known that it has unbounded support. The case of α-stable laws is easier: they have a smooth density except possibly at finitely many points [6], their density at the origin can be computed and this law can be interpreted as the spectral measure of the adjacency matrix of the PWIT [11]. But in general, simple properties such as the existence of

an absolutely continuous part are difficult, see e.g. [17] in the case of the Erdös-Rényi graph.

3 Central Limit Theorem for Linear Statistics

One can push the previous arguments to study the fluctuations of the linear statistics. Let us first consider light tails matrices, in fact matrices satisfying (2) and the fluctuations of

$$M_k^N := \sum_{i=1}^N \left(\lambda_i^k - \mathbb{E}[\lambda_i^k]\right) \tag{12}$$

Then, it was shown in [5], see also [24] for the Gaussian case, that M_k^N converges in distribution towards a Gaussian variable whose covariance depends on the fourth moment of the entries. Again, such convergence can be generalized to heavier tails by replacing taking as test functions smooth enough bounded test functions instead of moment: it was indeed shown, see [27, 31], that $\sum_{i=1}^N f(\lambda_i) - \mathbb{E}[\sum_{i=1}^N f(\lambda_i)]$ converges in distribution to a Gaussian variable provided f is $C^{1/2+\varepsilon}$, $\varepsilon > 0$. One can also recenter with respect to $N\int f(x)d\sigma(x)$, see e.g. [9], inducing in general a non trivial mean to the limiting Gaussian variable.

Hence, we see that the spectral measure of light tails matrices fluctuates much less than the empirical measure of independent variables which is of order $1/\sqrt{N}$ and not $1/N$. The situation changes drastically when one considers heavy tails matrices, and in fact as soon as the fourth moment of the entries is infinite. If one considers α stable laws with $\alpha \in (2,4)$, the fluctuations are of order $N^{-1+\alpha/4}$ [10]. For heavy tails matrices satisfying (6) or (8), the fluctuations are of size $1/\sqrt{N}$ [9], as for independent variables. Test functions are assumed to be smooth enough in these cases, and centering in general holds with respect to expectation (additional hypothesis concerning the errors in (6) or (8) are required otherwise). To give some heuristics of the proof of such central limit theorem, let us take $f = (z - .)^{-1}$ for $z \in \mathbb{C}\backslash\mathbb{R}$. Then, recall that Schur complement formula shows that

$$(z - \mathbf{A})^{-1}_{ii} \simeq \frac{1}{z - Y_i^N(z)} \text{ with } Y_i^N(z) = \sum_{j\neq i} A_{ij}^2 (z - \mathbf{A}^{(i)})^{-1}_{jj}\,.$$

The $(Y_i^N(z))_{1\leq i\leq N}$ converge (jointly for finite marginals) towards independent $\alpha/2$-stable laws with parameter ρ_z given by (11). Hence, the diagonal elements of the resolvent behave like independent equidistributed random variables, so that their sum, once renormalized by $\sqrt{N}$, converges towards a Gaussian variable.

4 Local Law

In order to get more local information, one would like to be able to take less smooth functions in the previous result, in fact functions which are supported in intervals going to zero as N goes to infinity. This idea was first developed for light tails matrices by Erdös-Schlein-Yau [22]. It reads as follows. Assume that μ, the law of the entries, have stretched exponential decay, i.e. there exists $\alpha > 0$ and $C < \infty$ such that for all $t \geq 0$

$$\mu(|x| \geq \sqrt{N}^{-1} t) \leq Ce^{-t^\alpha}. \tag{13}$$

Let for $I \subset]-2, 2[$, N_I be the number of eigenvalues in I. Then for all $\kappa \in (0, 2)$,all $\eta > \frac{(\log N)^4}{N}$ sufficiently small, there exists $c > 0$ such that we have for all $\delta \leq c\kappa$,

$$\mathbb{P}\left(\sup_{|E|\leq 2-\kappa} \left| \frac{N_{[E-\eta,E+\eta]}}{2N\eta} - \rho_{sc}(E) \right| > \frac{(\log N)^c}{\sqrt{\eta}N} |I| \right) \leq N^{-\log\log N}. \tag{14}$$

Such estimates were shown to hold under much weaker hypothesis afterwards and it was extended to the neighborhood of $\{-2\}$ and $\{2\}$, the boundary of the support of the semi-circle, see e.g. [19, 20, 26] or [32]. This allowed to prove that the eigenvalues are rigid, that is do not fluctuate much around their deterministic limit. Indeed, if we now order the eigenvalues $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_N$ and let γ_i^N be the ith quantile given by $\sigma([-2, \gamma_i^N]) = i/N$, then, with probability greater than $1 - N^{-N}$, for all i

$$|\lambda_i - \gamma_i^N| \leq (\log N)^2 N^{-2/3} \min\{(N-i)^{-1/3}, i^{-1/3}\}.$$

Of course, one can not expect the eigenvalues to be as rigid in the heavy tails case since this would contradict the central limit theorems of the previous section. However, one could still expect the local law to be true inside the bulk: in [9], corresponding to entries decaying like $x^{-\alpha}$ for some $\alpha \in (2, 4)$, it was shown that global fluctuations hold in the scale $N^{-\alpha/4}$ whereas local law inside the bulk was derived in [1]. Hence, in this case, large eigenvalues should be less rigid, creating large fluctuations. For heavier tails, local laws have not yet been established except for the case of α-stable entries [12]. The following result was proved if the A_{ij} are α-stable variables: for all $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(itA_{11})] = \exp(-\frac{1}{N} w_\alpha |t|^\alpha), \tag{15}$$

for some $0 < \alpha < 2$ and $w_\alpha = \pi/(\sin(\pi\alpha/2)\Gamma(\alpha))$. We let μ_α be the equilibrium measure and put

$$\rho = \begin{cases} \frac{1}{2} & \text{if } \frac{8}{5} \le \alpha < 2 \\ \frac{\alpha}{8-3\alpha} & \text{if } 1 < \alpha < \frac{8}{5} \\ \frac{\alpha}{2+3\alpha} & \text{if } 0 < \alpha \le 1. \end{cases} \tag{16}$$

Then, there exists a finite set $\mathcal{E}_\alpha \subset \mathbb{R}$ such that if $K \subset \mathbb{R}\backslash\mathcal{E}_\alpha$ is a compact set and $\delta > 0$, the following holds. There are constants $c_0, c_1 > 0$ such that for all integers $N \ge 1$, if $I \subset K$ is an interval of length $|I| \ge c_1 N^{-\rho}(\log N)^2$, then

$$\left|N_I - N\mu_\alpha(I)\right| \le \delta N|I|, \tag{17}$$

with probability at least $1 - 2\exp\left(-c_0 N\delta^2|I|^2\right)$.

In both light and heavy tails, the main point is to estimate the Stieltjes transform $G_N(z) = \frac{1}{N}\sum_{i=1}^N (z-\lambda_i)^{-1}$ for z going to the real axis: $z = E + i\eta$ with η of order nearly as good as N^{-1} for light tails, $N^{-\rho}$ for heavy tails. This is done by showing that G_N is characterized approximately by a closed set of equations. In the case of lights tails, one has simply a quadratic equation for G_N and needs to show that the error terms remain small as z approaches the real line. In the heavy tails case, the equations are much more complicated, see (10), and therefore more difficult to handle. Even in the α-stable case it is not clear what should be the optimal local law. We believe ρ should be at least equal to $1/2$ for all $\alpha \in (1, 2)$. Similar questions are completely open for other heavy tails matrices.

5 Localization and Delocalization of the Eigenvectors

Based on the local law, it was shown that the eigenvectors of Wigner matrices with light entries (for instance with sub exponential tail) are strongly delocalized [22, 23], for any $p \in (2, \infty]$ and $\epsilon > 0$, with high probability,

$$\max_{1\le k\le N} \|u_k\|_p = O(N^{1/p-1/2+\epsilon}), \tag{18}$$

where for $u \in \mathbb{R}^n$, $\|u\|_p = \left(\sum_{i=1}^n |u_i|^p\right)^{1/p}$ and $\|u\|_\infty = \max|u_i|$. This phenomenon seems to be quite robust and continues to hold even if a fraction of the entries vanish. For instance, if the entries vanish outside a band around the diagonal of width W, it is conjectured that the eigenvectors remain delocalized as long as $W \gg \sqrt{N}$, but start being localized when $W \ll \sqrt{N}$. Universality was shown for $W \simeq N$, see [15].

It was shown in [12] that the eigenvectors of matrices with α-stable entries are also delocalized if $\alpha \in (1,2)$: there is a finite set $\mathcal{E}_\alpha$ such if $K \subset \mathbb{R}\backslash\mathcal{E}_\alpha$ is a compact set, for any $\varepsilon > 0$, with high probability,

$$\max\left\{\|u_k\|_\infty : 1 \leq k \leq N, \lambda_k \in K\right\} = O(N^{-\delta+\epsilon}), \tag{19}$$

where $\delta = (\alpha-1)/((2\alpha) \vee (8-3\alpha))$. Since $\|u\|_p \leq \|u\|_\infty^{1-2/p}\|u\|_2^{2/p}$, it implies that the L^p-norm of the eigenvectors is $O(N^{2\delta/p-\delta+o(1)})$. Notice that when $\alpha \to 2$, then $\delta \to 1/4$ and it does not match with (18): we expect that this result could be improved.

However, for $\alpha \in (0,1)$, we observe a new phenomenon, closer to what can be observed for random Schrodinger operators, see e.g. [3]: eigenvectors are delocalized if they correspond to eigenvalues which are small, but are localized if they correspond to large eigenvalues. In [14], Bouchaud and Cizeau conjectured the existence of a mobility edge, $E_\alpha > 0$ where this transition occurs (a value for E_α is predicted in [34]). However, the sense of localization/delocalization has to be precised. In [12, 13], we considered

$$P_I(k) = \frac{1}{|\Lambda_I|}\sum_{u\in\Lambda_I}\langle u, e_k\rangle^2, Q_I = N\sum_{k=1}^{N} P_I(k)^2 \in [1,N].$$

and showed that for $\alpha \in (0,2/3)$, for $I = [E-\eta, E+\eta]$ with η going to zero with N, Q_I goes to infinity if E is large enough, whereas it is bounded for small enough E. This localization/delocalization of the eigenvectors should be related with the local fluctuations of the spectrum. Bouchaud and Cizeau conjectured that the small eigenvalues should behave like the eigenvalues of the Gaussian ensemble when $\alpha \in (1,2)$. Also, for $\alpha \in (0,1)$ and large eigenvalues, one expects a Poisson distribution. However, for the two remaining regimes, they predict something between Poisson and Sine-kernel. In [34], the authors predict a phase transition at a mobility edge between the localized and delocalized regimes. While this article was under print, it was shown in [2] that in the regime of delocalization, the local statistics are given by the GOE statistics and that, then, the eigenvectors are completely delocalized in the sense that (19) holds with the optimal rate $\delta = 1/2$.

6 Heavy-Tailed Operators in Free Probability

Another important feature of random matrices is their role in free probability, as a toy example of matrices whose large dimension limit are free. Free probability is a theory of non-commutative variables equipped with a notion of freeness. Let us consider self-adjoint non-commutative variables $X_1, \ldots, X_d$. We equip the set of

polynomials in these non-commutative variables with the involution

$$(zX_{i_1}X_{i_2}\cdots X_{i_k})^* = \bar{z}X_{i_k}\cdots X_{i_1}\,.$$

Distributions of d self adjoint variables are simply linear functions τ on this set of polynomials in non-commutative variables such that

$$\tau(PP^*) \geq 0, \quad \tau(1) = 1, \qquad \tau(PQ) = \tau(QP)\,,$$

for all choices of polynomials P, Q. Freeness is a condition on the joint distribution of non-commutative variables. For instance, we say that $X_1, \ldots, X_d$ are free under τ iff

$$\tau(P_1(X_{i_1})\cdots P_\ell(X_{i_\ell})) = 0 \tag{20}$$

as soon as $\tau(P_j(X_j)) = 0$ for all j and $i_j \neq i_{j+1}$, $1 \leq j \leq \ell - 1$. The latter property was introduced by Voiculescu and named freeness, as it is related with the usual notion of free generators of a group. Taking d independent Wigner matrices $\mathbf{X}_1^N, \ldots, \mathbf{X}_d^N$ with light tails, one finds that for all choices of $i_1, \ldots, i_k \in \{1, \ldots, d\}^k$,

$$\lim_{N\to\infty} \frac{1}{N}\,\mathrm{Tr}(\mathbf{X}_{i_1}^N \cdots \mathbf{X}_{i_k}^N) = \sigma^d(X_{i_1}\cdots X_{i_k}) \qquad a.s$$

where σ^d is uniquely described by saying that the moments of a single X_i are given by the Catalan numbers, and their joint moments satisfy (20). Voiculescu also showed that matrices $\mathbf{Y}_j = U_j D_j U_j^*$ with deterministic matrices D_j and independent Haar distributed orthogonal matrices satisfy at the large N limit the freeness property (20). Hence, matrices become asymptotically free if the position of their eigenvectors are "sufficiently" independent. One could then wonder whether Wigner matrices with heavy tails are also asymptotically free. All these matrices share the invariance by multiplication by permutation matrices. It is clear that matrices conjugated by independent permutation matrices are not asymptotically free. Indeed, for instance if one takes two diagonal matrices with given spectral measure, it will have a different joint law if it is conjugated by unitary matrices than if it is conjugated by permutations (which does not change the law) since then they will commute. Similarly, it is not enough to know the spectral measure of a heavy tail matrix to derive the joint law of several of them. In the one matrix case, this could already be guessed in view of the additional parameter ρ_z. In fact, this parameter appears naturally as the large N limit of $\frac{1}{N}\sum_{k=1}^N \Phi(t(z - \mathbf{A})_{ii}^{-1})$ which is not a function of the spectral measure. To remedy this point, another non-commutative framework was introduced by C. Male: the distribution of traffics and their free product [28]. Traffics distributions are now linear maps of a set of functionals that generalize the non-commutative polynomials, called graph polynomials. Namely, if we are given d $N \times N$ self-adjoint random matrices

$(A_1^N, \ldots, A_d^N)$, a finite connected graph $G = (V, E)$ and γ a map from E into $\{1, \ldots, d\}$, we define the observables given by

$$\Phi_{A^N}(G, \gamma) = \mathbb{E}\left[\frac{1}{N} \sum_{\phi: V \mapsto \{1,\ldots,N\}} \prod_{e=(v,w)\in E} A_{\gamma(e)}(\phi(w), \phi(v))\right],$$

where the sum is taken over all maps and $N \geq |V|$. For instance if G is a cycle $V = \{v_1, \ldots, v_k\}$, $E = \{e_\ell = (v_\ell, v_{\ell+1})\}_{1\leq \ell \leq k}$ with $v_{k+1} = v_1$, we get the trace of the word $A_{\gamma(e_1)} \cdots A_{\gamma(e_k)}$. If V is as before, but $E = \{e_\ell = (v_\ell, v_{\ell+1})\}_{1\leq \ell \leq k} \cup \{e_{\ell+k} = (v_\ell, v_{\ell+1})\}_{1\leq \ell \leq k}$ while $\gamma(e_\ell) = 1$ for $\ell \leq k$ and 2 for $\ell \geq k+1$, we get the trace of the kth moment of the Hadamard product $A_1^N \circ A_2^N$. More generally, we can obtain all the the normalized trace of Hadamard products of polynomials in the matrices $A_1^N, \ldots, A_d^N$

$$\mathbb{E}\left[\frac{1}{N} \mathrm{Tr}(P_1(A_1^N, \ldots, A_d^N) \circ \cdots \circ P_k(A_1^N, \ldots, A_d^N))\right].$$

The collection of all $\Phi_{A^N}(G, \gamma)$ defines the distribution of the traffics $(A_1^N, \ldots, A_d^N)$. A sequence $(A_1^N, \ldots, A_d^N)$ of matrices converges in traffics iff $\Phi_{A^N}(G, \gamma)$ converges for all finite connected graphs G and all map γ. The model of heavy Wigner matrices was the initial motivation to introduce it: matrices satisfying (6) can be seen to converge in traffic. Traffic distribution comes together with the notion of traffic independence, which is more complicated than freeness in the sense that it involves non algebraic (combinatorial) formulas (see [28, Definition 3.10]). However, it prescribes uniquely the traffic distribution of two families $\mathbf{A}$ and $\mathbf{B}$ from the traffic distributions of $\mathbf{A}$ and $\mathbf{B}$. One can see that traffic independence does not imply free independence. Let us consider two asymptotically traffic independent families of matrices $\mathbf{A}_N$ and $\mathbf{B}_N$ (that is with traffic distribution which converges towards a distribution of two traffic independent families). If

$$\kappa(A_N, P) = \frac{1}{N} \mathrm{Tr}\left[P(\mathbf{A}_N) \circ P^*(\mathbf{A}_N)\right] - |\frac{1}{N} \mathrm{Tr}\, P(\mathbf{A}_N)|^2$$

does not go to zero for some polynomial P and the same hold for $\mathbf{B}_N$, then $\mathbf{A}_N$ and $\mathbf{B}_N$ are not asymptotically freely independent [28, Section 3.3]. This criterion applies for heavy Wigner matrices, which shows in particular that heavy Wigner matrices are not asymptotically freely independent, and not asymptotically freely independent with diagonal matrices. On the contrary, if $\kappa(A_N, P)$ and $\kappa(B_N, P)$ tend to zero for all polynomial P, then A_N and B_N are asymptotically free independent [16, Section 3.2]. This is the case of adjacency matrices of Erdos-Renyi matrices with parameter $\frac{c_N}{N}$ when c_N goes to infinity [29]. Traffic independence is difficult to manipulate, still we can deduce from it a system of equations which characterizes the limiting distribution of independent heavy Wignerand deterministic diagonal matrices. It involves again limits of normalized trace of

Hadamard products of polynomials in matrices. It implies another characterization of the spectrum of a single heavy Wigner matrix in term of the maps $G(\lambda)^k = \frac{1}{N}\sum_i \left[(\lambda - X)^{-1}_{ii}\right]^k$ [29].

References

1. Aggarwal, A.: Bulk universality for generalized wigner matrices with few moments. arXiv:1612.00421 (2016)
2. Aggarwal, A., Lopatto, P., Yau, H.-T.: GOE statistics for Levy matrices. https://arxiv.org/abs/1806.07363 (2018)
3. Aizenman, M., Warzel, S.: Disorder-induced delocalization on tree graphs. In: Exner, P. (ed.) Mathematical Results in Quantum Physics, pp. 107–109. World Scientific Publication, Hackensack (2011). MR 2885163
4. Anderson, G.W., Guionnet, A., Zeitouni, O.: An Introduction to Random Matrices. Cambridge Studies in Advanced Mathematics, vol. 118. Cambridge University Press, Cambridge (2010). MR 2760897
5. Anderson, G.W., Zeitouni, O.: A CLT for a band matrix model. Probab. Theory Rel. Fields **134**, 283–338 (2005). MR 2222385
6. Belinschi, S., Dembo, A., Guionnet, A.: Spectral measure of heavy tailed band and covariance random matrices. Commun. Math. Phys. **289**(3), 1023–1055 (2009)
7. Ben Arous, G., Guionnet, A.: The spectrum of heavy tailed random matrices. Commun. Math. Phys. **278**(3), 715–751 (2008)
8. Benaych-Georges, F., Guionnet, A.: Central limit theorem for eigenvectors of heavy tailed matrices. Electron. J. Probab. **19**(54), 27 (2014). MR 3227063
9. Benaych-Georges, F., Guionnet, A., Male, C.: Central limit theorems for linear statistics of heavy tailed random matrices. Commun. Math. Phys. **329**(2), 641–686 (2014). MR 3210147
10. Benaych-Georges, F., Maltsev, A.: Fluctuations of linear statistics of half-heavy-tailed random matrices. Stoch. Process. Appl. **126**(11), 3331–3352 (2016). MR 3549710
11. Bordenave, C., Caputo, P., Chafaï, D.: Spectrum of large random reversible Markov chains: heavy-tailed weights on the complete graph. Ann. Probab. **39**(4), 1544–1590 (2011)
12. Bordenave, C., Guionnet, A.: Localization and delocalization of eigenvectors for heavy-tailed random matrices. Probab. Theory Relat. Fields **157**(3–4), 885–953 (2013). MR 3129806
13. Bordenave, C., Guionnet, A.: Delocalization at small energy for heavy-tailed random matrices. Commun. Math. Phys. **354**(1), 115–159 (2017). MR 3656514
14. Bouchaud, J.-P., Cizeau, P.: Theory of Lévy matrices. Phys. Rev. E **3**, 1810–1822 (1994)
15. Bourgade, P., Erdos, L., Yau, H.-T., Yin, J.: Universality for a class of random band matrices. Adv. Theor. Math. Phys. **21**(3), 739–800 (2017). MR 3695802
16. Cébron, G., Dahlqvist, A., Male, C: Universal constructions for space of traffics (2016). arXiv:1601.00168
17. Bordenave, C., Sen, A., Virág, B.: Mean quantum percolation. J. Eur. Math. Soc. (JEMS) **19**(12), 3679–3707 (2017). MR 3730511
18. Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. Int. Math. Res. Not. (IMRN) (3), 436–479 (2010). MR 2587574
19. Erdős, L.: Universality of Wigner random matrices: a survey of recent results. Uspekhi Mat. Nauk **66**(3)(399), 67–198 (2011). MR 2859190
20. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. Electron. J. Probab. **18**(59), 58 (2013). MR 3068390
21. Erdős, L., Péché, S., Ramírez, J.A., Schlein, B., Yau, H.-T.: Bulk universality for Wigner matrices. Commun. Pure Appl. Math. **63**(7), 895–925 (2010). MR 2662426

22. Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. Commun. Math. Phys. **287**(2), 641–655 (2009)
23. Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. Ann. Probab. **37**(3), 815–852 (2009)
24. Johansson, K.: On fluctuations of eigenvalues of random Hermitian matrices. Duke Math. J. **91**, 151–204 (1998)
25. Johansson, K.: Universality for certain Hermitian Wigner matrices under weak moment conditions. Ann. Inst. Henri Poincaré Probab. Stat. **48**(1), 47–79 (2012). MR 2919198
26. Lee, J.O., Yin, J.: A necessary and sufficient condition for edge universality of Wigner matrices. Duke Math. J. **163**(1), 117–173 (2014). MR 3161313
27. Lytova, A., Pastur, L.: Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. Ann. Probab. **37**(5), 1778–1840 (2009). MR 2561434
28. Male, C.: Traffics distributions and independence: the permutation invariant matrices and the notions of independence. arXiv:1111.4662 (2011)
29. Male, C.: The limiting distributions of large heavy Wigner and arbitrary random matrices. J. Funct. Anal. **272**(1), 1–46 (2017). MR 3567500
30. Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. Commun. Math. Phys. **207**(3), 697–733 (1999). MR 1727234
31. Sosoe, P., Wong, P.: Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices. Adv. Math. **249**, 37–87 (2013). MR 3116567
32. Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics up to the edge. Commun. Math. Phys. **298**(2), 549–572 (2010)
33. Tao, T., Vu, V.: The Wigner-Dyson-Mehta bulk universality conjecture for Wigner matrices. Electron. J. Probab. **16**(77), 2104–2121 (2011). MR 2851058
34. Tarquini, E., Biroli, G., Tarzia, M.: Level statistics and localization transitions of levy matrices. Phys. Rev. Lett. **116**, 010601 (2015)
35. Wigner, E.P.: On the distribution of the roots of certain symmetric matrices. Ann. Math. **67**, 325–327 (1958)
36. Wishart, J.: The generalized product moment distribution in samples from a normal multivariate population. Biometrika **20A**, 32–52 (1928)
37. Zakharevich, I.: A generalization of Wigner's law. Commun. Math. Phys. **268**, 403–414 (2006)

An Analyst's Take on the BPHZ Theorem

Martin Hairer

Abstract We provide a self-contained formulation of the BPHZ theorem in the Euclidean context, which yields a systematic procedure to "renormalise" otherwise divergent integrals appearing in generalised convolutions of functions with a singularity of prescribed order at their origin. We hope that the formulation given in this article will appeal to an analytically minded audience and that it will help to clarify to what extent such renormalisations are arbitrary (or not). In particular, we do not assume any background whatsoever in quantum field theory and we stay away from any discussion of the physical context in which such problems typically arise.

1 Introduction

The BPHZ renormalisation procedure named after Bogoliubov, Parasiuk, Hepp and Zimmerman [1, 17, 21] (but see also the foundational results by Dyson and Salam [8, 9, 18, 19]) provides a consistent way to renormalise probability amplitudes associated to Feynman diagrams in perturbative quantum field theory (pQFT). The main aim of this article is to provide an analytical result, Theorem 3.1 below, which is a general form of the "BPHZ theorem" in the Euclidean context. To a large extent, this theorem has been part of the folklore of mathematical physics since the publication of the abovementioned works (see for example the article [11] which gives rather sharp analytical bounds and is close in formulation to our statement, as well as the series of articles [4–6] which elucidate some of the algebraic aspects of the theory, but focus on dimensional regularisation which is not available in the general context considered here), but it seems difficult to find precise analytical

M. Hairer (✉)
Mathematics Institute, Imperial College, London, UK
e-mail: m.hairer@imperial.ac.uk

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_16

statements in the literature that go beyond the specific context of pQFT. One reason seems to be that, in the context of the perturbative expansions arising in pQFT, there are three related problems. The first is to control the small-scale behaviour of the integrands appearing in Feynman diagrams (the "ultraviolet behaviour"), the second is to control their large scale ("infrared") behaviour, and the final problem is to show that the renormalisation required to deal with the first problem can be implemented by modifying (in a scale-dependent way) the finitely many coupling constants appearing in the Lagrangian of the theory at hand, so that one still has a physical theory.

The approach we take in the present article is purely analytic and completely unrelated to any physical theory, so we do not worry about the potential physical interpretation of the renormalisation procedure. We do however show in Sect. 3.3 that it has a number of very nice mathematical properties so that the renormalised integrals inherit many natural properties from their unrenormalised counterparts. We also completely discard the infrared problem by assuming that all the kernels ("propagators") under consideration are compactly supported. For the reader who might worry that this could render our main result all but useless, we give a simple separate argument showing how kernels with algebraic decay at infinity can be dealt with as well. Note also, that contrary to much of the related theoretical and mathematical physics literature, all of our arguments take place in configuration space, rather than in Fourier space. In particular, the analysis presented in this article shares similarities with a number of previous works, see for example [7, 10, 11] and references therein.

The approach taken here is informed by some results recently obtained in the context of the analysis of rough stochastic PDEs in [2, 3, 15]. Indeed, the algebraic structure described in Sects. 2.3 and 2.4 below is very similar to the one described in [2, 15], with the exception that there is no "positive renormalisation" in the present context. In this sense, this article can be seen as a perhaps gentler introduction to these results, with the content of Sect. 2 roughly parallel to [2], while the content of Sect. 3 is rather close to that of [3]. In particular, Sect. 2 is rather algebraic in nature and allows to conceptualise the structure of the counterterms appearing in the renormalisation procedure, while Sect. 3 is rather analytical in nature and contains the multiscale analysis underpinning our main continuity result, Theorem 3.1. Finally, in Sect. 4, we deal with kernels exhibiting only algebraic decay at infinity. While the conditions given in this section are sharp in the absence of any cancellations in the large-scale behaviour, we do not introduce an analogue of the "positive renormalisation" of [3], so that the argument remains relatively concise.

2 An Analytical Form of the BPHZ Theorem

Fix a countable set $\mathfrak{L}$ of labels, a map $\deg\colon \mathfrak{L} \to \mathbf{R}$, and an integer dimension $d > 0$. We assume that the set of labels has a distinguished element which we denote by $\delta \in \mathfrak{L}$ satisfying $\deg \delta = -d$ and that, for every multiindex k, there is an injective map $\mathfrak{t} \mapsto \mathfrak{t}^{(k)}$ on $\mathfrak{L}$ with $\mathfrak{t}^{(0)} = \mathfrak{t}$ and such that

$$(\mathfrak{t}^{(k)})^{(\ell)} = \mathfrak{t}^{(k+\ell)}\,, \qquad \deg \mathfrak{t}^{(k)} = \deg \mathfrak{t} - |k|\,. \tag{1}$$

We also set $\mathfrak{L}_\star = \mathfrak{L} \setminus \{\delta^{(k)} : k \in \mathbf{N}^d\}$ and we assume that there is a finite set $\mathfrak{L}_0 \subset \mathfrak{L}$ such that every element of $\mathfrak{L}$ is of the form $\mathfrak{t}^{(k)}$ for some $k \in \mathbf{N}^d$ and some $\mathfrak{t} \in \mathfrak{L}_0$. We then give the following definition.

Definition 2.1 A *Feynman diagram* is a finite directed graph $\Gamma = (\mathcal{V}, \mathcal{E})$ endowed with the following additional data:

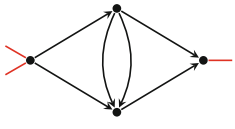
- An ordered set of distinct vertices $\mathcal{L} = \{[1], \dots, [k]\} \subset \mathcal{V}$ such that each vertex in $\mathcal{L}$ has exactly one outgoing edge (called a "leg") and no incoming edge, and such that each connected component of Γ contains at least one leg. We will frequently use the notation $\mathcal{V}_\star = \mathcal{V} \setminus \mathcal{L}$, as well as $\mathcal{E}_\star \subset \mathcal{E}$ for the edges that are not legs.
- A decoration $\mathfrak{t}\colon \mathcal{E} \to \mathfrak{L}$ of the edges of Γ such that $\mathfrak{t}(e) \in \mathfrak{L}_\star$ if and only if $e \in \mathcal{E}_\star$.

We will always use the convention of [16] that e_- and e_+ are the source and target of an edge e, so that $e = (e_- \to e_+)$. We also label legs in the same way as the corresponding element in $\mathcal{L}$, i.e. we call the unique edge incident to the vertex $[j]$ the jth leg of Γ. The way we usually think of Feynman diagrams is as labelled graphs $(\mathcal{V}_\star, \mathcal{E}_\star)$ with a number of legs attached to them, where the legs are ordered and each leg is assigned a d-dimensional multiindex. An example of Feynman diagram with three legs is shown in Fig. 1, with legs drawn in red and decorations suppressed. We do not draw the arrows on the legs since they are always incoming by definition. In this example, $|\mathcal{V}| = 7$ and $|\mathcal{V}_\star| = 4$.

Write now $\mathbf{S} = \mathbf{R}^d$, and assume that we are given a kernel $K_\mathfrak{t}\colon \mathbf{S} \to \mathbf{S}$ for every $\mathfrak{t} \in \mathfrak{L}_\star$, such that $K_\mathfrak{t}$ exhibits a behaviour of order $\deg \mathfrak{t}$ at the origin but is smooth otherwise. For simplicity, we also assume that these kernels are all compactly supported, say in the unit ball. More precisely, we assume that for every $\mathfrak{t} \in \mathfrak{L}_\star$ and every d-dimensional multiindex k there exists a constant C such that one has the bound

$$|D^k K_\mathfrak{t}(x)| \le C|x|^{\deg \mathfrak{t} - |k|}\mathbf{1}_{|x|\le 1}\,, \qquad \forall x \in \mathbf{S}\,. \tag{2}$$

Fig. 1 A Feynman diagram



We also extend K to all of $\mathfrak{L}$ by using the convention that $K_\delta = \delta$, a Dirac mass at the origin, and we impose that for every multiindex k and label $\mathfrak{t} \in \mathfrak{L}$, one has

$$K_{\mathfrak{t}^{(k)}} = D^k K_{\mathfrak{t}} \,. \tag{3}$$

Note that (2) is compatible with (1) so that non-trivial (i.e. not just vanishing or smooth near the origin) kernel assignments do actually exist. To some extent it is also compatible with the convention $K_\delta = \delta$ and $\deg \delta = -d$ since the "delta function" on $\mathbf{R}^d$ is obtained as a distributional limit of functions satisfying a uniform bound of the type (2) with $\deg \mathfrak{t} = -d$. Given all this data, we would now like to associate to each Feynman diagram Γ with k legs a distribution $\Pi\Gamma$ on $\mathbf{S}^k$ by setting

$$\big(\Pi\Gamma\big)(\varphi) = \int_{\mathbf{S}^{\mathcal{V}}} \prod_{e \in \mathcal{E}} K_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \varphi(x_{[1]}, \dots, x_{[k]})\, dx \,. \tag{4}$$

Note that of course $\Pi\Gamma$ does not just depend on the combinatorial data $\Gamma = (\mathcal{V}, \mathcal{E}, \mathcal{L}, \mathfrak{t})$, but also on the analytical data $(K_{\mathfrak{t}})_{\mathfrak{t} \in \mathfrak{L}_\star}$. We sometimes suppress the latter dependency on our notation in order to keep it light, but it will be very useful later on to also allow ourselves to vary the kernels $K_{\mathfrak{t}}$. We call the map Π a "valuation".

The problem is that on the face of it, the definition (4) does not always make sense. The presence of the (derivatives of) delta functions is not a problem: writing $v_i \in \mathcal{V}_\star$ for the unique vertex such that $([i] \to v_i) \in \mathcal{E}$ and ℓ_i for the multiindex such that the label of this leg is $\delta^{(\ell_i)}$, we can rewrite (4) as

$$\big(\Pi\Gamma\big)(\varphi) = \int_{\mathbf{S}^{\mathcal{V}_\star}} \prod_{e \in \mathcal{E}_\star} K_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \big(D_1^{\ell_1} \cdots D_k^{\ell_k} \varphi\big)(x_{v_1}, \dots, x_{v_k})\, dx \,. \tag{5}$$

The problem instead is the possible lack of integrability of the integrand appearing in (5). For example, the simplest nontrivial Feynman diagram with two legs is given by $\Gamma = \overset{0}{\bullet} \xrightarrow{\mathfrak{t}} \overset{0}{\bullet}$ which, by (5), should be associated to the distribution

$$\big(\Pi\Gamma\big)(\varphi) = \int_{\mathbf{S}^2} K_{\mathfrak{t}}(y_1 - y_0) \varphi(y_0, y_1)\, dy \,. \tag{6}$$

If it happens that $\deg \mathfrak{t} < -d$, then $K_{\mathfrak{t}}$ is non-integrable in general, so that this integral may not converge. It is then natural to modify our definition, but "as little as possible". In this case, we note that if the test function φ happens to vanish near the diagonal $y_1 = y_0$, then the singularity of $K_{\mathfrak{t}}$ does not matter and (6) makes perfect sense. We would therefore like to find a distribution $\Pi\Gamma$ which agrees with (6) on such test functions but still yields finite values for *every* test function φ. One way of achieving this is to set

$$\big(\Pi\Gamma\big)(\varphi) = \int_{\mathbf{S}^2} K_{\mathfrak{t}}(y_1 - y_0) \Big(\varphi(y_0, y_1) - \sum_{|k| + \deg \mathfrak{t} \le -d} \frac{(y_1 - y_0)^k}{k!} D_2^k \varphi(y_0, y_0)\Big)\, dy \,. \tag{7}$$

At first glance, this doesn't look very canonical since it seems that the variables y_0 and y_1 no longer play a symmetric role in this expression. However, it is an easy exercise to see that the *same* distribution can alternatively also be written as

$$\big(\Pi\Gamma\big)(\varphi) = \int_{\mathbf{S}^2} K_{\mathfrak{t}}(y_1 - y_0)\Big(\varphi(y_0, y_1) - \sum_{|k|+\deg \mathfrak{t} \le -d} \frac{(y_0 - y_1)^k}{k!} D_1^k \varphi(y_1, y_1)\Big)\, dy\ .$$

The BPHZ theorem is a far-reaching generalisation of this construction. To formalise what we mean by this, write $\mathcal{K}_\infty^-$ for the space of all smooth kernel assignments as above (compactly supported in the unit ball and satisfying (3)). When endowed with the system of seminorms given by the minimal constants in (2), its completion $\mathcal{K}_0^-$ is a Fréchet space.

With these notations, a "renormalisation procedure" is a map $K \mapsto \Pi^K$ turning a kernel assignment $K \in \mathcal{K}_0^-$ into a valuation Π^K. The purpose of the BPHZ theorem is to argue that the following question can be answered positively.

Main question: Is there a consistent renormalisation procedure such that, for every Feynman diagram, $\Pi\Gamma$ can be interpreted as a "renormalised version" of (4)?

As stated, this is a very loose question since we have not specified what we mean by a "consistent" renormalisation procedure and what properties we would like a valuation to have in order to be a candidate for an interpretation of (4). One important property we would like a good renormalisation procedure to have is the continuity of the map $K \mapsto \Pi^K$. In this way, we can always reason on smooth kernel assignments $K \in \mathcal{K}_\infty^-$ and then "only" need to show that the procedure under consideration extends continuously to all of $\mathcal{K}_0^-$. Furthermore, we would like Π^K to inherit as many properties as possible from its interpretation as the formal expression (4). Of course, as already seen, the "naïve" renormalisation procedure given by (4) itself does *not* have the required continuity property, so we will have to modify it.

2.1 Consistent Renormalisation Procedures

The aim of this section is to collect and formalise a number of properties of (4) which then allows us to formulate precisely what we mean by a "consistent" renormalisation procedure. Let us write $\mathcal{T}$ for the free (real) vector space generated by all Feynman diagrams. This space comes with a natural grading and we write $\mathcal{T}_k \subset \mathcal{T}$ for the subspace generated by diagrams with k legs. Note that $\mathcal{T}_0 \approx \mathbf{R}$ since there is exactly one Feynman diagram with 0 legs, which is the empty one.

Write $\mathcal{S}_k$ for the space of all distributions on $\mathbf{S}^k$ that are translation invariant in the sense that, for $\eta \in \mathcal{S}_k$, $h \in \mathbf{S}$, and any test function φ, one has $\eta(\varphi) = \eta(\varphi \circ \tau_h)$ where $\tau_h(y_1, \dots, y_k) = (y_1 + h, \dots, y_k + h)$. We will write $\mathcal{S}_k^{(c)} \subset \mathcal{S}_k$ for the subset of "compactly supported" distributions in the sense that there exists a compact set $\mathfrak{K} \subset \mathbf{S}^k/\mathbf{S}$ such that $\eta(\varphi) = 0$ as soon as $\operatorname{supp}\varphi \cap \mathfrak{K} = \emptyset$.

Remark 2.2 Compactly supported distributions can be tested against any smooth function φ with the property that for any $x \in \mathbf{S}^k$, the set $\{h \in \mathbf{S} : \varphi(\tau_h(x)) \neq 0\}$ is compact.

Note that $\mathcal{S}_1 \approx \mathbf{R}$ since translation invariant distributions in one variable are naturally identified with constant functions. We will use the convention $\mathcal{S}_0 \approx \mathbf{R}$ by identifying "functions in 0 variables" with $\mathbf{R}$. We also set $\mathcal{S} = \bigoplus_{k\geq 0} \mathcal{S}_k$, so that a valuation Π can be viewed as a linear map $\Pi : \mathcal{T} \to \mathcal{S}$ which respects the respective graduations of these spaces.

Note that the symmetric group $\mathfrak{S}_k$ in k elements acts naturally on $\mathcal{T}_k$ by simply permuting the order of the legs. Similarly, $\mathfrak{S}_k$ acts on $\mathcal{S}_k$ by permuting the arguments of the test functions. Given two Feynman diagrams $\Gamma_1 \in \mathcal{T}_k$ and $\Gamma_2 \in \mathcal{T}_\ell$, we then write $\Gamma_1 \bullet \Gamma_2 \in \mathcal{T}_{k+\ell}$ for the Feynman diagram given by the disjoint union of Γ_1 and Γ_2. Here, we renumber the ℓ legs of Γ_2 in an order-preserving way from $k+1$ to $k+\ell$, so that although $\Gamma_1 \bullet \Gamma_2 \neq \Gamma_2 \bullet \Gamma_1$ in general, one has $\Gamma_1 \bullet \Gamma_2 = \sigma_{k,\ell}(\Gamma_2 \bullet \Gamma_1)$, where $\sigma_{k,\ell} \in \mathfrak{S}_{k+\ell}$ is the permutation that swaps $(1, \dots, \ell)$ and $(\ell+1, \dots, \ell+k)$. Given distributions $\eta_1 \in \mathcal{S}_k$ and $\eta_2 \in \mathcal{S}_\ell$, we write $\eta_1 \bullet \eta_2 \in \mathcal{S}_{k+\ell}$ for the distribution such that

$$\big(\eta_1 \bullet \eta_2\big)(\varphi_1 \otimes \varphi_2) = \eta_1(\varphi_1)\eta_2(\varphi_2) \; .$$

Similarly to above, one has $\eta_1 \bullet \eta_2 = \sigma_{k,\ell}(\eta_2 \bullet \eta_1)$. We extend $\bullet$ by linearity to all of $\mathcal{T}$ and $\mathcal{S}$ respectively, thus turning these spaces into (non-commutative) algebras. This allows us to formulate the first property we would like to retain.

Property 1 *A consistent renormalisation procedure should produce valuations* Π *that are graded algebra morphisms from* $\mathcal{T}$ *to* $\mathcal{S}$ *and such that, for every Feynman diagram* Γ *with* k *legs and every* $\sigma \in \mathfrak{S}_k$, *one has* $\Pi\sigma(\Gamma) = \sigma(\Pi\Gamma)$. *Furthermore* $\Pi\Gamma \in \mathcal{S}_k^{(c)}$ *if* Γ *is connected with* k *legs.*

Similarly, consider a Feynman diagram Γ with $k \geq 2$ legs such that the label of the kth leg is δ and such that the connected component of Γ containing $[k]$ contains at least one other leg. Let $\mathrm{Del}_k \Gamma$ be the Feynman diagram identical to Γ, but with the kth leg removed. If the label of the kth leg is $\delta^{(m)}$ with $m \neq 0$, we set $\mathrm{Del}_k \Gamma = 0$. If we write ι_k for the natural injection of smooth functions on $\mathbf{S}^{k-1}$ to functions on $\mathbf{S}^k$ given by $(\iota_k\varphi)(x_1, \dots, x_k) = \varphi(x_1, \dots, x_{k-1})$, we have the following property for (4) which is very natural to impose on our valuations..

Property 2 *A consistent renormalisation procedure should produce valuations* Π *such that for any connected* Γ *with* k *legs, one has* $(\Pi\,\mathrm{Del}_k\,\Gamma)(\varphi) = (\Pi\Gamma)(\iota_k\varphi)$ *for all compactly supported test functions* φ *on* $\mathbf{S}^{k-1}$.

(Note that the right hand side is well-defined by Remark 2.2 even though $\iota_k\varphi$ is no longer compactly supported.) To formulate our third property, it will be useful to have a notation for our test functions. We write $\mathcal{D}_k$ for the set of all $\mathcal{C}^\infty$ functions on $\mathbf{S}^k$ with compact support. It will be convenient to consider the following subspaces of $\mathcal{D}_k$. Let $\mathcal{A}$ be a collection of subsets of $\{1, \dots, k\}$ such that every set $A \in \mathcal{A}$

contains at least two elements. Then, we write $\mathscr{D}_k^{(\mathscr{A})} \subset \mathscr{D}_k$ for the set of such functions φ which vanish in a neighbourhood of the set $\Delta_k^{(\mathscr{A})} \subset \mathbf{S}^k$ given by

$$\Delta_k^{(\mathscr{A})} = \{y \in \mathbf{S}^k \,:\, \exists A \in \mathscr{A} \text{ with } y_i = y_j \ \forall i, j \in A\} \,. \tag{8}$$

Because of this definition, we also call a collection $\mathscr{A}$ as above a "collision set". Note that in particular one has $\mathscr{D}_k^{(\emptyset)} = \mathscr{D}_k$.

A first important question to address then concerns the conditions under which the expression (4) converges. A natural notion then is that of the degree of a subgraph of a Feynman diagram. In this article, we define a subgraph $\bar{\Gamma} \subset \Gamma$ to be a subset $\bar{\mathcal{E}}$ of the collection $\mathcal{E}_\star$ of internal edges and a subset $\bar{\mathcal{V}} \subset \mathcal{V}_\star$ of the internal vertices such that $\bar{\mathcal{V}}$ consists precisely of those vertices incident to at least one edge in $\bar{\mathcal{E}}$. (In particular, isolated nodes are not allowed in $\bar{\Gamma}$.) Given such a subgraph $\bar{\Gamma}$, we then set

$$\deg \bar{\Gamma} \stackrel{\text{def}}{=} \sum_{e \in \bar{\mathcal{E}}} \deg \mathfrak{t}(e) + d(|\bar{\mathcal{V}}| - 1) \,. \tag{9}$$

We define the degree of the full Feynman diagram Γ in exactly the same way, with $\bar{\mathcal{E}}$ and $\bar{\mathcal{V}}$ replaced by $\mathcal{E}_\star$ and $\mathcal{V}_\star$. One then has the following result initially due to Weinberg [20]. See also [16, Thm A.3] for the proof of a slightly more general statement which is also notationally closer to the setting considered here.

Proposition 2.3 *If Γ is a Feynman diagram with k legs such that $\deg \bar{\Gamma} > 0$ for every subgraph $\bar{\Gamma} \subset \Gamma$, then the integral in (5) is absolutely convergent for every $\varphi \in \mathscr{D}_k$.* □

We will henceforth call a subgraph $\bar{\Gamma} \subset \Gamma$ *divergent* if $\deg \bar{\Gamma} \leq 0$. A virtually identical proof actually yields the following refined statement which tells us very precisely where exactly there is a need for renormalisation.

Proposition 2.4 *Let Γ be a Feynman diagram with k legs and let $\mathscr{A}$ be a collision set such that, for every connected divergent subgraph $\bar{\Gamma} \subset \Gamma$, there exists $A \in \mathscr{A}$ such that every leg in A is adjacent to $\bar{\Gamma}$. Then (5) is absolutely convergent for every $\varphi \in \mathscr{D}_k^{(\mathscr{A})}$.*

Remark 2.5 Here and below we say that an edge e is adjacent to a subgraph $\bar{\Gamma} \subset \Gamma$ (possibly itself consisting only of a single edge) if e is not an edge of $\bar{\Gamma}$, but shares a vertex with such an edge.

Proof Since the main idea will be useful in the general result, we sketch it here. Note first that we can assume without loss of generality that, for every $A \in \mathscr{A}$, the vertices of $\mathcal{V}_\star$ to which the legs in A are attached are all distinct, since otherwise (5) vanishes identically for $\varphi \in \mathscr{D}_k^{(\mathscr{A})}$.

The key remark is that, for every configuration of points $x \in \mathbf{S}^{\mathcal{V}_\star}$ we can find a binary tree T with leaves given by $\mathcal{V}_\star$ and a label $\mathbf{n}_u \in \mathbf{N}$ for every inner vertex u

of T in such a way that $\mathbf{n}$ is increasing when going from the root to the leaves of T and, for any $v, \bar{v} \in \mathcal{V}_\star$, one has

$$C^{-1}2^{-\mathbf{n}_u} \le \|x_v - x_{\bar{v}}\| \le C2^{-\mathbf{n}_u} , \tag{10}$$

where $u = v \wedge \bar{v}$ is the least common ancestor of v and $\bar{v}$ in T. Here, the constant C only depends on the size of $\mathcal{V}_\star$. (Simply take for T the minimal spanning tree of the point configuration.) Writing $\mathbf{T} = (T, \mathbf{n})$ for this data, we then let $D_{\mathbf{T}} \subset \mathbf{S}^{\mathcal{V}_\star}$ be the set of configurations giving rise to the data $\mathbf{T}$. By analogy with the construction of [17], we call $D_{\mathbf{T}}$ a "Hepp sector".

Remark 2.6 While the type of combinatorial data $(T, \mathbf{n})$ used to index Hepp sectors is identical to that appearing in "Gallavotti-Nicolo trees" [12, 13] and the meaning of the index $\mathbf{n}$ is similar in both cases, there does not appear to be a direct analogy between the terms indexed by this data in both cases.

Remark 2.7 Thanks to the tree structure of T, the quantity $d_{\mathbf{T}}$ given by $d_{\mathbf{T}}(v, \bar{v}) = 2^{-\mathbf{n}_u}$ as above is an ultrametric.

Writing $\mathbf{n}(e)$ for the value of $\mathbf{n}_{e^\uparrow}$, with $e^\uparrow = e_- \wedge e_+$, the integrand of (5) is then bounded by some constant times $\prod_{e \in \mathcal{E}_\star} 2^{-\mathbf{n}(e) \deg \mathfrak{t}(e)}$. Identifying T with its set of internal nodes, one can also show that the measure of $D_{\mathbf{T}}$ is bounded by $\prod_{u \in T} 2^{-d\mathbf{n}_u}$. Finally, by the definition of $\mathcal{D}_k^{(\mathcal{A})}$, there exists a constant N_0 such that the integrand vanishes on sets $D_{\mathbf{T}}$ such that $\sup_{A \in \mathcal{A}} \mathbf{n}_{A^\uparrow} \ge N_0$, where $A^\uparrow$ is the least common ancestor in T of the collection of elements of $\mathcal{V}_\star$ incident to the legs in A. Writing

$$\mathcal{T}_{\mathcal{A}} = \{(T, \mathbf{n}) \,:\, \sup_{A \in \mathcal{A}} \mathbf{n}_{A^\uparrow} < N_0\} ,$$

we conclude that (5) is bounded by some constant multiple of

$$\sum_{\mathbf{T} \in \mathcal{T}_{\mathcal{A}}} \prod_{u \in T} 2^{-\eta_u} , \qquad \eta = d + \sum_{e \in \mathcal{E}_\star} \mathbf{1}_{e^\uparrow} \deg \mathfrak{t}(e) . \tag{11}$$

We now note that the assumption on $\mathcal{A}$ guarantees that, for every node $u \in T$, one has either $\sum_{v \ge u} \eta_v > 0$, or there exists some $A \in \mathcal{A}$ such that $u \le A^\uparrow$. In the latter case, $\mathbf{n}_u$ is bounded from above by N_0. Furthermore, as a consequence of the fact that each connected component of Γ has at least one leg and the kernels $K_{\mathfrak{t}}$ are compactly supported, (5) vanishes on all Hepp sectors with some $\mathbf{n}_u$ sufficiently negative. Combining these facts, and performing the sum in (11) "from the leaves inwards" as in [16, Lem. A.10], it is then straightforward to see that it does indeed converge, as claimed. □

In other words, Proposition 2.4 tells us that the only region in which the integrand of (5) diverges in a non-integrable way consists of an arbitrarily small neighbourhood of those points x for which there exists a divergent subgraph

$\bar{\Gamma} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$ such that $x_u = x_v$ for all vertices $u, v \in \bar{\mathcal{V}}$. It is therefore very natural to impose the following.

Property 3 *A consistent renormalisation procedure should produce valuations* Π *that agree with* (4) *for test functions and Feynman diagrams satisfying the assumptions of Proposition* 2.4.

Finally, a natural set of relations of the canonical valuation Π given by (4) which we would like to retain is those given by integration by parts. In order to formulate this, it is convenient to introduce the notion of a *half-edge*. A half-edge is a pair (e, v) with $e \in \mathcal{E}$ and $v \in \{e_+, e_-\}$. It is said to be *incoming* if $v = e_+$ and *outgoing* if $v = e_-$. Given an edge e, we also write $e_\leftarrow$ and $e_\rightarrow$ for the two half-edges (e, e_-) and (e, e_+). Given a Feynman diagram Γ, a half-edge (e, v), and $k \in \mathbf{N}^d$, we then write $\partial^k_{(e,v)}\Gamma$ for the element of $\mathcal{T}$ obtained from Γ by replacing the decoration $\mathfrak{t}$ of the edge e by $\mathfrak{t}^{(k)}$ and then multiplying the resulting Feynman diagram by $(-1)^{|k|}$ if the half-edge (e, v) is outgoing. We then write $\partial\mathcal{T}$ for the smallest subspace of $\mathcal{T}$ such that, for every Feynman diagram Γ, every $i \in \{1, \dots, d\}$ and every inner vertex $v \in \mathcal{V}_\star$ of Γ, one has

$$\sum_{e \sim v} \partial^{\delta_i}_{(e,v)}\Gamma \in \partial\mathcal{T} \ , \tag{12}$$

where $e \sim v$ signifies that the edge e is incident to the vertex v and δ_i is the ith canonical element of $\mathbf{N}^d$. By integration by parts, it is immediate that if the kernels $K_\mathfrak{t}$ are all smooth, then the canonical valuation (4) satisfies $\Pi\partial\mathcal{T} = 0$. It is therefore natural to impose the following.

Property 4 *A consistent renormalisation procedure should produce valuations* Π *that vanish on* $\partial\mathcal{T}$.

Setting $\mathcal{H} = \mathcal{T}/\partial\mathcal{T}$, we can therefore consider a valuation as a map $\Pi\colon \mathcal{H} \to \mathcal{S}$. Note that since $\partial\mathcal{T}$ is an ideal of $\mathcal{T}$ which respects its grading, $\mathcal{H}$ is again a graded algebra. Furthermore, since $\partial\mathcal{T}$ is invariant under the action of the symmetric group, $\mathfrak{S}_k$ acts naturally on $\mathcal{H}_k$. In particular, Property 1 can be formulated in $\mathcal{H}$ rather than $\mathcal{T}$ and it is not difficult to see that the deletion operation Del_k introduced in Property 2 also makes sense on $\mathcal{H}$. This motivates the following definition.

Definition 2.8 A valuation $\Pi\colon \mathcal{H} \to \mathcal{S}$ is *consistent* for the kernel assignment K if it satisfies Properties 1, 2 and 3.

2.2 Some Algebraic Operations on Feynman Diagrams

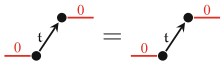
In order to satisfy Property 3, we will consider valuations that differ from the canonical one only by counterterms of the same form, but with some of the factors of (4) corresponding to divergent subgraphs replaced by a suitable derivative of a delta function, just like what we did in (7).

These counterterms can again be encoded into Feynman diagrams with the same number of legs as the original diagram, multiplied by a suitable weight. We are therefore looking for a procedure which, given a smooth kernel assignment $K \in \mathcal{K}_\infty^-$, builds a linear map $M^K : \mathcal{T} \to \mathcal{T}$ such that if we define a "renormalised" valuation $\hat{\Pi}^K$ by

$$\hat{\Pi}^K \Gamma = \Pi^K M^K \Gamma \,, \tag{13}$$

with Π^K the canonical valuation given by (4), then $K \mapsto \hat{\Pi}^K$ is a renormalisation procedure which extends continuously to all of $\mathcal{K}_0^-$. We would furthermore like M^K to differ from the identity only by terms of the form described above, obtained by contracting divergent subgraphs to a derivative of a delta function.

The procedure (7) is exactly of this form with

M^K  $= \mathrm{Id} - \sum_{|k|+\deg \mathfrak{t} \le -d} c_k \cdot$, $\qquad c_k = \frac{1}{k!} \int_S x^k K_{\mathfrak{t}}(x)\, dx \,.$

(14)

Note that the condition $\deg \mathfrak{t} \le -d$ which is required for M^K to differ from the identity is precisely the condition that the subgraph $\bullet \xrightarrow{\mathfrak{t}} \bullet$ is divergent, which then guarantees that this example satisfies Property 3.

It is natural to index the constants appearing in the terms of such a renormalisation map by the corresponding subgraphs that were contracted. These subgraphs then have no legs anymore, but may require additional decorations describing the powers of x appearing in the expression for c_k above. We therefore give the following definition, where the choice of terminology is chosen to be consistent with the QFT literature.

Definition 2.9 A *vacuum diagram* consists of a Feynman diagram $\Gamma = (\mathcal{V}, \mathcal{E})$ with exactly one leg per connected component, endowed additionally with a node decoration $\mathfrak{n} : \mathcal{V}_\star \to \mathbf{N}^d$. We also impose that each leg has label δ. We say that a connected vacuum diagram is *divergent* if $\deg \Gamma \le 0$, where

$$\deg \Gamma = \sum_{e \in \mathcal{E}} \mathfrak{t}(e) + \sum_{v \in \mathcal{V}} |\mathfrak{n}(v)| + d(|\mathcal{V}| - 1) \,.$$

We extend this to arbitrary vacuum diagrams by imposing that $\deg(\Gamma_1 \bullet \Gamma_2) = \deg \Gamma_1 + \deg \Gamma_2$.

One should think of a connected vacuum diagram Γ as encoding the constant

$$\Pi_-^K \Gamma \stackrel{\text{def}}{=} \int_{S^{\mathcal{V}_\star \setminus \{v_\star\}}} \prod_{e \in \mathcal{E}_\star} K_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \prod_{w \in \mathcal{V}_\star} (x_w - x_{v_\star})^{\mathfrak{n}(w)}\, dx \tag{15}$$

where $v_\star$ is the element of $\mathcal{V}_\star$ that has the unique leg attached to it. This is then extended multiplicatively to all vacuum diagrams. In view of this, it is also natural to ignore the ordering of the legs for vacuum diagrams, and we will always do this from now on.

Write now $\hat{\mathcal{T}}_-$ for the algebra of all vacuum diagrams such that each connected component has at least one internal edge and by $\mathcal{T}_- \subset \hat{\mathcal{T}}_-$ for the subalgebra generated by those diagrams such that each connected components is divergent. Since we ignored the labelling of legs, the product $\bullet$ turns $\hat{\mathcal{T}}_-$ into a commutative algebra. Note that if we write $\mathcal{J}_+ \subset \hat{\mathcal{T}}_-$ for the ideal generated by all vacuum diagrams Γ with $\deg \Gamma > 0$, then we have a natural isomorphism

$$\mathcal{T}_- \approx \hat{\mathcal{T}}_- / \mathcal{J}_+ \,.$$

Similarly to above, it is natural to identify vacuum diagrams related to each other by integration by parts, but also those related by changing the location of the leg(s). In order to formalise this, we reinterpret a connected vacuum diagram as above as a Feynman diagram "with 0 legs", but with one of the vertices being distinguished, which is of course completely equivalent, and we write it as $(\Gamma, v_\star, \mathfrak{n})$. With this notation, we define $\partial\hat{\mathcal{T}}_-$ as the smallest ideal of $\hat{\mathcal{T}}_-$ such that, for every connected $(\Gamma, v_\star, \mathfrak{n})$ one has the following.

- For every vertex $v \in \mathcal{V} \setminus \{v_\star\}$ and every $i \in \{1, \dots, d\}$, one has

$$\sum_{e \sim v} (\partial^{\delta_i}_{(e,v)} \Gamma, v_\star, \mathfrak{n}) + \mathfrak{n}(v)_i (\Gamma, v_\star, \mathfrak{n} - \delta_i \mathbf{1}_v) \in \partial\hat{\mathcal{T}}_- \,, \tag{16}$$

  where $\mathbf{1}_v$ denotes the indicator function of $\{v\}$.
- One has

$$\sum_{e \sim v_\star} (\partial^{\delta_i}_{(e,v_\star)} \Gamma, v_\star, \mathfrak{n}) - \sum_{v \in \bar{\mathcal{V}}} \mathfrak{n}(v)_i (\Gamma, v_\star, \mathfrak{n} - \delta_i \mathbf{1}_v) \in \partial\hat{\mathcal{T}}_- \,, \tag{17}$$

- For every vertex $v \in \mathcal{V}$, one has

$$(\Gamma, v_\star, \mathfrak{n}) - \sum_{\mathfrak{m} \colon \mathcal{V} \to \mathbf{N}^d} (-1)^{|\mathfrak{m}|} \binom{\mathfrak{n}}{\mathfrak{m}} (\Gamma, v, \mathfrak{n} - \mathfrak{m} + \Sigma\mathfrak{m} \mathbf{1}_{v_\star}) \in \partial\hat{\mathcal{T}}_- \,, \tag{18}$$

  where $\Sigma\mathfrak{m} = \sum_u \mathfrak{m}(u)$ and we use the convention $\mathfrak{m}! = \prod_{u \in \mathcal{V}} \prod_{i=1}^d \mathfrak{m}(u)_i!$ to define the binomial coefficients, with the additional convention that the coefficient vanishes unless $\mathfrak{m} \le \mathfrak{n}$ everywhere.

Remark 2.10 One can verify that if $K \in \mathcal{H}^-_\infty$ and Π^K_- is given by (15), then $\partial\hat{\mathcal{T}}_- \in \ker \Pi^K_-$. In the case of (16) and (17), this is because the integrand is then a total derivative with respect to $(x_v)_i$ and $(x_{v_\star})_i$ respectively. In the case of (18), this can be seen by writing $(x_w - x_{v_\star})^{\mathfrak{n}(w)} = ((x_w - x_v) - (x_{v_\star} - x_v))^{\mathfrak{n}(w)}$ and applying the multinomial theorem.

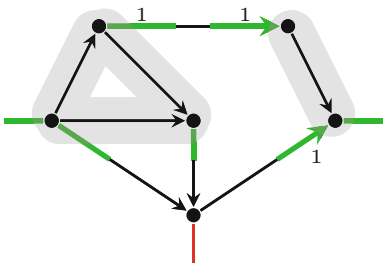


Fig. 2 Example of a subgraph (shaded) and its boundary (green)

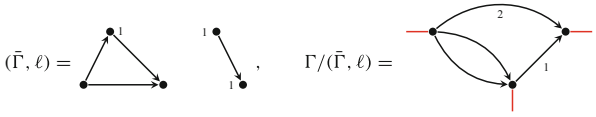
Remark 2.11 The expressions (16) and (17) are consistent with (12) in the special case $\mathfrak{n} = 0$. Considering the case $v = v_\star$ in (18), it is also straightforward to verify that $(\Gamma, v_\star, \mathfrak{n}) \in \partial\hat{\mathcal{T}}_-$ as soon as $\mathfrak{n}(v_\star) \neq 0$.

As before, we then write $\hat{\mathscr{H}}_-$ as a shorthand for $\hat{\mathcal{T}}_-/\partial\hat{\mathcal{T}}_-$ and similarly for $\mathscr{H}_-$. (This is well-defined since $\partial\hat{\mathcal{T}}_-$ does not mix elements of different degree.) As a consequence of Remark 2.10, we see that every $K \in \mathcal{H}_\infty^-$ yields a character Π_-^K of $\hat{\mathscr{H}}_-$ and therefore also of $\mathscr{H}_-$.

Given a Feynman diagram Γ and a subgraph $\bar{\Gamma} \subset \Gamma$, we can (and will) identify $\bar{\Gamma}$ with an element of $\hat{\mathscr{H}}_-$, obtained by setting all the node decorations to 0. By (18) we do not need to specify where we attach leg(s) to $\bar{\Gamma}$ since these elements are all identified in $\hat{\mathscr{H}}_-$. We furthermore write $\partial\bar{\Gamma}$ for the set of all *half-edges* adjacent to $\bar{\Gamma}$. Figure 2 shows an example of a Feynman diagram with a subgraph $\bar{\Gamma}$ shaded in grey and $\partial\bar{\Gamma}$ indicated in green. Legs can also be part of $\partial\bar{\Gamma}$ as is the case in our example, but they can *not* be part of $\bar{\Gamma}$ by our definition of a subgraph. Note also that the edge joining the two vertices at the top appears as two distinct half-edges in $\partial\bar{\Gamma}$. Given furthermore a map $\ell\colon \partial\bar{\Gamma} \to \mathbf{N}^d$ (canonically extended to vanish on all other half-edges of Γ), we then define the following two objects.

- A vacuum diagram $(\bar{\Gamma}, \pi\ell)$ which consists of the graph $\bar{\Gamma}$ endowed with the edge decoration inherited from Γ, as well as the node decoration $\mathfrak{n} = \pi\ell$ given by $(\pi\ell)(v) = \sum_{e\,:\,(e,v)\in\partial\Gamma} \ell(e, v)$.
- A Feynman diagram $\Gamma/(\bar{\Gamma}, \ell)$ obtained by contracting the connected components of $\bar{\Gamma}$ to nodes and applying ℓ to the resulting diagram in the sense that, for edges $e \in \mathcal{E} \setminus \bar{\mathcal{E}}$ adjacent to $\partial\bar{\Gamma}$ and with label (in Γ) given by $\mathfrak{t}$, we replace their label by $\mathfrak{t}^{(\ell(e_\leftarrow)+\ell(e_\rightarrow))}$.

In the example of Fig. 2, where non-zero values of ℓ are indicated by small labels, we have



where a label k on an edge means that if it had a decoration $\mathfrak{t}$ in Γ, then it now has a decoration $\mathfrak{t}^{(k)}$. Given a map $\ell\colon \partial\bar\Gamma \to \mathbf{N}^d$ as above, we also write "out ℓ" as a shorthand for the restriction of ℓ to outgoing half-edges. With these notations at hand, we define a map $\Delta\colon \mathcal{T} \to \mathcal{H}_- \otimes \mathcal{H}$ by

$$\Delta\Gamma = \sum_{\bar\Gamma\subset\Gamma} \sum_{\ell\colon \partial\bar\Gamma\to\mathbf{N}^d} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!}(\bar\Gamma, \pi\ell) \otimes \Gamma/(\bar\Gamma, \ell)\;, \tag{19}$$

where we use the same conventions for factorials as in (18). Note that since the right hand side is identified with an element of $\mathcal{H}_- \otimes \mathcal{H}$, this sum is finite. Indeed, unless $(\bar\Gamma, \pi\ell) \in \mathcal{T}_-$, which only happens for finitely many choices of ℓ, the corresponding factor is identified with 0 in $\mathcal{H}_-$.

Remark 2.12 For any fixed Γ this sum is actually finite since there are only finitely many subgraphs and since, for large enough ℓ, $(\bar\Gamma, \pi\ell)$ is no longer in $\mathcal{T}_-$.

Remark 2.13 The factor $(-1)^{|\operatorname{out}\ell|}$ appearing here encodes the fact that for an edge e, having $\ell(e, u) = k$ means that in the resulting Feynman diagram $\Gamma/(\bar\Gamma, \ell)$, one would like to replace the factor $K_{\mathfrak{t}}(x_{e_+} - x_{e_-})$ by its kth derivative with respect to x_u, which is precisely what happens when one replaces the corresponding connected component of $\bar\Gamma$ by a derivative of a delta function. In the case when $u = e_-$, namely when the half-edge is outgoing, this is indeed the same as $(-1)^{|k|}(D^k K_{\mathfrak{t}})(x_{e_+} - x_{e_-})$, while the factor $(-1)^{|k|}$ is absent for incoming half-edges.

It turns out that one has the following.

Proposition 2.14 *The map Δ is well-defined as a map from $\mathcal{H}$ to $\mathcal{H}_- \otimes \mathcal{H}$.*

Before we start our proof, recall the following version of the Chu-Vandermonde identity

Lemma 2.15 *Given finite sets S, $\bar S$ and maps $\pi : S \to \bar S$ and $\ell\colon S \to \mathbf{N}$, we define $\pi_\star\ell\colon \bar S \to \mathbf{N}$ by $\pi_\star\ell(x) = \sum_{y\in\pi^{-1}(x)} \ell(y)$.*

Then, for every finite set S and every $k\colon S \to \mathbf{N}$, one has the identity

$$\sum_{\ell\,:\,\pi_\star\ell} \binom{k}{\ell} = \binom{\pi_\star k}{\pi_\star\ell}\;,$$

where the sum runs over all possible choices of ℓ such that $\pi_\star\ell$ is fixed. □

Proof (of Proposition 2.14) We first show that for $\Gamma \in \mathcal{T}$ the right hand side of (19) is well-defined as an element of $\mathcal{H}_- \otimes \mathcal{H}$, which is a priori not obvious since we did not specify where the legs of $(\bar\Gamma, \ell)$ are attached. Our aim therefore is to show that, for any fixed $L \in \mathbf{N}^d$, the expression

$$\sum_{\substack{\ell\colon \partial\bar\Gamma\to\mathbf{N}^d\\ \Sigma\ell = L}} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!}(\bar\Gamma, v, \pi\ell) \otimes \Gamma/(\bar\Gamma, \ell) \tag{20}$$

is independent of $v \in \bar{V}$ in $\mathscr{H}_- \otimes \mathscr{H}$. By Remark 2.11, we can restrict the sum over ℓ to those values such that ℓ vanishes on the set A_v of all half-edges incident to v since $(\bar{\Gamma}, v, \ell) = 0$ in $\mathscr{H}_-$ for those ℓ for which this is not the case. Fixing some arbitrary $u \neq v$ and using (18) as well as Lemma 2.15, we then see that (20) equals

$$\sum_{\substack{\ell\colon \partial\bar{\Gamma}\setminus A_v \to \mathbf{N}^d \\ \Sigma\ell = L}} \sum_{m\colon \partial\bar{\Gamma} \to \mathbf{N}^d} \frac{(-1)^{|\operatorname{out}\ell| + |m|}}{\ell!} \binom{\ell}{m} (\bar{\Gamma}, u, \pi\ell - \pi m + \Sigma m \mathbf{1}_v) \otimes \Gamma/(\bar{\Gamma}, \ell) \,.$$

Writing $k = \ell - m$, we rewrite this expression as

$$\sum_{\substack{k\colon \partial\bar{\Gamma}\setminus A_v \to \mathbf{N}^d \\ \Sigma k \le L}} \sum_{\substack{m\colon \partial\bar{\Gamma}\setminus A_v \to \mathbf{N}^d \\ \Sigma m = L - \Sigma k}} \frac{(-1)^{|\operatorname{out}k| + |\operatorname{out}m| + |m|}}{k!m!} (\bar{\Gamma}, u, \pi k + \Sigma m \mathbf{1}_v) \otimes \Gamma/(\bar{\Gamma}, k + m) \,.$$

At this stage we note that, as a consequence of (12), we have for every subset $A \subset \partial\bar{\Gamma}$ and every $M \in \mathbf{N}^d$ the identity

$$\sum_{\substack{m\colon \partial\bar{\Gamma}\setminus A \to \mathbf{N}^d \\ \Sigma m = M}} \frac{(-1)^{|\operatorname{out}m|}}{m!} \Gamma/(\bar{\Gamma}, k + m) = \sum_{\substack{n\colon A \to \mathbf{N}^d \\ \Sigma n = M}} \frac{(-1)^{|n| + |\operatorname{out}n|}}{n!} \Gamma/(\bar{\Gamma}, k + n) \,.$$

Inserting this into the above expression and noting that for functions n supported on A_v one has $\pi n = \Sigma n \mathbf{1}_v$, we conclude that it equals

$$\sum_{\substack{k\colon \partial\bar{\Gamma}\setminus A_v \to \mathbf{N}^d \\ \Sigma k \le L}} \sum_{\substack{n\colon A_v \to \mathbf{N}^d \\ \Sigma n = L - \Sigma k}} \frac{(-1)^{|\operatorname{out}k| + |\operatorname{out}n|}}{k!n!} (\bar{\Gamma}, u, \pi k + \pi n) \otimes \Gamma/(\bar{\Gamma}, k + n) \,.$$

Setting $\ell = k + n$ and noting that $k!n! = (k+n)!$ since k and n have disjoint support, we see that this is indeed equal to (20) with v replaced by u, as claimed.

It remains to show that Δ is well-defined on $\mathscr{H}$, namely that $\Delta\tau = 0$ in $\mathscr{H}_- \otimes \mathscr{H}$ for $\tau \in \partial\mathscr{T}$. Choose a Feynman diagram Γ, an inner vertex $v \in V_\star$, an index $i \in \{1, \dots, d\}$, and a subgraph $\bar{\Gamma} \subset \Gamma$. Writing $\bar{A}_v$ for the half-edges in $\bar{\Gamma}$ adjacent to v and A_v for the remaining half-edges adjacent to v (so that $A_v \subset \partial\bar{\Gamma}$), it suffices to show that

$$\begin{aligned} &\sum_{h \in A_v} \sum_{\substack{\ell\colon \partial\bar{\Gamma} \to \mathbf{N}^d \\ (\bar{\Gamma}, \ell) \in \mathscr{T}_-}} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!} (\bar{\Gamma}, \pi\ell) \otimes \partial_h^{\delta_i} \Gamma/(\bar{\Gamma}, \ell) \\ &\quad + \sum_{h \in \bar{A}_v} \sum_{\substack{\ell\colon \partial\bar{\Gamma} \to \mathbf{N}^d \\ \partial_{i,v}(\bar{\Gamma}, \ell) \in \mathscr{T}_-}} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!} \partial_h^{\delta_i} (\bar{\Gamma}, \pi\ell) \otimes \Gamma/(\bar{\Gamma}, \ell) = 0 \end{aligned} \tag{21}$$

in $\mathcal{H}_- \times \mathcal{H}$, where we used the shorthand notation $\partial_{i,v}(\bar\Gamma, \ell) \in \mathcal{T}_-$ for the condition $\partial_h^{\delta_i}(\bar\Gamma, \ell) \in \mathcal{T}_-$, which is acceptable since this condition does not depend on which half-edge h one considers. If v is not contained in $\bar\Gamma$, then the second term vanishes and A_v consists exactly of all edges adjacent to v in $\Gamma/(\bar\Gamma, \ell)$, so that the first term vanishes as well by (12). If v is contained in $\bar\Gamma$, then we attach the leg of the corresponding connected component $\bar\Gamma_0$ of $\bar\Gamma$ to v itself, so that in particular the sum over ℓ can be restricted to values supported on $\partial\bar\Gamma \setminus A_v$. By (17), the second term is then equal to

$$\sum_{h\in\partial\bar\Gamma_0\setminus A_v} \sum_{\substack{\ell\colon \partial\bar\Gamma\setminus A_v\to\mathbf{N}^d\\ \partial_{i,v}(\bar\Gamma,\ell)\in\mathcal{T}_-}} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!}\ell(h)_i(\bar\Gamma, v, \pi(\ell - \delta_i\mathbf{1}_h)) \otimes \Gamma/(\bar\Gamma, \ell)\ ,$$

which can be rewritten as

$$\sum_{h\in\partial\bar\Gamma_0\setminus A_v} \sum_{\substack{\ell\colon \partial\bar\Gamma\to\mathbf{N}^d\\ (\bar\Gamma,\ell)\in\mathcal{T}_-}} \frac{(-1)^{|\operatorname{out}\ell|+\delta_{h\in\operatorname{out}}}}{\ell!}(\bar\Gamma, v, \pi\ell) \otimes \Gamma/(\bar\Gamma, \ell + \delta_i\mathbf{1}_h)\ .$$

Inserting this into (21), we conclude that this expression equals

$$\sum_{h\in\partial\bar\Gamma_0} \sum_{\substack{\ell\colon \partial\bar\Gamma\to\mathbf{N}^d\\ (\bar\Gamma,\ell)\in\mathcal{T}_-}} \frac{(-1)^{|\operatorname{out}\ell|}}{\ell!}(\bar\Gamma, \pi\ell) \otimes \partial_h^{\delta_i}\Gamma/(\bar\Gamma, \ell)$$

which vanishes in $\mathcal{H}_- \otimes \mathcal{H}$ by (12) since the half-edges in $\partial\bar\Gamma_0$ are precisely all the half-edges adjacent in $\Gamma/(\bar\Gamma, \ell)$ to the node that $\bar\Gamma_0$ was contracted to.

For any element $g\colon \mathcal{H}_- \to \mathbf{R}$ of the dual of $\mathcal{H}_-$, we now have a linear map $M^g\colon \mathcal{H} \to \mathcal{H}$ by

$$M^g\Gamma = (g \otimes \mathrm{id})\Delta\Gamma\ ,$$

which leads to a valuation $\Pi_g^K\colon \mathcal{H} \to \mathcal{S}$ by setting

$$\Pi_g^K = \Pi^K \circ M^g \tag{22}$$

as in (13), with Π^K the canonical valuation (4). Note that this is well-defined since $\Pi^K\partial\mathcal{T} = 0$, as already remarked. In particular, we can also view Π_g^K as a map from $\mathcal{T}$ to $\mathcal{S}$.

For any choice of g (depending on the kernel assignment K), such a valuation then automatically satisfies Properties 3 and 4, since these were encoded in the definition of the space $\mathcal{H}$, as well as Property 2 since the action of Δ commutes with the operation of "amputation of the kth leg" on the subspace on which the

latter is defined. In general, such a valuation may fail to satisfy Property 1, but if we restrict ourselves to elements $g : \mathscr{H}_- \to \mathbf{R}$ that are also characters, one has

$$M^g(\Gamma_1 \bullet \Gamma_2) = (M^g\Gamma_1) \bullet (M^g\Gamma_2) \,.$$

Since $\partial\mathscr{T}$ is an ideal, this implies that the valuation Π_g^K is multiplicative as a map from $\mathscr{T}$ to $\mathcal{S}$, as required by Property 1. We have therefore shown the following.

Proposition 2.16 *For every character $g : \mathscr{H}_- \to \mathbf{R}$, the valuation Π_g^K is consistent for K in the sense of Definition 2.8.* □

Writing $\mathscr{G}_-$ for the space of characters of $\mathscr{T}_-$, it is therefore natural to define a "consistent renormalisation procedure" as a map $\mathscr{R} : \mathscr{K}_\infty^- \to \mathscr{G}_-$ such that the map

$$K \mapsto \hat{\Pi}^K = \Pi^K \circ M^{\mathscr{R}(K)} \,, \tag{23}$$

where Π^K denotes the canonical valuation given by (4), extends continuously to all of $\mathscr{K}_0^-$. Our question now turns into the question whether such a map exists.

Remark 2.17 We do certainly *not* want to impose that $\mathscr{R}$ extends continuously to all of $\mathscr{K}_0^-$ since this would then imply that Π^K extends to all of $\mathscr{K}_0^-$ which is obviously false.

2.3 A Hopf Algebra

In this subsection, we address the following point. We have seen that every character g of $\mathscr{H}_-$ allow us to build a new valuation Π_g from the canonical valuation Π associated to a smooth kernel assignment. We can then take a second character h and build a new valuation $\Pi_g \circ M^h$. It is natural to ask whether this would give us a genuinely new valuation or whether this valuation is again of the form $\Pi_{\bar{g}}$ for some character $\bar{g}$. In other words, does $\mathscr{G}_-$ have a group structure, so that $g \mapsto M^g$ is a left action of this group on the space of all valuations?

In order to answer this question, we first define a map $\Delta^- : \hat{\mathscr{T}}_- \to \mathscr{H}_- \otimes \hat{\mathscr{H}}_-$ in a way very similar to the map Δ, but taking into account the additional labels $\mathfrak{n}$:

$$\Delta^-(\Gamma, v_\star, \mathfrak{n}) = \sum_{\bar{\Gamma} \subset \Gamma} \sum_{\substack{\bar{\ell} : \partial\bar{\Gamma} \to \mathbf{N}^d \\ \bar{\mathfrak{n}} : \bar{\mathscr{V}} \to \mathbf{N}^d}} \frac{(-1)^{|\operatorname{out}\bar{\ell}|}}{\bar{\ell}!} \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} (\bar{\Gamma}, \bar{\mathfrak{n}} + \pi\bar{\ell}) \otimes (\Gamma, v_\star, \mathfrak{n} - \bar{\mathfrak{n}})/(\bar{\Gamma}, \bar{\ell}) \,. \tag{24}$$

Here, we define $(\Gamma, v_\star, \mathfrak{n})/(\bar{\Gamma}, \bar{\ell})$ similarly to before, with the node-label of the quotient graph obtained by summing over the labels of all the nodes that get contracted to the same node. If $\bar{\Gamma}$ completely contains one (or several) connected components of Γ, then this definition could create graphs that contain isolated

nodes, which is forbidden by our definition of $\hat{\mathcal{T}}_-$. Given (15), it is natural to identify isolated nodes with vanishing node-label with the empty diagram $\mathbf{1}$, while we identify those with non-vanishing node-labels with 0. In particular, it follows that

$$\Delta^- \tau = \tau \otimes \mathbf{1} + \mathbf{1} \otimes \tau + \Delta' \tau \ ,$$

where each of the terms appearing in $\Delta'\tau$ is such that both factors contain at least one edge.

Note the strong similarity with [2, Def. 3.3] which looks formally almost identical, but with graphs replaced by trees. As before, one then has

Proposition 2.18 *The map Δ^- is well-defined both as a map $\hat{\mathcal{H}}_- \to \mathcal{H}_- \otimes \hat{\mathcal{H}}_-$ and a map $\mathcal{H}_- \to \mathcal{H}_- \otimes \mathcal{H}_-$.* □

It follows immediately from the definitions that Δ^- is multiplicative. What is slightly less obvious is that it also has a nice coassociativity property as follows.

Proposition 2.19 *The identities*

$$(\Delta^- \otimes \mathrm{id})\Delta = (\mathrm{id} \otimes \Delta)\Delta \ , \qquad (\Delta^- \otimes \mathrm{id})\Delta^- = (\mathrm{id} \otimes \Delta^-)\Delta^- \tag{25}$$

hold between maps $\mathcal{B} \to \mathcal{H}_- \otimes \mathcal{B} \otimes \mathcal{B}$ for $\mathcal{B} = \mathcal{H}$ in the case of the first identity and for $\mathcal{B} \in \{\mathcal{H}_-, \hat{\mathcal{H}}_-\}$ in the case of the second one.

Proof We only verify the second identity since the first one is essentially a special case of the second one. The difference is the presence of legs, which are never part of the subgraphs appearing in the definition of Δ, but otherwise play the same role as a "normal" edge.

Fix now a Feynman diagram Γ as well as two subgraphs Γ_1 and Γ_2 with the property that each connected component of Γ_1 is either contained in Γ_2 or vertex-disjoint from it. We also write $\bar{\Gamma} = \Gamma_1 \cup \Gamma_2$ and $\Gamma_{1,2} = \Gamma_1 \cap \Gamma_2$. There is then a natural bijection between the terms appearing in $(\Delta^- \otimes \mathrm{id})\Delta^-$ and those appearing in $(\mathrm{id} \otimes \Delta^-)\Delta^-$ obtained by noting that first extracting $\bar{\Gamma}$ from Γ and then extracting Γ_1 from $\bar{\Gamma}$ is the same as first extracting Γ_1 from Γ and then extracting $\Gamma_2 / \Gamma_{1,2}$ from Γ / Γ_1. It therefore remains to show that the labellings and combinatorial factors appearing for these terms are also the same. This in turn is a consequence from a generalisation of the Chu-Vandermonde identity and can be obtained in almost exactly the same way as [2, Prop. 3.9].

If we write $\mathbf{1}$ for the empty vacuum diagram and $\mathbf{1}^*$ for the element of $\mathcal{G}_-$ that vanishes on all non-empty diagrams, then we see that $(\mathcal{H}_-, \Delta^-, \bullet, \mathbf{1}, \mathbf{1}^*)$ is a bialgebra. Since it also graded (by the number of edges of a diagram) and connected (the only diagram with 0 edges is the empty one), it is a Hopf algebra so that $\mathcal{G}_-$ is indeed a group with product

$$f \circ g \stackrel{\text{def}}{=} (f \otimes g)\Delta^- \ ,$$

and inverse $g^{-1} = g\mathcal{A}$, where $\mathcal{A}$ is the antipode. The first identity in (25) then implies that the map $g \mapsto M^g = (g \otimes \mathrm{id})\Delta$ does indeed yield a group action on the space of valuations, thus answering positively the question asked at the start of this section.

2.4 Twisted Antipodes and the BPHZ Theorem

An arbitrary character g of $\mathcal{H}_-$ is uniquely determined by its value on connected vacuum diagrams Γ with $\deg \Gamma \leq 0$. Comparing (14) with (19), this would suggest that a natural choice of renormalisation procedure $\mathcal{R}$ is given by simply setting

$$\mathcal{R}(K)\Gamma = -\Pi_-^K \Gamma \,,$$

as this would indeed reproduce the expression (7). Unfortunately, while this choice does yield valuations that extend continuously to all kernel assignments in $\mathcal{K}_0^-$ for a class of "simple" Feynman diagrams, it fails to do so for all of them.

Following [5, 6], a more sophisticated guess would be to set $\mathcal{R}(K)\Gamma = \Pi_-^K \mathcal{A}\Gamma$, for $\mathcal{A}$ the antipode of $\mathcal{H}_-$ endowed with the Hopf algebra structure described in the previous section. The reason why this identity also fails to do the trick can be illustrated with the following example. Consider the case $d = 1$ and two labels with $|\mathfrak{t}_1| = -1/3$ and $|\mathfrak{t}_2| = -4/3$. Drawing edges decorated with $\mathfrak{t}_1$ in black and edges decorated with $\mathfrak{t}_2$ in blue, we then consider

$\Gamma =$  ,

which has degree $\deg \Gamma = 0$. Since Γ has only one leg, the naive valuation $\Pi^K \Gamma$ can be identified with the real number

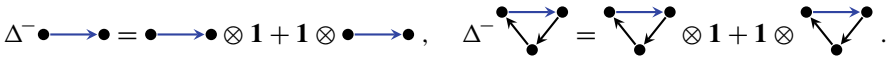
$$\Pi^K \Gamma = (K_1 * K_2 * K_1)(0) \,,$$

where we wrote $K_i \stackrel{\text{def}}{=} K_{\mathfrak{t}_i}$ and $*$ denotes convolution. Since this might diverge for a generic kernel assignment in $\mathcal{K}_0^-$, even if K_2 is replaced by its renormalised version, there appears to be no good canonical renormalised value for $\hat{\Pi}^K \Gamma$, so we would expect to just have $\hat{\Pi}^K \Gamma = 0$.

Let's see what happens instead if we choose the renormalisation procedure $\mathcal{R}(K)\Gamma = \Pi_-^K \mathcal{A}\Gamma$. It follows from the definition of Δ that

$\Delta\Gamma = \mathbf{1} \otimes$  $+$, (26)

since and  are the only subgraphs of negative degree, but their degree remains above -1 so that no node-decorations are added. Note furthermore that in $\mathcal{H}_-$ one has the identities



The reason why there is no additional term analogous to the middle term of (26) appearing in the second identity is that the corresponding factor would be of positive degree and therefore vanishes when viewed as an element of $\mathcal{H}_-$. As a consequence, we have $\mathcal{A}\tau = -\tau$ in both cases, so that the first and last terms of (26) cancel out and we are eventually left with

$$\hat{\Pi}^K \Gamma = -(K_1 * K_1)(0) \cdot K_2(0) \,,$$

which is certainly not desirable since it might diverge as well.

The way out of this conundrum is to define a *twisted antipode* $\hat{\mathcal{A}} : \mathcal{H}_- \to \hat{\mathcal{H}}_-$ which is defined by a relation very similar to that defining the antipode, but this time guaranteeing that the renormalised valuation vanishes on those diagrams that encode "potentially diverging constants" as above. Here, the renormalised valuation is defined by setting

$$\mathcal{R}(K)\Gamma = \Pi_-^K \hat{\mathcal{A}} \Gamma \,, \tag{27}$$

where Π_-^K is defined by (15). Writing $\mathcal{M} : \hat{\mathcal{H}}_- \otimes \hat{\mathcal{H}}_- \to \hat{\mathcal{H}}_-$ for the product, we define $\hat{\mathcal{A}}$ to be such that

$$\mathcal{M}(\hat{\mathcal{A}} \otimes \mathrm{id})\Delta^- \Gamma = 0 \,, \tag{28}$$

for every non-empty connected vacuum diagram $\Gamma \in \hat{\mathcal{H}}_-$ with $\deg \Gamma \leq 0$. At first sight, this looks exactly like the definition of the antipode. The difference is that the map Δ^- in the above expression goes from $\hat{\mathcal{H}}_-$ to $\mathcal{H}_- \otimes \hat{\mathcal{H}}_-$, so that no projection onto diverging diagrams takes place on the right factor. If we view $\mathcal{H}_-$ as a subspace of $\hat{\mathcal{H}}_-$, then the antipode satisfies the identity

$$\mathcal{M}(\mathcal{A} \otimes \pi_-)\Delta^- \Gamma = 0 \,,$$

where $\pi_- : \hat{\mathcal{H}}_- \to \mathcal{H}_-$ is the projection given by quotienting by the ideal $\mathcal{I}_+$ generated by diagrams with strictly positive degree. We have the following simple lemma.

Lemma 2.20 *There exists a unique map $\hat{\mathcal{A}} : \mathcal{H}_- \to \hat{\mathcal{H}}_-$ satisfying (28). Furthermore, the map Π^K_{BPHZ} given by (23) with $\mathcal{R}(K) = \Pi_-^K \hat{\mathcal{A}}$ is indeed a valuation.*

Proof The existence and uniqueness of $\hat{\mathcal{A}}$ is immediate by performing an induction over the number of edges. Defining $\Delta'^{(k)} : \mathcal{H}_- \to \hat{\mathcal{H}}_-^{\otimes(k+1)}$ inductively by $\Delta'^{(0)} = \iota$ and then

$$\Delta'^{(k+1)} = (\Delta'^{(k)} \otimes \mathrm{id})\Delta'\iota \,,$$

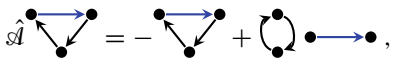
where $\iota : \mathcal{H}_- \to \hat{\mathcal{H}}_-$ is the canonical injection, one obtains the (locally finite) Neumann series

$$\hat{\mathcal{A}} = \sum_{k\geq 0} (-1)^{k+1} \mathcal{M}^{(k)} \Delta'^{(k)} \,, \tag{29}$$

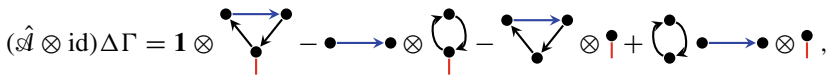
where $\mathcal{M}^{(k)} : \hat{\mathcal{H}}_-^{\otimes(k+1)} \to \hat{\mathcal{H}}_-$ is the multiplication operator. The uniqueness also immediately implies that $\hat{\mathcal{A}}$ is multiplicative, so that $\mathcal{R}(K)$ as defined above is indeed a character for every $K \in \mathcal{K}_\infty^-$.

Definition 2.21 We call the renormalisation procedure defined by $\mathcal{R}(K) = \Pi_-^K \hat{\mathcal{A}}$ the "BPHZ renormalisation".

It follows from (29) that in the above example the twisted antipode satisfies



so that



which makes it straightforward to verify that indeed $\Pi^K_{\text{BPHZ}}\Gamma = 0$. The following general statement should make it clear that this is indeed the "correct" way of renormalising Feynman diagrams.

Proposition 2.22 *The BPHZ renormalisation is characterised by the fact that, for every $k \geq 1$ and every connected Feynman diagram Γ with k legs and $\deg\Gamma \leq 0$, there exists a constant C such that if φ is a test function on $\mathbf{S}^k$ of the form $\varphi = \varphi_0 \cdot \varphi_1$ such that φ_1 depends only on $x_1 + \ldots + x_k$, φ_0 depends only on the differences of the x_i, and there exists a polynomial P with $\deg P + \deg\Gamma \leq 0$ and*

$$\varphi_0(x_1, \ldots, x_k) = P(x_2 - x_1, \ldots, x_k - x_1) \,, \qquad |x| \leq C \,, \tag{30}$$

then $\big(\Pi^K_{\text{BPHZ}}\Gamma\big)(\varphi) = 0$.

Remark 2.23 One way to interpret this statement is that, once we have defined $\Pi^K_{\text{BPHZ}}\Gamma$ for test functions in $\mathcal{D}_k^{(\mathcal{A})}$ with $\mathcal{A} = \{\{1, \ldots, k\}\}$, the canonical way of extending it to all test functions is to subtract from it the linear combination of

derivatives of delta functions which has precisely the same effect when testing it against all polynomials of degree at most $-\deg\Gamma$.

Proof The statement follows more or less immediately from the following observation. Take a valuation of the form Π_g^K as in (22) for some $K \in \mathcal{K}_\infty^-$ and some $g \in \mathcal{G}_-$. Fixing the Feynman diagram Γ from the statement, we write $\partial\Gamma = \{[1], \dots, [k]\}$ for its k legs, and we fix a function $\mathfrak{n}\colon \partial\Gamma \to \mathbf{N}^d$ with $|\mathfrak{n}| + \deg\Gamma \le 0$. Write furthermore $\bar{\mathfrak{n}}\colon \partial\Gamma \to \mathbf{N}^d$ for the function such that the ℓth leg has label $\delta^{(\bar{\mathfrak{n}}([\ell]))}$. We assume without loss of generality that $\bar{\mathfrak{n}}([1]) = 0$ since we can always reduce ourselves to this case by (12).

Let then P be given by

$$P(x) = P_{\mathfrak{n}}(x) = \prod_{i=[2]}^{[k]} (x_i - x_{[1]})^{\mathfrak{n}(i)} ,$$

let φ_0 be as in (30), and let φ_1 be a test function depending only on the sums of the coordinates and integrating to 1. We then claim that, writing $\bar\Gamma \subset \Gamma$ for the maximal subgraph where we only discarded the legs and $v_\star$ for the vertex of Γ incident to the first leg, one has

$$\big(\Pi_g^K \Gamma\big)(\varphi) = \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} (g \otimes \Pi_-^K)\Delta^-(\bar\Gamma, v_\star, \pi(\mathfrak{n} - \bar{\mathfrak{n}})) .$$

(In particular $\big(\Pi_g^K \Gamma\big)(\varphi) = 0$ unless $\bar{\mathfrak{n}} \le \mathfrak{n}$.) Indeed, comparing (5)–(15), it is clear that this is the case when $g = \mathbf{1}^*$, noting that

$$D_2^{\bar{\mathfrak{n}}([2])} \cdots D_k^{\bar{\mathfrak{n}}([k])} P_{\mathfrak{n}} = \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} P_{\mathfrak{n}-\bar{\mathfrak{n}}} . \tag{31}$$

The general case then follows by comparing the definitions of Δ and Δ^-, noting that by (31) the effect of the label $\bar{\mathfrak{n}}$ in (24) is exactly the same of that of the components of ℓ supported on the "legs" in (19). In other words, when comparing the two expressions one should set $\bar\ell(h) = \ell(h)$ for the half-edges h that are not legs and $\bar{\mathfrak{n}}(v) = \sum \ell(e, v)$, where the sum runs over all legs (if any) adjacent to v.

The claim now follows immediately from the definition of the twisted antipode and the BPHZ renormalisation:

$$\begin{aligned}\big(\Pi_{\text{BPHZ}}^K \Gamma\big)(\varphi) &= \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} (\Pi_-^K \hat{\mathcal{A}} \otimes \Pi_-^K)\Delta^-(\bar\Gamma, v_\star, \pi(\mathfrak{n} - \bar{\mathfrak{n}})) \\ &= \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} \Pi_-^K \mathcal{M}(\hat{\mathcal{A}} \otimes \mathrm{id})\Delta^-(\bar\Gamma, v_\star, \pi(\mathfrak{n} - \bar{\mathfrak{n}})) = 0 ,\end{aligned}$$

since the degrees of Γ and of $(\bar\Gamma, v_\star, \pi(\mathfrak{n}-\bar{\mathfrak{n}}))$ agree (and are negative) by definition. □

3 Statement and Proof of the Main Theorem

We now have all the definitions in place in order to be able to state the BPHZ theorem.

Theorem 3.1 *The valuation Π^K_{BPHZ} is consistent for K and extends continuously to all $K \in \mathcal{K}_0^-$.*

By Proposition 2.16, we only need to show the continuity part of the statement. Before we turn to the proof, we give an explicit formula for the valuation Π^K_{BPHZ} instead of the implicit characterisation given by (28). This is nothing but Zimmermann's celebrated "forest formula".

3.1 Zimmermann's Forest Formula

So what are these "forests" appearing in the eponymous formula? Given any Feynman diagram Γ, the set $\mathcal{G}_\Gamma^-$ of all *connected* vacuum diagrams $\bar\Gamma \subset \Gamma$ with $\deg \bar\Gamma \le 0$ is endowed with a natural partial order given by inclusion. A subset $\mathcal{F} \subset \mathcal{G}_\Gamma^-$ is called a "forest" if any two elements of $\mathcal{F}$ are either comparable in $\mathcal{G}_\Gamma^-$ or vertex-disjoint as subgraphs of Γ.

Given a forest $\mathcal{F}$ and a subgraph $\bar\Gamma \in \mathcal{F}$, we say that $\bar\Gamma_1$ is a *child* of $\bar\Gamma$ if $\bar\Gamma_1 < \bar\Gamma$ and there exists no $\bar\Gamma_2 \in \mathcal{F}$ with $\bar\Gamma_1 < \bar\Gamma_2 < \bar\Gamma$. Conversely, we then say that $\bar\Gamma$ is $\bar\Gamma_1$'s parent. (The forest structure of $\mathcal{F}$ guarantees that its elements have at most one parent.) An element without children is called a *leaf* and one without parent a *root*. If we connect parents to their children in $\mathcal{F}$, then it does indeed form a forest with arrows pointing away from the roots and towards the leaves. We henceforth write $\mathcal{F}_\Gamma^-$ for the set of all forests for Γ.

Given a diagram Γ, we now consider the space $\mathcal{T}_\Gamma$ generated by all diagrams $\hat\Gamma$ such that each connected component has *either* at least one leg *or* a distinguished vertex $v_\star$, but not both. We furthermore endow $\hat\Gamma$ with an $\mathbf{N}^d$-valued vertex decoration $\mathfrak{n}$ supported on the leg-less components and, most importantly, with a bijection $\tau : \hat{\mathcal{E}} \to \mathcal{E}$ between the edges of $\hat\Gamma$ and those of Γ, such that legs get mapped to legs. The operation of discarding τ yields a natural injection $\mathcal{T}_\Gamma \hookrightarrow \hat{\mathcal{T}}_- \otimes \mathcal{T}$ by keeping the components with a distinguished vertex in the first factor and those with legs in the second factor. (The space $\mathcal{T}_\Gamma$ itself however is not a tensor product due to the constraint that τ is a bijection, which exchanges information between the two factors.) We can also define $\partial\mathcal{T}_\Gamma$ analogously to (12) and (16), (17), and (18), so that $\mathcal{H}_\Gamma = \mathcal{T}_\Gamma / \partial\mathcal{T}_\Gamma$ naturally injects into $\hat{\mathcal{H}}_- \otimes \mathcal{H}$.

Given a connected subgraph $\gamma \subset \Gamma$, we then define a contraction operator $\mathcal{C}_\gamma$ acting on $\mathcal{H}_\Gamma$ in the following way. Given an element $(\hat\Gamma, \mathfrak{n}) \in \mathcal{T}_\Gamma$, we write $\hat\gamma$ for the subgraph of $\hat\Gamma$ such that τ is a bijection between the edges of $\hat\gamma$ and those of γ.

If $\hat\gamma$ is not connected, then we set $\mathscr{C}_\gamma(\hat\Gamma, \mathfrak{n}) = 0$. Otherwise, we set as in (24)

$$\mathscr{C}_\gamma(\hat\Gamma, \mathfrak{n}) = \sum_{\substack{\bar\ell\colon \partial\gamma \to \mathbf{N}^d \\ \bar{\mathfrak{n}}\colon \mathscr{V}_\gamma \to \mathbf{N}^d}} \frac{(-1)^{|\operatorname{out}\bar\ell|}}{\bar\ell!} \mathbf{1}_{\deg(\hat\gamma, \bar{\mathfrak{n}} + \pi\bar\ell) \le 0} \binom{\mathfrak{n}}{\bar{\mathfrak{n}}} (\hat\gamma, \bar{\mathfrak{n}} + \pi\bar\ell) \cdot (\hat\Gamma, \mathfrak{n} - \bar{\mathfrak{n}}) / (\hat\gamma, \bar\ell)\ , \tag{32}$$

with the obvious bijections between the edges of $\hat\gamma \cdot \hat\Gamma / \hat\gamma$ and those of Γ. This time we explicitly include the restriction to terms such that $\deg(\hat\gamma, \bar{\mathfrak{n}} + \pi\bar\ell) \le 0$, which replaces the projection to $\mathscr{H}_-$ in (24). An important fact is then the following.

Lemma 3.2 *Let γ_1, γ_2 be two subgraphs of Γ that are vertex-disjoint and let $\hat\Gamma \in \mathscr{T}_\Gamma$ be such that $\hat\gamma_1$ and $\hat\gamma_2$ are vertex disjoint. Then $\mathscr{C}_{\gamma_1}\mathscr{C}_{\gamma_2}\hat\Gamma = \mathscr{C}_{\gamma_2}\mathscr{C}_{\gamma_1}\hat\Gamma$.* □

We will use the natural convention that $\emptyset \in \mathscr{F}_\Gamma^-$. For any $\mathscr{F} \in \mathscr{F}_\Gamma^-$, we then write $\mathscr{C}_\mathscr{F}\Gamma$ for the element of $\mathscr{H}_\Gamma$ defined recursively in the following way. For $\mathscr{F} = \emptyset$, we set $\mathscr{C}_\emptyset\Gamma = \Gamma$. For non-empty $\mathscr{F}$, we write $\varrho(\mathscr{F}) \subset \mathscr{F}$ for the set of roots of $\mathscr{F}$ and we set recursively

$$\mathscr{C}_\mathscr{F}\Gamma = \mathscr{C}_{\mathscr{F}\setminus\varrho(\mathscr{F})} \prod_{\gamma\in\varrho(\mathscr{F})} \mathscr{C}_\gamma\Gamma\ .$$

The order of the product doesn't matter by Lemma 3.2, since the roots of $\mathscr{F}$ are all vertex-disjoint. With these notations at hand, Zimmermann's forest formula [21] then reads

Proposition 3.3 *The BPHZ renormalisation procedure is given by the identity*

$$(\hat{\mathscr{A}} \otimes \mathrm{id})\Delta\Gamma = \mathscr{R}\Gamma \overset{def}{=} \sum_{\mathscr{F}\in\mathscr{F}_\Gamma^-} (-1)^{|\mathscr{F}|} \mathscr{C}_\mathscr{F}\Gamma\ , \tag{33}$$

where we implicitly use the injection $\mathscr{H}_\Gamma \hookrightarrow \hat{\mathscr{H}}_- \otimes \mathscr{H}$ for the right hand side.

Proof This follows from the representation (29). Another way of seeing it is to first note that $\mathscr{R}$ is indeed of the form $(\mathscr{B} \otimes \mathrm{id})\Delta\Gamma$ for *some* $\mathscr{B}\colon \mathscr{H}_- \to \hat{\mathscr{H}}_-$ and to then make use of the characterisation (28) of the twisted antipode $\hat{\mathscr{A}}$. This implies that it suffices to show that $\mathscr{R}\Gamma = 0$ for every connected Γ with a distinguished vertex and a node-labelling such that $\deg\Gamma \le 0$.

The idea is to observe that $\mathscr{F}_\Gamma^-$ can be partitioned into two disjoint sets that are in bijection with each other: those that contain Γ itself and the complement $\hat{\mathscr{F}}_\Gamma^-$ of those forest that don't. Furthermore, it follows from the definition that $\mathscr{C}_\Gamma\Gamma = \Gamma$, so that

$$\sum_{\mathscr{F}\in\mathscr{F}_\Gamma^-} (-1)^{|\mathscr{F}|}\mathscr{C}_\mathscr{F}\Gamma = \sum_{\mathscr{F}\in\hat{\mathscr{F}}_\Gamma^-} (-1)^{|\mathscr{F}|}\big(\mathscr{C}_\mathscr{F}\Gamma - \mathscr{C}_{\mathscr{F}\cup\{\Gamma\}}\Gamma\big) = \sum_{\mathscr{F}\in\hat{\mathscr{F}}_\Gamma^-} (-1)^{|\mathscr{F}|}\big(\mathscr{C}_\mathscr{F}\Gamma - \mathscr{C}_\mathscr{F}\Gamma\big)\ ,$$

which vanishes thus completing the proof. □

In order to analyse (33), it will be very convenient to have ways of resumming its terms in order to make cancellations more explicit. These resummations are based on the following trivial identity. Given a finite set A and operators X_i with $i \in A$, one has

$$\prod_{i\in A}(\mathrm{id} - X_i) = \sum_{B\subset A}(-1)^B \prod_{j\in B} X_j \;, \tag{34}$$

provided that the order in which the operators are composed is the same in each term and that the empty product is interpreted as the identity. The right hand side of this expression is clearly reminiscent of (33) while the left hand side encodes cancellations if the X_i are close to the identity in some sense. If $\mathcal{G}_\Gamma^-$ itself happens to be a forest, then $\mathcal{F}_\Gamma^-$ consists simply of all subsets of $\mathcal{G}_\Gamma^-$, so that one can indeed write

$$(\hat{\mathcal{A}} \otimes \mathrm{id})\Delta\Gamma = \mathcal{R}_{\mathcal{G}_\Gamma^-}\Gamma \;, \tag{35}$$

where $\mathcal{R}_\mathcal{F}\Gamma$ is defined by $\mathcal{R}_\emptyset\Gamma = \Gamma$ and then via the recursion

$$\mathcal{R}_\mathcal{F}\Gamma = \mathcal{R}_{\mathcal{F}\setminus\varrho(\mathcal{F})}\prod_{\gamma\in\varrho(\mathcal{F})}(\mathrm{id} - \mathcal{C}_\gamma)\Gamma \;. \tag{36}$$

In general however this is not the case, and this is precisely the problem of "overlapping divergences". In order to deal with this, we introduce the following variant of (35) which still works in the general case. To formulate it, we introduce the notion of a "forest interval" $\mathbb{M}$ for Γ which is a subset of $\mathcal{F}_\Gamma^-$ of the form $[\underline{\mathbb{M}}, \overline{\mathbb{M}}]$ in the sense that it consists precisely of all those forests $\mathcal{F} \in \mathcal{F}_\Gamma^-$ such that $\underline{\mathbb{M}} \subset \mathcal{F} \subset \overline{\mathbb{M}}$. An alternative description of $\mathbb{M}$ is that there is a forest $\delta(\mathbb{M}) = \overline{\mathbb{M}}\setminus\underline{\mathbb{M}}$ disjoint from $\underline{\mathbb{M}}$ and such that $\mathbb{M}$ consists of all forests of the type $\underline{\mathbb{M}} \cup \mathcal{F}$ with $\mathcal{F} \subset \delta(\mathbb{M})$. Given a forest interval, we define an operation $\mathcal{R}_\mathbb{M}$ which renormalises all subgraphs in $\delta(\mathbb{M})$ and contracts those subgraphs in $\underline{\mathbb{M}}$. In other words, we set $\mathcal{R}_\mathbb{M} = \mathcal{R}_\mathbb{M}^{\overline{\mathbb{M}}}$, where $\mathcal{R}_\mathbb{M}^\mathcal{F}$ is defined recursively by

$$\mathcal{R}_\mathbb{M}^\mathcal{F}\Gamma = \mathcal{R}_\mathbb{M}^{\mathcal{F}\setminus\varrho(\mathcal{F})}\prod_{\gamma\in\varrho(\mathcal{F})}\mathcal{C}_\gamma^\sharp\Gamma \;, \qquad \mathcal{C}_\gamma^\sharp = \begin{cases} \mathrm{id} - \mathcal{C}_\gamma & \text{if } \gamma \in \delta(\mathbb{M}), \\ -\mathcal{C}_\gamma & \text{otherwise.} \end{cases}$$

This definition is consistent with (36) in the sense that one has $\mathcal{R}_\mathcal{F} = \mathcal{R}_\mathbb{M}$ for $\mathbb{M} = [\emptyset, \mathcal{F}]$. Combining Proposition 3.3 with (34), we then obtain the following alternative characterisation of our renormalisation map.

Lemma 3.4 *Let Γ be a Feynman diagram and let $\mathcal{P}$ be a partition of $\mathcal{F}_\Gamma^-$ consisting of forest intervals. Then, one has the identity $(\hat{\mathcal{A}} \otimes \mathrm{id})\Delta\Gamma = \sum_{\mathbb{M}\in\mathcal{P}}\mathcal{R}_\mathbb{M}\Gamma$.* □

3.2 *Proof of the BPHZ Theorem, Theorem 3.1*

We now have all the ingredients in place to prove Theorem 3.1. We only need to show that for every (connected) Feynman diagram Γ there are constants C_Γ and N_Γ such that for every test function φ with compact support in the ball of radius 1 one has the bound

$$\big|\big(\Pi^K_{\text{BPHZ}}\Gamma\big)(\varphi)\big| \le C_\Gamma \prod_{e\in\mathscr{E}} |K_{\mathfrak{t}(e)}|_{N_\Gamma} \sup_{|k|\le N_\Gamma} \|D^{(k)}\varphi\|_{L^\infty} \,, \tag{37}$$

where $|K_{\mathfrak{t}}|_N$ denotes the smallest constant C such that (2) holds for all $|k| \le N$.

The proof of (37) follows the same lines as that of the main result in [3], but with a number of considerable simplifications:

- There is no "positive renormalisation" in the present context so that we do not need to worry about overlaps between positive and negative renormalisations. As a consequence, we also do not make any claim on the behaviour of (37) when rescaling the test function. In general, it is *false* that (37) obeys the naive power-counting when φ is replaced by φ^λ and $\lambda \to 0$ as in [16, Lem. A.7].
- The BPHZ renormalisation procedure studied in the present article is directly formulated at the level of graphs. In [2, 3] on the other hand, it is formulated at the level of trees (which are the objects indexing a suitable family of stochastic processes) and then has to be translated into a renormalisation procedure on graphs which, depending on how trees are glued together in order to form these graphs, creates additional "useless" terms.
- We only consider kernels with a single argument, corresponding to "normal" edges in our graphs, while [3] deals with non-Gaussian processes which then gives rise to Feynman diagrams containing some "multiedges".

We therefore only give an overview of the main steps, but we hope that the style of our exposition is such that the interested reader will find it possible to fill in the missing details without undue effort.

As in the proof of Proposition 2.4, we break the domain of integration into Hepp sectors $D_{\mathbf{T}}$ and we estimate terms separately on each sector. The main trick is then to resum the terms as in Lemma 3.4, but by using a partition $\mathscr{P}_{\mathbf{T}}$ that is adapted to the Hepp sector $\mathbf{T}$ in such a way that the occurrences of $(\mathrm{id} - \mathscr{C}_\gamma)$ create cancellations that are useful on $D_{\mathbf{T}}$.

In order to formulate this, it is convenient to write all the terms appearing in the definition of $\Pi^K_{\text{BPHZ}}\Gamma$ as integrals over the same set of variables. For this, we henceforth fix a connected Feynman diagram Γ once and for all, together with an arbitrary total order for its vertices.

We then define the space $\hat{\mathcal{T}}_\Gamma$ generated by connected Feynman diagrams $\bar\Gamma$ with edges *and vertices* in bijection with those of Γ via a map $\tau\colon (\bar{\mathscr{E}}, \bar{\mathscr{V}}) \to (\mathscr{E}, \mathscr{V})$, together with a vertex labelling $\mathfrak{n}$, as well as a map $\mathfrak{d}\colon \bar{\mathscr{E}} \to \mathbf{N}$ which vanishes on all legs of $\bar\Gamma$. The goal of this map is to allow us to keep track on which parts of Γ were

contracted, as well as the structure of nested contractions: $\mathfrak{d}$ measures how "deep" a given edge lies within nested contractions. In particular, it is natural to impose that $\mathfrak{d}$ vanishes on legs since they are never contracted. We furthermore impose that for every $j > 0$, every connected component $\hat{\gamma}$ of $\mathfrak{d}^{-1}(j)$ has the following two properties.

- The highest vertex $v_\star(\hat{\gamma})$ of $\hat{\gamma}$ has an incident edge e with $\mathfrak{d}(e) < j$. (Here, "highest" refers to the total order we fixed on vertices of Γ, which is transported to $\bar{\Gamma}$ by the bijection between vertices of Γ and $\bar{\Gamma}$.)
- All edges e incident to a vertex of $\hat{\gamma}$ other than $v_\star(\hat{\gamma})$ satisfy $\mathfrak{d}(e) \geq j$.

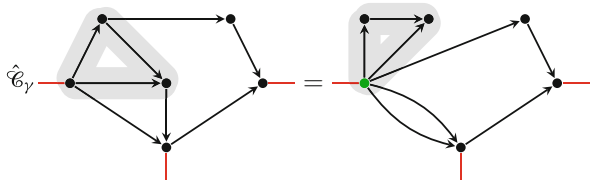
Writing $\bar{\mathcal{V}}^c \subset \bar{\mathcal{V}}$ for those vertices v with at least one edge e incident to v such that $\mathfrak{d}(e) > 0$, we also impose that $\mathfrak{n}(v) = 0$ for $v \notin \bar{\mathcal{V}}^c$. We view Γ itself as an element of $\hat{\mathcal{T}}_\Gamma$ by setting $\mathfrak{d} \equiv 0$. Note that this data defines a map $v \mapsto v_\star$ from $\bar{\mathcal{V}}^c$ to $\bar{\mathcal{V}}^c$ such that $v \mapsto v_\star(\hat{\gamma})$ for $\hat{\gamma}$ the connected component of $\mathfrak{d}^{-1}(j)$ with the lowest possible value of j containing v.

For $\gamma \subset \Gamma$ as above, we then define maps $\hat{\mathcal{C}}_\gamma$ on $\hat{\mathcal{T}}_\Gamma$ similarly to (32). This time however, we set $\hat{\mathcal{C}}_\gamma \bar{\Gamma} = 0$ unless the following conditions are met.

- The graph $\tau^{-1}(\gamma) \subset \bar{\Gamma}$ is connected.
- For every edge e adjacent to $\tau^{-1}(\gamma)$, one has $\mathfrak{d}(e) \leq \inf_{\hat{e} \in \hat{\mathcal{E}}} \mathfrak{d}(\hat{e})$.

We also restrict the sum over labels ℓ supported on edges with $\mathfrak{d}(e) = \inf_{\hat{e} \in \hat{\mathcal{E}}} \mathfrak{d}(\hat{e})$. In order to remain in $\hat{\mathcal{T}}_\Gamma$, instead of extracting $\hat{\gamma} = \tau^{-1}(\gamma)$, we reconnect the edges of $\bar{\Gamma}$ adjacent to $\hat{\gamma}$ to the highest vertex $\hat{v}$ of $\hat{\gamma}$ and we increase $\mathfrak{d}(e)$ by 1 on all edges e of $\hat{\gamma}$. We similarly define elements $\hat{\mathcal{R}}_{\mathbb{M}}\Gamma$ as above with every instance of $\mathcal{C}_\gamma$ replaced by $\hat{\mathcal{C}}_\gamma$. We also view Γ itself as an element of $\hat{\mathcal{T}}_\Gamma$ by setting both $\mathfrak{d}$ and $\mathfrak{n}$ to 0.

Let us illustrate this by taking for Γ the diagram of Fig. 2 and for γ the triangle shaded in grey. In this case, assuming that the order on our vertices is such that the first vertex is the leftmost one and that the degree of γ is above -1 so that no node-decorations are needed, we have



with $\mathfrak{d}(v)$ equal to 1 in the shaded region of the diagram on the right. The green node then denotes the element $v_\star$ for all the nodes v in that region. This time, it follows in virtually the same way as the proof of Proposition 2.19 that if γ_1 and γ_2 are either vertex disjoint or such that one is included in the other, then the operators $\hat{\mathcal{C}}_{\gamma_1}$ and

$\hat{\mathcal{C}}_{\gamma_2}$ commute. In particular, we can simply write

$$\hat{\mathcal{R}}_{\mathbb{M}}\Gamma = \Big(\prod_{\gamma \in \delta(\mathbb{M})} (\mathrm{id} - \hat{\mathcal{C}}_\gamma) \prod_{\bar{\gamma} \in \underline{\mathbb{M}}} (-\hat{\mathcal{C}}_{\bar{\gamma}}) \Big) \Gamma \,, \tag{38}$$

without having to worry about the order of the operations as in (36).

For every $K \in \mathcal{K}_\infty^-$ and every test function φ, we then have a linear map $\mathcal{W}^K : \hat{\mathcal{T}}_\Gamma \to \mathcal{C}^\infty(\mathbf{S}^{\mathcal{V}_\star})$ given by

$$\begin{aligned} \big(\mathcal{W}^K \bar{\Gamma}\big)(x) = \prod_{e \in \bar{\mathcal{E}}_\star} K_{\mathfrak{t}(e)}(x_{\tau(e_+)} - x_{\tau(e_-)}) \prod_{v \in \bar{\mathcal{V}}_\star} (x_{\tau(v)} - x_{\tau(v_\star)})^{\mathbf{n}(v)} \\ \times \big(D_1^{\ell_1} \cdots D_k^{\ell_k} \varphi\big)(x_{v_1}, \dots, x_{v_k}) \,, \end{aligned}$$

where $\tau : \bar{\mathcal{V}} \cup \bar{\mathcal{E}} \to \mathcal{V} \cup \mathcal{E}$ is the bijection between edges and vertices of $\bar{\Gamma}$ and those of Γ, v_i are the vertices to which the k legs of Γ are attached, and ℓ_i are the corresponding multiindices as in (5). With this notation, our definitions show that, for every partition $\mathcal{P}$ of $\mathcal{F}_\Gamma^-$ into forest intervals, one has

$$\big(\Pi^K_{\text{BPHZ}}\Gamma\big)(\varphi) = \sum_{\mathbb{M} \in \mathcal{P}} \int_{\mathbf{S}^{\mathcal{V}_\star}} \big(\mathcal{W}^K \hat{\mathcal{R}}_{\mathbb{M}}\Gamma\big)(x)\, dx \,.$$

We bound this rather brutally by

$$\begin{aligned} \big|\big(\Pi^K_{\text{BPHZ}}\Gamma\big)(\varphi)\big| &\le \sum_{\mathbf{T}} \sum_{\mathbb{M} \in \mathcal{P}_{\mathbf{T}}} \int_{D_{\mathbf{T}}} \big|\big(\mathcal{W}^K \hat{\mathcal{R}}_{\mathbb{M}}\Gamma\big)(x)\big|\, dx \\ &\le \sum_{\mathbf{T}} \sum_{\mathbb{M} \in \mathcal{P}_{\mathbf{T}}} \sup_{x \in D_{\mathbf{T}}} \big|\big(\mathcal{W}^K \hat{\mathcal{R}}_{\mathbb{M}}\Gamma\big)(x)\big| \prod_{u \in T} 2^{-d\mathbf{n}_u} \,. \end{aligned} \tag{39}$$

At this stage, we would like to make a smart choice for the partition $\mathcal{P}_{\mathbf{T}}$ which allows us to obtain a summable bound for this expression. In order to do this, we would like to guarantee that a cancellation $(\mathrm{id} - \hat{\mathcal{C}}_\gamma)$ appears for all of the subgraphs γ that are such that the length of all adjacent edges (as measured by the quantity $|x_{\tau(e_+)} - x_{\tau(e_-)}|$) is much greater than the diameter of γ (measured in the same way).

In order to achieve this, we first note that by Proposition 3.11 and (56) below, we can restrict ourselves in (39) to the case where $\mathcal{P}_{\mathbf{T}}$ is a partition of the subset $\hat{\mathcal{F}}_\Gamma^- \subset \mathcal{F}_\Gamma^-$ of all forests containing only subgraphs that are full in Γ. (Recall that a subgraph $\bar{\gamma} \subset \Gamma$ is full in Γ if it is induced by a subset of the vertices of $\bar{\Gamma}$ in the sense that it consists of all edges of Γ connecting two vertices of the subset in question.) We then consider the following construction. For any forest $\mathcal{F} \in \hat{\mathcal{F}}_\Gamma^-$, write $\mathfrak{K}_{\mathcal{F}}\Gamma$ for the Feynman diagram obtained by performing the contractions of $\hat{\mathcal{C}}_{\mathcal{F}}\Gamma$. (So that $\hat{\mathcal{C}}_{\mathcal{F}}\Gamma$ is a linear combination of terms obtained from $\mathfrak{K}_{\mathcal{F}}\Gamma$ by adding

node-labels $\mathfrak{n}$ and the corresponding derivatives on incident edges.) As above, write τ for the corresponding bijection between edges and vertices of $\mathfrak{K}_{\mathcal{F}}\Gamma$ and those of Γ. Given a Hepp sector $\mathbf{T} = (T, \mathbf{n})$ for Γ and an edge e of Γ, we then write $\mathrm{scale}^{\mathcal{F}}_{\mathbf{T}}(e) = \mathbf{n}(v_e)$, where $v_e = \tau(\tau^{-1}(e)_-) \wedge \tau(\tau^{-1}(e)_+)$ is the common ancestor in T of the two vertices incident to e, but when viewed as an edge of $\mathfrak{K}_{\mathcal{F}}\Gamma$. (Since we only consider forests consisting of full subgraphs, $\tau^{-1}(e)_-$ and $\tau^{-1}(e)_+$ are distinct, so this is well-defined.) Given $\gamma \in \mathcal{F}$, we then set

$$\mathrm{int}^{\mathcal{F}}_{\mathbf{T}}(\gamma) = \inf_{e \in \mathcal{E}^{\mathcal{F}}_{\gamma}} \mathrm{scale}^{\mathcal{F}}_{\mathbf{T}}(e)\,, \qquad \mathrm{ext}^{\mathcal{F}}_{\mathbf{T}}(\gamma) = \sup_{e \in \partial\mathcal{E}^{\mathcal{F}}_{\gamma}} \mathrm{scale}^{\mathcal{F}}_{\mathbf{T}}(e)\,,$$

where $\mathcal{E}^{\mathcal{F}}_{\gamma}$ denotes the edges belonging to γ, but *not* to any of the children of γ in $\mathcal{F}$, while $\partial\mathcal{E}^{\mathcal{F}}_{\gamma}$ denotes the edges adjacent to γ and belonging to the parent $\mathscr{A}(\gamma)$ of γ in $\mathcal{F}$ (with the convention that if γ has no parent, then $\mathscr{A}(\gamma) = \Gamma$). With these notations, we then make the following definition.

Definition 3.5 Fix a Hepp sector $\mathbf{T}$. Given a forest $\mathcal{F} \in \hat{\mathcal{F}}^-_{\Gamma}$, we say that $\gamma \in \mathcal{F}$ is *safe in* $\mathcal{F}$ if $\mathrm{ext}^{\mathcal{F}}_{\mathbf{T}}(\gamma) \geq \mathrm{int}^{\mathcal{F}}_{\mathbf{T}}(\gamma)$ and that it is *unsafe in* $\mathcal{F}$ otherwise. Given a forest $\mathcal{F}$ and a subgraph $\gamma \in \mathcal{G}^-_{\Gamma}$, we say that γ is *safe/unsafe for* $\mathcal{F}$ if $\mathcal{F} \cup \{\gamma\} \in \hat{\mathcal{F}}^-_{\Gamma}$ and γ is safe/unsafe in $\mathcal{F} \cup \{\gamma\}$. Finally, we say that a forest $\mathcal{F}$ is safe if every $\gamma \in \mathcal{F}$ is safe in $\mathcal{F}$.

The following remark is then crucial.

Lemma 3.6 *Let $\mathcal{F}_s \in \hat{\mathcal{F}}^-_{\Gamma}$ be a safe forest and write $\mathcal{F}_u$ for the collection of all $\gamma \in \mathcal{G}^-_{\Gamma}$ that are unsafe for $\mathcal{F}_s$. Then, one has $\mathcal{F}_s \cup \mathcal{F}_u \in \hat{\mathcal{F}}^-_{\Gamma}$ and furthermore every γ in $\mathcal{F}_s/\mathcal{F}_u$ is safe/unsafe in $\mathcal{F}_s \cup \mathcal{F}_u$.*

Proof Fix $\mathcal{F}_s$ and write again τ for the corresponding bijection between edges and vertices of $\mathfrak{K}_{\mathcal{F}_s}\Gamma$ and those of Γ. For each $\gamma \in \mathcal{F}_s$, write $\mathcal{V}^{\mathcal{F}_s}_{\gamma} \subset \mathcal{V}$ for the set of vertices of the form $\tau(\tau^{-1}(e)_{\pm})$ for $e \in \mathcal{E}^{\mathcal{F}_s}_{\gamma}$, as well as $v^{\mathcal{F}_s}_{\star,\gamma} \in \mathcal{V}^{\mathcal{F}_s}_{\gamma}$ for the highest one of these vertices. (This is the vertex that edges outside of γ were reconnected to by the operation $\mathfrak{K}_{\mathcal{F}_s}$.) We also write $\partial\mathcal{V}^{\mathcal{F}_s}_{\gamma} \subset \mathcal{V}$ for all vertices of the form $\tau(\tau^{-1}(e)_{\pm})$ for $e \in \partial\mathcal{E}^{\mathcal{F}_s}_{\gamma}$ that are *not* in $\mathcal{V}^{\mathcal{F}_s}_{\gamma}$.

With this notation, $\mathrm{int}^{\mathcal{F}}_{\mathbf{T}}(\gamma) = \mathbf{n}((\mathcal{V}^{\mathcal{F}_s}_{\gamma})^{\uparrow})$ and there exists a vertex $w \in \partial\mathcal{V}^{\mathcal{F}_s}_{\gamma}$ for $\mathscr{A}(\gamma)$ the parent of γ in $\mathcal{F}_s$ (with the convention as above) such that $\mathrm{ext}^{\mathcal{F}}_{\mathbf{T}}(\gamma) = \mathbf{n}(v^{\mathcal{F}_s}_{\star,\gamma} \wedge w)$. Since both $(\mathcal{V}^{\mathcal{F}_s}_{\gamma})^{\uparrow}$ and $v^{\mathcal{F}_s}_{\star,\gamma} \wedge w$ lie on the path connecting the root of T to $v_{\star,\gamma}$, it follows from the definition of a safe forest that one necessarily has $v^{\mathcal{F}_s}_{\star,\gamma} \wedge w > (\mathcal{V}^{\mathcal{F}_s}_{\gamma})^{\uparrow}$.

Let now $\bar\gamma \in \mathcal{G}^-_{\Gamma} \setminus \mathcal{F}_s$ be such that $\mathcal{F}_s \cup \{\bar\gamma\} \in \hat{\mathcal{F}}^-_{\Gamma}$ and set $\mathcal{V}_{\bar\gamma} = \mathcal{V}^{\mathcal{F}_s \cup \{\bar\gamma\}}_{\bar\gamma}$ as well as $\partial\mathcal{V}_{\bar\gamma} = \partial\mathcal{V}^{\mathcal{F}_s \cup \{\bar\gamma\}}_{\bar\gamma}$. It follows from the definitions that $\bar\gamma \in \mathcal{F}_u$ if and only if none of the descendants of $\mathcal{V}^{\uparrow}_{\bar\gamma}$ in T belongs to $\partial\mathcal{V}_{\bar\gamma}$. As a consequence of this characterisation, any two graphs $\gamma_1, \gamma_2 \in \mathcal{F}_u$ are either vertex-disjoint, or one of them is included in the other one. Indeed, assume by contradiction that neither is

included in the other one and that their intersection $\gamma_\cap$ contains at least one vertex. Writing $\hat\gamma_\cap$ for one of the connected components of $\gamma_\cap$, there exist edges e_i in γ_i that are adjacent to $\hat\gamma_\cap$: otherwise, since the γ_i are connected, one of them would be contained in $\hat\gamma_\cap$. Write v_i for the vertex of e_i that does not belong to $\hat\gamma_\cap$. Such a vertex exists since otherwise it would not be the case that $\hat\gamma_\cap$ is full in $\gamma^\uparrow = \mathscr{A}(\gamma_1) = \mathscr{A}(\gamma_2)$. Since γ_1 is unsafe, it follows that v_2 is not a descendent of $(\mathscr{V}_{\hat\gamma_\cap} \cup \{v_1\})^\uparrow$, so that in particular, for every vertex $v \in \hat\gamma_\cap$, one has $v_1 \wedge v > v_2 \wedge v$. The same argument with the roles of γ_1 and γ_2 reversed then leads to a contradiction.

This shows that $\mathscr{F}_s \cup \mathscr{F}_u$ is indeed again a forest so that it remains to show the last statement. We will show a slightly stronger statement namely that, given an arbitrary forest $\mathscr{F}$, the property of $\gamma \in \mathscr{F}$ being safe or unsafe does not change under the operation of adding to $\mathscr{F}$ a graph $\bar\gamma$ that is unsafe for $\mathscr{F}$. Given the definitions, there are three potential cases that could affect the "safety" of γ: either $\bar\gamma \subset \gamma$, or $\gamma \subset \bar\gamma$, or $\bar\gamma \subset \mathscr{A}(\gamma)$ and there exists an edge e adjacent to both γ and $\bar\gamma$. We consider these three cases separately and we write $\bar{\mathscr{F}} = \mathscr{F} \cup \{\bar\gamma\}$.

In the case $\bar\gamma \subset \gamma$, it follows from the ultrametric property and the fact that $\bar\gamma$ is unsafe that $\mathrm{int}_{\mathbf{T}}^{\bar{\mathscr{F}}}(\gamma) = \mathrm{int}_{\mathbf{T}}^{\mathscr{F}}(\gamma)$ whence the desired property follows. In the case $\gamma \subset \bar\gamma$, it is $\mathrm{ext}_{\mathbf{T}}^{\mathscr{F}}(\gamma)$ which could potentially change since $\partial\mathscr{C}_\gamma^{\mathscr{F}}$ becomes smaller when adding $\bar\gamma$. Note however that by the ultrametric property, combined with the fact that $\bar\gamma$ is unsafe, the edges e in $\partial\mathscr{C}_\gamma^{\mathscr{F}} \setminus \partial\mathscr{C}_\gamma^{\bar{\mathscr{F}}}$ satisfy $\mathrm{scale}_{\mathbf{T}}^{\mathscr{F}}(e) = \mathrm{scale}_{\mathbf{T}}^{\bar{\mathscr{F}}}(e)$. Furthermore, again as a consequence of $\bar\gamma$ being unsafe, one has $\mathrm{scale}_{\mathbf{T}}^{\bar{\mathscr{F}}}(e) < \mathrm{scale}_{\mathbf{T}}^{\bar{\mathscr{F}}}(\bar e)$ for every edge $\bar e$ in $\bar\gamma$ which is not in γ, so in particular for $\bar e \in \partial\mathscr{C}_\gamma^{\bar{\mathscr{F}}}$. This shows again that $\mathrm{ext}_{\mathbf{T}}^{\mathscr{F}}(\gamma) = \mathrm{ext}_{\mathbf{T}}^{\bar{\mathscr{F}}}(\gamma)$ as required. The last case can be dealt with in a very similar way, thus concluding the proof. □

As a corollary of the proof, we see that the definition of the notion of "safe forest" as well as the construction of $\mathscr{F}_u$ given a safe forest $\mathscr{F}_s$ only depend on the topology of the tree T and not on the specific scale assignment $\mathbf{n}$. It also follows that, given an arbitrary $\mathscr{F} \in \hat{\mathscr{F}}_\Gamma^-$, there exists a unique way of writing $\mathscr{F} = \mathscr{F}_s \cup \mathscr{F}_u$ with $\mathscr{F}_s$ a safe forest and $\mathscr{F}_u$ being unsafe for $\mathscr{F}_s$ (and equivalently for $\mathscr{F}$). In particular, writing $\mathscr{F}_\Gamma^{(s)}(T)$ for the collection of safe forests for the tree T, the collection $\mathscr{P}_{\mathbf{T}} = \{[\mathscr{F}_s, \mathscr{F}_s \cup \mathscr{F}_u] : \mathscr{F}_s \in \mathscr{F}_\Gamma^{(s)}(T)\}$ where, for any $\mathscr{F}_s$, the forest $\mathscr{F}_u$ is defined as in Lemma 3.6, forms a partition of $\hat{\mathscr{F}}_\Gamma^-$ into forest intervals. It then follows from (39) that

$$\big|\big(\Pi_{\mathrm{BPHZ}}^K \Gamma\big)(\varphi)\big| \le \sum_T \sum_{\mathscr{F}_s \in \mathscr{F}_\Gamma^{(s)}(T)} \sum_{\mathbf{n}} \sup_{x \in D_{\mathbf{T}}} \big|\big(\mathscr{W}^K \hat{\mathscr{R}}_{[\mathscr{F}_s, \mathscr{F}_s \cup \mathscr{F}_u]} \Gamma\big)(x)\big| \prod_{v \in T} 2^{-d\mathbf{n}_v} ,$$

where $\mathbf{n}$ runs over all monotone integer labels for T and the construction of $\mathscr{F}_u$ given $\mathscr{F}_s$ and T is as above. We note that the first two sums are finite, so that as in the proof of Proposition 2.4 it is sufficient, for any given choice of T and safe forest $\mathscr{F}_s$, to find a collection real-valued function $\{\eta_i\}_{i \in I}$ (for some *finite* index set I) on

the interior vertices of T such that

$$\sum_{\mathbf{n}} \sup_{x\in D_{\mathbf{T}}} \big|\big(\mathscr{W}^K \hat{\mathscr{R}}_{[\mathcal{F}_s,\mathcal{F}_s\cup\mathcal{F}_u]}\Gamma\big)(x)\big| \prod_{v\in T} 2^{-d\mathbf{n}_v} \le \sum_{i\in I}\sum_{\mathbf{n}}\prod_{v\in T} 2^{-\eta_i(v)\mathbf{n}_v}\,, \tag{40}$$

and such that

$$\sum_{w\ge v} \eta_i(w) > 0\,, \qquad \forall v\in T\,, \quad \forall i\in I\,, \tag{41}$$

which then guarantees that the above expression converges.

Before we turn to the construction of the η_i, let us examine in a bit more detail the structure of the graph $\mathfrak{K}_{\mathcal{F}}\Gamma = (\mathcal{V}_{\mathcal{F}}, \mathcal{E}_{\mathcal{F}})$. Writing τ for the bijection between $\mathfrak{K}_{\mathcal{F}}\Gamma$ and Γ, every $\gamma \in \mathcal{F}$ yields a subgraph $\mathfrak{K}(\gamma) = (\mathcal{V}_\gamma, \mathcal{E}_\gamma)$ of $\mathfrak{K}_{\mathcal{F}}\Gamma$ whose edge set is given by the preimage under τ of the edge set of $\gamma \setminus \bigcup \mathcal{C}(\gamma)$, where $\mathcal{C}(\gamma)$ denotes the set of all children of γ in $\mathcal{F}$. Furthermore, $\mathfrak{K}(\gamma)$ is connected by exactly one vertex to $\mathfrak{K}(\bar\gamma)$, for $\bar\gamma \in \mathcal{C}(\gamma)\cup\{\mathscr{A}(\gamma)\}$, and it is disconnected from $\mathfrak{K}(\bar\gamma)$ for all other elements $\gamma\in\mathcal{F}$. This is also the case if γ is a root of $\mathcal{F}$, so that $\mathscr{A}(\gamma) = \Gamma$ by our usual convention, if we set $\mathfrak{K}(\Gamma)$ to be the preimage in $\mathfrak{K}_{\mathcal{F}}\Gamma$ of the complement of all roots of $\mathcal{F}$. We henceforth write $v_\star(\gamma)$ for the unique vertex connecting $\mathfrak{K}(\gamma)$ to $\mathfrak{K}(\mathscr{A}(\gamma))$ and we write $\mathcal{V}_\gamma^\star = \mathcal{V}_\gamma \setminus \{v_\star(\gamma)\}$, so that one has a partition $\mathcal{V}_{\mathcal{F}} = \mathcal{V}_\Gamma \sqcup \bigsqcup_{\gamma\in\mathcal{F}} \mathcal{V}_\gamma^\star$.

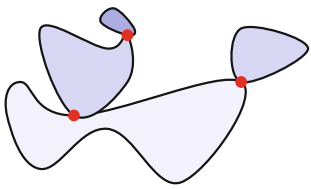
In this way, the tree structure of $\mathcal{F}$ is reflected in the topology of $\mathfrak{K}_{\mathcal{F}}\Gamma$, as illustrated in Fig. 3, where each $\mathfrak{K}(\gamma)$ is stylised by a coloured shape, with parents having lighter shades than their children and connecting vertices drawn in red. Recall that we also fixed a total order on the vertices of Γ (and therefore those of $\mathfrak{K}_{\mathcal{F}}\Gamma$) and that the construction of $\mathfrak{K}_{\mathcal{F}}\Gamma$ implies that the corresponding order on $\{v_\star(\gamma)\}_{\gamma\in\mathcal{F}}$ is compatible with the partial order on $\mathcal{F}$ given by inclusion. For $e\in\mathcal{E}_{\mathcal{F}}$, write $M_e \subset \{+,-\}$ for those ends such that $\tau(e)_\bullet \neq \tau(e_\bullet)$ for $\bullet\in M_e$ and set

$$\mathcal{E}_{\mathcal{F}}^\bullet = \{(e,\bullet) : e\in\mathcal{E}_{\mathcal{F}}^m,\ \bullet\in M_e\}\,.$$

Then, by the construction of $\mathfrak{K}_{\mathcal{F}}\Gamma$, for every $\bullet\in M_e$ there exists a unique $\gamma_\bullet(e)\in\mathcal{F}$ and vertex $e_\circ\in\mathcal{V}_{\mathcal{F}}$ such that

$$e_\bullet = v_\star(\gamma_\bullet(e))\,, \quad e_\circ = \tau^{-1}(\tau(e)_\bullet) \in \mathcal{V}_{\gamma_\bullet(e)}\,, \quad e\in\mathcal{E}_{\mathscr{A}(\gamma_\bullet(e))}\,. \tag{42}$$

Fig. 3 Structure of $\mathfrak{K}_{\mathcal{F}}\Gamma$



Given $\ell\colon \mathcal{E}^\bullet_{\mathcal{F}} \to \mathbf{N}^d$, we then define a canonical basis element $\mathcal{D}^\ell_{\mathcal{F}}\Gamma \in \hat{\mathcal{T}}_\Gamma$ by

$$\mathcal{D}^\ell_{\mathcal{F}}\Gamma = (\mathfrak{K}_{\mathcal{F}}\Gamma, \mathfrak{t}^{(\ell)}, \pi\ell)\ ,$$

where $\mathfrak{t}^{(\ell)}$ is the edge-labelling given by $\mathfrak{t}^{(\ell)}(e) = \mathfrak{t}(\tau(e)) + \sum_{\bullet\in M_e} \ell(e,\bullet)$, with $\mathfrak{t}$ the original edge-labelling of Γ, and $\pi\ell$ is the node-labelling given by $\pi\ell(v) = \sum\{\ell(e,\bullet) : e_\circ = v\}$. Given $\gamma \in \mathcal{F}$ and ℓ as above, we also set $\ell(\gamma) = \sum\{|\ell(e,\bullet)| : \gamma_\bullet(e) = \gamma\}$.

We now return to the bound (40) and first consider the special case when $\mathcal{F}_s$ is a safe forest such that $\mathcal{F}_u = \emptyset$. By (32) and (38), $\hat{\mathcal{R}}_{\mathcal{F}_s}\Gamma$ can then be written as

$$\hat{\mathcal{R}}_{\mathcal{F}_s}\Gamma = (-1)^{|\mathcal{F}_s|} \sum_{\ell\colon \mathcal{E}^\bullet_{\mathcal{F}_s}\to\mathbf{N}^d} \frac{(-1)^{\ell_{\mathrm{out}}}}{\ell!}\mathcal{D}^\ell_{\mathcal{F}_s}\Gamma\ , \tag{43}$$

where $\ell_{\mathrm{out}} = \sum\{|\ell(e,\bullet)| : \bullet = -\}$ and the sum in (43) is restricted to those choices of ℓ such that, for every $\gamma \in \mathcal{F}_s$, one has $\deg\gamma + \ell(\gamma) \le 0$.

In this case, we take as the index set I appearing in (41) all those functions ℓ appearing in the sum (43) (recall that the sum is restricted to finitely many such functions) and we set

$$\eta_\ell(u) = d + \sum_{e\in\mathcal{E}_{\mathcal{F}_s}} \mathfrak{t}^{(\ell)}(e)\mathbf{1}_{e^\uparrow}(u) + \sum_{(e,\bullet)\in\mathcal{E}^\bullet_{\mathcal{F}_s}} |\ell(e)|\mathbf{1}_{(e,\bullet)^\uparrow}(u)\ ,$$

where, for $e \in \mathcal{E}_{\mathcal{F}_s}$, $e^\uparrow$ denotes the node of T given by $\tau(e_-) \wedge \tau(e_+)$ and, for $(e,\bullet) \in \mathcal{E}^\bullet_{\mathcal{F}_s}$, $(e,\bullet)^\uparrow$ denotes the node $\tau(e_\circ) \wedge \tau(e_\bullet)$.

It follows from the definition of $\mathcal{W}^K$ that this choice does indeed satisfy (40). We now claim that as a consequence of the fact that $\mathcal{F}_s$ is such that $\mathcal{F}_u = \emptyset$, it also satisfies (41). Assume by contradiction that there exists a node u of T and a labelling ℓ such that $a \stackrel{\mathrm{def}}{=} \sum_{v\ge u}\eta_\ell(v) \le 0$. Write $\mathcal{V}_0 \subset \mathcal{V}_{\mathcal{F}_s}$ for the vertices v such that $\tau(v) \ge u$ in T and $\Gamma_0 = (\mathcal{E}_0, \mathcal{V}_0) \subset \mathfrak{K}_{\mathcal{F}_s}\Gamma$ for the corresponding subgraph. In general, Γ_0 does not need to be connected, so we write $\Gamma_0^{(i)} = (\mathcal{E}_0^{(i)}, \mathcal{V}_0^{(i)})$ for its connected components. We then set

$$a_i \stackrel{\mathrm{def}}{=} |\mathcal{V}_0^{(i)}| - 1 + \sum_{e\in\mathcal{E}_0^{(i)}} \mathfrak{t}^{(\ell)}(e) + \sum_{(e,\bullet)\in\mathcal{E}^\bullet_{\mathcal{F}_s}} |\ell(e)|\mathbf{1}_{\{e_\bullet,e_\circ\}\subset\mathcal{V}_0^{(i)}}\ ,$$

so that $\sum_i a_i \le a$, with equality if Γ_0 happens to be connected. Since $a \le 0$, there exists i such that $a_i \le 0$. Furthermore, i can be chosen such that $|\mathcal{V}_0^{(i)}| \ge 2$, since $|\mathcal{V}_0| \ge 2$ and we would otherwise have $a = |\mathcal{V}_0| - 1 \ge 1$.

Set $\mathcal{V}_{0,\gamma} = \mathcal{V}_0 \cap \mathcal{V}_\gamma$ and let $\mathcal{F}_s^{(i)} \subset \mathcal{F}_s \cup \{\Gamma\}$ be the subtree consisting of those γ such that either $\mathcal{E}_\gamma \cap \mathcal{E}_0^{(i)} \neq \emptyset$ or $v_\star(\gamma) \in \mathcal{V}_0^{(i)}$ (or both). We also break a_i into

contributions coming from each $\gamma \in \mathcal{F}_s^{(i)}$ by setting

$$a_{i,\gamma} \stackrel{\text{def}}{=} |\mathcal{V}_{0,\gamma}| - 1 + \sum_{e \in \mathcal{E}_\gamma \cap \mathcal{E}_0} \mathfrak{t}^{(\ell)}(e) + \sum_{(e,\bullet) \in \mathcal{E}^\bullet_{\mathcal{F}_s}} \mathbf{1}_{\gamma_\bullet(e)=\gamma} |\ell(e)| \mathbf{1}_{\{e_\bullet, e_\circ\} \subset \mathcal{V}_0^{(i)}} \,. \tag{44}$$

We claim that $\sum_\gamma a_{i,\gamma} = a_i$: recalling that one always has $\Gamma \in \mathcal{F}_s^{(i)}$ by definition, the only part which is not immediate is that $\sum_\gamma (|\mathcal{V}_{0,\gamma}| - 1) = |\mathcal{V}_0^{(i)}| - 1$. This is a consequence of the fact that in the sum $\sum_\gamma |\mathcal{V}_{0,\gamma}|$, each "connecting vertex" is counted double. Since $\mathcal{F}_s^{(i)}$ is a tree, the number of these equals $|\mathcal{F}_s^{(i)}| - 1$, whence the claim follows.

We introduce the following terminology. An element $\gamma \in \mathcal{F}_s \cup \{\Gamma\}$ is said to be "full" if $\mathcal{E}_\gamma \cap \mathcal{E}_0^{(i)} = \mathcal{E}_\gamma$, "empty" if $\mathcal{E}_\gamma \cap \mathcal{E}_0^{(i)} = \emptyset$, and "normal" otherwise. We also set $a_{i,\gamma} = 0$ for all empty γ with $\mathcal{V}_{0,\gamma} = \emptyset$. Recall furthermore the definition of $\deg \gamma$ for $\gamma \in \mathcal{F}_s$ given in (9) and the definition of $\ell(\gamma)$ given above. With this terminology, we then have the following.

Lemma 3.7 *A full subgraph γ cannot have an empty parent and one has*

$$\begin{aligned} a_{i,\gamma} &= \deg \gamma + \ell(\gamma) - \sum_{\bar\gamma \in \mathcal{C}(\gamma)} (\deg \bar\gamma + \ell(\bar\gamma)) && \text{if } \gamma \text{ is full,} \\ a_{i,\gamma} &= 0 && \text{if } \gamma \text{ is empty,} \\ a_{i,\gamma} &> - \sum_{\bar\gamma \in \mathcal{C}_\star(\gamma)} (\deg \bar\gamma + \ell(\bar\gamma)) && \text{if } \gamma \text{ is normal,} \end{aligned} \tag{45}$$

where $\mathcal{C}_\star(\gamma)$ consists of those children $\bar\gamma$ of γ such that $v_\star(\bar\gamma) \in \mathcal{V}_0^{(i)}$.

Before we proceed to prove Lemma 3.7, let us see how this leads to a contradiction. By (43), one has $\deg \bar\gamma + \ell(\bar\gamma) \le 0$ for every $\gamma \in \mathcal{F}_s$ and a fortiori $\deg \bar\gamma < 0$. Furthermore, since $|\mathcal{V}_0^{(i)}| \ge 2$, there exists at least one subgraph γ which is either full or normal. Since full subgraphs can only have parents that are either full or normal and since Γ itself cannot be full (since legs are never contained in $\mathcal{E}_0^{(i)}$), we have at least one normal subgraph. Since each of the negative terms $\deg \gamma + \ell(\gamma)$ appearing in the right hand side of the bound of $a_{i,\gamma}$ for γ full is compensated by a corresponding term in its parent, and since we use the strict inequality appearing for normal γ at least once, we conclude that one has indeed $\sum_\gamma a_{i,\gamma} > 0$ as required.

Proof of Lemma 3.7 Let us first show that the bounds (45) hold. If γ is empty, one has either $\gamma \notin \mathcal{F}_s^{(i)}$ in which case $\mathcal{V}_{0,\gamma} = \emptyset$ and $a_{i,\gamma} = 0$ by definition, or $\mathcal{V}_{0,\gamma} = v_\star(\gamma)$ in which case $a_{i,\gamma} = 0$ by (44). If γ is full, then it follows immediately from the definition of $\deg \gamma$ that one would have $a_{i,\gamma} = \deg \gamma - \sum_{\bar\gamma \in \mathcal{C}(\gamma)} \deg \bar\gamma$ if it weren't for the presence of the labels ℓ. If γ is full then, whenever $(e, \bullet)$ is such that $\gamma_\bullet(e) = \gamma$, one also has $\{e_\bullet, e_\circ\} \subset \mathcal{V}_0^{(i)}$ by (42) and the definition of being full. Similarly, one has $e \in \mathcal{E}_\gamma \cap \mathcal{E}_0$ whenever $\gamma_\bullet(e) \in \mathcal{C}(\gamma)$. The first identity in

(45) then follows from the fact that each edge with $\gamma_\bullet(e) = \gamma$ contributes $|\ell(e)|$ to the last term in (44) while each edge with $\gamma_\bullet(e) \in \mathscr{C}(\gamma)$ contributes $-|\ell(e)|$ to the penultimate term.

Regarding the last identity in (45), given a normal subgraph γ, write $\hat{\gamma}$ for the subgraph of Γ with edge set given by

$$\hat{\mathscr{E}} = \tau(\mathscr{E}_\gamma \cap \mathscr{E}_0^{(i)}) \cup \bigcup_{\bar{\gamma} \in \mathscr{C}_\star(\gamma)} \mathscr{E}(\bar{\gamma}) \,. \tag{46}$$

In exactly the same way as for a full subgraph, one then has $a_{i,\gamma} \geq \deg \hat{\gamma} - \sum_{\bar{\gamma} \in \mathscr{C}_\star(\gamma)} (\deg \bar{\gamma} + \ell(\bar{\gamma}))$. The reason why this is an inequality and not an equality is that we may have additional positive contributions coming from those $\ell(e)$ with $\gamma_m(e) = \gamma$ and such that $e_\circ \in \mathscr{V}_0^{(i)}$, while we do not have any negative contributions from those $\ell(e)$ with $\gamma_m(e) \in \mathscr{C}_\star(\gamma)$ but $e \notin \mathscr{E}_0^{(i)}$. The claim then follows from the fact that one necessarily has $\deg \hat{\gamma} > 0$ by the assumption that $\mathscr{F}_u = \emptyset$. Indeed, it follows from its definition and the construction of the Hepp sector T that the subgraph Γ_0 satisfies that $\mathrm{scale}_{\mathbf{T}}^{\mathscr{F}_s}(e) > \mathrm{scale}_{\mathbf{T}}^{\mathscr{F}_s}(e)$ for every edge $e \in \mathscr{E}_0$ and every edge $\bar{e}$ adjacent to Γ_0 in $\mathfrak{K}_{\mathscr{F}_s}\Gamma$, so that one would have $\hat{\gamma} \in \mathscr{F}_u$ otherwise.

It remains to show that if γ is a full subgraph, then it cannot have empty parents. This follows in essentially the same way as above, noting that if it were the case that γ has an empty parent, then it would be unsafe in $\mathscr{F}_s$, in direct contradiction with the fact that $\mathscr{F}_s$ is a safe forest. □

In order to complete the proof of Theorem 3.1, it remains to consider the general case when $\mathscr{F}_u \neq \emptyset$. In this case, setting $\mathbb{M} = [\mathscr{F}_s, \mathscr{F}_s \cup \mathscr{F}_u]$, we have

$$\hat{\mathscr{R}}_{\mathbb{M}}\Gamma = (-1)^{|\mathscr{F}_s|} \sum_{\ell \colon \mathscr{E}_{\mathscr{F}_s}^m \to \mathbf{N}^d} \frac{(-1)^{\ell_{\mathrm{out}}}}{\ell!} \Big(\prod_{\gamma \in \mathscr{F}_u} (\mathrm{id} - \hat{\mathscr{C}}_\gamma) \Big) \mathscr{D}_{\mathscr{F}_s}^\ell \Gamma \,, \tag{47}$$

with the sum over ℓ restricted in the same ways as before. Again, we bound each term in this sum separately, so that our index set I consists again of the subset of functions $\ell \colon \mathscr{E}_{\mathscr{F}_s}^m \to \mathbf{N}^d$ such that $\deg \gamma + \ell(\gamma) < 0$ for every $\gamma \in \mathscr{F}_s$, but this time each of these summands is still comprised of several terms generated by the action of the operators $\hat{\mathscr{C}}_\gamma$ for the "unsafe" graphs γ.

For any $\gamma \in \mathscr{F}_u$, we define a subgraph $\mathfrak{K}(\gamma)$ of $\mathfrak{K}_{\mathscr{F}_s}\Gamma$ as before, with the children of γ being those in $\mathscr{F}_s \cup \{\gamma\}$ *not* in all of $\mathscr{F}_s \cup \mathscr{F}_u$. The definition of γ being "unsafe" then guarantees that there exists a vertex $\gamma^\uparrow$ in T such that $\tau(\mathscr{V}_\gamma) = \{v \in \mathscr{V} \,:\, v \geq \gamma^\uparrow\}$. We furthermore define

$$\gamma^{\uparrow\uparrow} = \sup\{e^\uparrow \,:\, e \in \mathscr{E}_{\mathscr{A}(\gamma)} \;\&\; e \sim \mathfrak{K}(\gamma)\} \,,$$

with "$\sim$" meaning "adjacent to", which is well-defined since all of the elements appearing under the sup lie on the path joining $\gamma^\uparrow$ to the root of T. In particular, one has $\gamma^\uparrow > \gamma^{\uparrow\uparrow}$. We also set $N(\gamma) = 1 + \lfloor -\deg \gamma \rfloor$ with the convention that $N(\gamma) = 0$ for $\gamma \notin \mathscr{F}_u$.

We claim that this time, if we set

$$\eta_\ell(u) = d + \sum_{e \in \mathscr{E}_{\mathscr{F}_s}} \mathfrak{t}^{(\ell)}(e) \mathbf{1}_{e^\uparrow}(u) + \sum_{e \in \mathscr{E}^m_{\mathscr{F}_s}} |\ell(e)| \mathbf{1}_{e_m^\uparrow}(u) + \sum_{\gamma \in \mathscr{F}_u} N(\gamma) \big(\mathbf{1}_{\gamma^\uparrow}(u) - \mathbf{1}_{\gamma^{\uparrow\uparrow}}(u) \big) \,, \tag{48}$$

then η_ℓ does indeed satisfy the required properties, which then concludes the proof. As before, we assume by contradiction that there is u such that $a = \sum_{v \geq u} \eta_\ell(v) \leq 0$ and we define, for each connected component $\Gamma_0^{(i)}$ of Γ_0,

$$a_i \stackrel{\text{def}}{=} |\mathscr{V}_0^{(i)}| - 1 + \sum_{e \in \mathscr{E}_0^{(i)}} \mathfrak{t}^{(\ell)}(e) + \sum_{e \in \mathscr{E}^m_{\mathscr{F}_s}} |\ell(e)| \mathbf{1}_{\{e_\bullet, e_\circ\} \subset \mathscr{V}_0^{(i)}} + \sum_{\gamma \in \mathscr{F}_u} N(\gamma) \mathbf{1}_{\mathfrak{K}(\gamma) = \Gamma_0^{(i)} \cap \mathfrak{K}(\mathscr{A}(\gamma))} \,.$$

It is less obvious than before to see that $\sum a_i \leq a$ because of the presence of the last term. Given $\gamma \in \mathscr{F}_u$, there are two possibilities regarding the corresponding term in (48). If $\gamma^\uparrow < u$ in T, then it does not contribute to a at all. Otherwise, $\tau^{-1}(\gamma)$ is included in Γ_0 and we distinguish two cases. In the first case, one has $\mathfrak{K}(\gamma) = \mathfrak{K}(\mathscr{A}(\gamma)) \cap \Gamma_0$. In this case, since the inclusion $\gamma \subset \mathscr{A}(\gamma)$ is strict, there is at least one edge in $\mathfrak{K}(\mathscr{A}(\gamma))$ adjacent to $\mathfrak{K}(\gamma)$. Since this edge is also adjacent to Γ_0, it follows that in this case $\gamma^{\uparrow\uparrow} < u$ so that we have indeed a contribution $N(\gamma)$ to a. In the remaining case, the corresponding term may or may not contribute to a, but if it does, then its contribution is necessarily positive, so we can discard it and still have $\sum a_i \leq a$ as required.

As before, we then write $a_i = \sum_{\gamma \in \mathscr{F}_s^{(i)}} a_{\gamma,i}$ with

$$a_{i,\gamma} \stackrel{\text{def}}{=} |\mathscr{V}_{0,\gamma}| - 1 + \sum_{e \in \mathscr{E}_\gamma \cap \mathscr{E}_0} \mathfrak{t}^{(\ell)}(e) + \sum_{e \in \mathscr{E}^m_{\mathscr{F}_s}} \mathbf{1}_{\gamma_m(e) = \gamma} |\ell(e)| \mathbf{1}_{\{e_\bullet, e_\circ\} \subset \mathscr{V}_0^{(i)}}$$
$$+ \sum_{\bar\gamma \in \mathscr{F}_u} N(\bar\gamma) \mathbf{1}_{\mathfrak{K}(\bar\gamma) = \Gamma_0^{(i)} \cap \mathfrak{K}(\gamma)} \,. \tag{49}$$

We claim that the statement of Lemma 3.7 still holds in this case. Indeed, the only case that requires a slightly different argument is that when γ is "normal". In this case, defining again $\hat\gamma$ as in (46), we have

$$a_{i,\gamma} \geq \deg \hat\gamma + N(\hat\gamma) - \sum_{\bar\gamma \in \mathscr{C}_\star(\gamma)} (\deg \bar\gamma + \ell(\bar\gamma)) \,,$$

since the last term in (49) contributes precisely when $\hat\gamma \in \mathscr{F}_u$ and then only the term with $\bar\gamma = \hat\gamma$ is selected by the indicator function. The remainder of the argument, including the fact that this then yields a contradiction with the assumption that $a \leq 0$, is then identical to before since one always has $\deg \hat\gamma + N(\hat\gamma) > 0$.

In order to complete the proof of our main theorem, it thus remains to show that the choice of η_ℓ given in (48) allows to bound from above the contribution of the

Hepp sector indexed by T, in the sense that the bound (40) holds. The only non-trivial part of this is the presence of a term

$$N(\gamma)\big(\mathbf{1}_{\gamma\uparrow}(u) - \mathbf{1}_{\gamma\uparrow\uparrow}(u)\big)$$

for each factor of $(1 - \hat{\mathscr{C}}_\gamma)$ in (47). This will be a consequence of the following bound.

Lemma 3.8 *Let $K_i : \mathbf{S} \to \mathbf{R}$ be kernels satisfying the bound (2) with $\deg \mathfrak{t} = -\alpha_i < 0$ for $i \in I$ with I a finite index set, and write $I_\star = I \sqcup \{\star\}$. Let furthermore $x_i, y_i \in \mathbf{S}$ such that $|x_i - x_j| \le \delta < \Delta \le |x_i - y_j|$ for all $i, j \in I_\star$ and let $N \ge 0$ be an integer. Then, one has the bound*

$$\Big|\prod_{i\in I} K_i(x_i - y_i) - \sum_{|\ell|<N} \frac{1}{\ell!} \prod_{i\in I} (x_i - x_\star)^{\ell_i} \big(D^{\ell_i} K_i\big)(x_\star - y_i)\Big| \lesssim \delta^N \Delta^{-N} \prod_{i\in I} |y_i - x_i|^{-\alpha_i} \,. \tag{50}$$

Proof The proof is a straightforward application of Taylor's theorem to the function $x \mapsto \prod_{i\in I} K_i(x_i)$ defined on $\mathbf{S}^I$. For example, the version given in [14, Prop. A.1] shows that for every $\tilde{\ell} : I \to \mathbf{N}^d$ with $|\tilde{\ell}| = N$, there exist measures $\mathbb{Q}_{\tilde{\ell}}$ on $\mathbf{S}^I$ with total variation $\frac{1}{\tilde{\ell}!} \prod_i |(x_i - x_\star)^{\tilde{\ell}_i}| \lesssim \delta^N$ and support in the ball of radius $K\delta$ around $(x_\star, \dots, x_\star)$ (for some K depending only on $|I|$ and d) such that

$$\prod_{i\in I} K_i(x_i - y_i) - \sum_{|\ell|<N} \frac{1}{\ell!} \prod_{i\in I} (x_i - x_\star)^{\ell_i} \big(D^{\ell_i} K_i\big)(x_\star - y_i) = \sum_{|\tilde{\ell}|=N} \int \prod_{i\in I} \big(D^{\tilde{\ell}_i} K_i\big)(z_i - y_i)\, \mathbb{Q}_{\tilde{\ell}}(dz) \tag{51}$$

If $\Delta > (K+1)\delta$, then the claim follows at once from the fact that

$$\big|\big(D^{\tilde{\ell}_i} K_i\big)(z_i - y_i)\big| \lesssim |z_i - y_i|^{-\alpha_i - |\tilde{\ell}_i|} \lesssim \Delta^{-|\tilde{\ell}_i|} |x_i - y_i|^{-\alpha_i} \,.$$

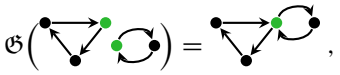
If $\Delta \le (K+1)\delta$ on the other hand, each term in the left hand side of (50) already satisfies the required bound individually.

It now remains to note that each occurence of $(1 - \hat{\mathscr{C}}_\gamma)$ in (47) produces precisely one factor of the type considered in Lemma 3.8, with the set I consisting of the edges in $\mathscr{A}(\gamma)$ adjacent to γ, $\delta = 2^{-\mathbf{n}(\gamma^\uparrow)}$ and $\Delta = 2^{-\mathbf{n}(\gamma^{\uparrow\uparrow})}$. The additional factor $\delta^{N(\gamma)} \Delta^{-N(\Gamma)}$ produced in this way precisely corresponds to the additional term $N(\gamma)\big(\mathbf{1}_{\gamma\uparrow}(u) - \mathbf{1}_{\gamma\uparrow\uparrow}(u)\big)$ in our definition of η. The only potential problem that could arise is when some edges are involved in the renormalisation of more than

one different subgraph. The explicit formula (51) however shows that this is not a problem. The proof of Theorem 3.1 is complete.

3.3 Properties of the BPHZ Valuation

In this section, we collect a few properties of the BPHZ valuation Π^K_{BPHZ}. In order to formulate the main tool for this, we first introduce a "gluing operator" $\mathfrak{G}\colon \hat{\mathcal{T}}_- \to \hat{\mathcal{T}}_-$ such that $\mathfrak{G}\Gamma$ is the connected vacuum diagram obtained by identifying all the marked vertices of Γ, for example



where the marked vertices are indicated in green. It follows from the definition (15) that the linear map Π^K_- satisfies the identity

$$\Pi^K_- \mathfrak{G}\tau = \Pi^K_- \tau\,, \qquad \tau \in \hat{\mathcal{T}}_-\,. \tag{52}$$

We claim that the same also holds for $\Pi^K_- \hat{\mathcal{A}}\pi$, where $\pi : \hat{\mathcal{T}}_- \to \mathcal{H}_-$ is the canonical projection.

Lemma 3.9 *One has* $\Pi^K_- \hat{\mathcal{A}}\pi\mathfrak{G}\tau = \Pi^K_- \hat{\mathcal{A}}\pi\tau$ *for all* $\tau \in \hat{\mathcal{T}}_-$.

Proof By induction on the number of connected components and since Π^K_-, $\hat{\mathcal{A}}$ and π are all multiplicative, it suffices to show that, for every element τ of the form $\tau = \gamma_1\gamma_2$ where the γ_i are connected and non-empty, one has the identity

$$\Pi^K_- \hat{\mathcal{A}}\pi\mathfrak{G}\tau = \Pi^K_- \hat{\mathcal{A}}\pi\gamma_1 \cdot \Pi^K_- \hat{\mathcal{A}}\pi\gamma_2 = \Pi^K_-\big(\hat{\mathcal{A}}\pi\gamma_1 \cdot \hat{\mathcal{A}}\pi\gamma_2\big)\,.$$

In particular, one has $\Pi^K_- \hat{\mathcal{A}}\tau = 0$ for every τ with $\deg\tau \le 0$ of the form $\mathfrak{G}(\gamma_1\gamma_2)$, as soon as one of the factors has strictly positive degree.

We will use the fact that, as a consequence of (28) combined with the definition of Δ^-, one has for connected $\sigma = (\Gamma, v_\star, \mathfrak{n})$ with $\deg\sigma \le 0$ the identity

$$\hat{\mathcal{A}}\sigma = -\sigma - \sum_{\substack{\bar\Gamma\subset\Gamma\\ \bar\Gamma\notin\{\circledcirc,\Gamma\}}} \mathscr{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar\Gamma}\sigma\,, \tag{53}$$

where we made use of the operators

$$\mathfrak{X}_{\bar\Gamma}\sigma = \sum_{\substack{\bar\ell\colon \partial\bar\Gamma\to\mathbf{N}^d\\ \bar{\mathfrak{n}}\colon \bar{\mathcal{V}}\to\mathbf{N}^d}} \frac{(-1)^{|\operatorname{out}\bar\ell|}}{\bar\ell!}\binom{\mathfrak{n}}{\bar{\mathfrak{n}}}(\bar\Gamma, \star, \bar{\mathfrak{n}} + \pi\bar\ell) \otimes (\Gamma, v_\star, \mathfrak{n} - \bar{\mathfrak{n}})/(\bar\Gamma, \bar\ell)$$

and $\star$ denotes some arbitrary choice of distinguished vertex. (Here, $\mathfrak{X}$ stands for "extract".) Note that the nonvanishing terms in (53) are always such that the degree of $\bar{\Gamma}$ (not counting node-decorations) is negative.

The proof of the lemma now goes by induction on the number of edges of $\tau = \gamma_1\gamma_2$. In the base case, each of the γ_i has one edge and there are two non-trivial cases. In the first case, $\deg \gamma_i \leq 0$ for both values of i. In this case, it follows from the above formula that, since $v_\star$ is the vertex in $\mathfrak{G}\tau$ at which both edges are connected and since $\hat{\mathcal{A}}\gamma_1 = -\gamma_i$, one has

$$\hat{\mathcal{A}}\mathfrak{G}\tau = -\mathfrak{G}\tau + 2\tau \;,$$

so that the claim follows from (52), combined with the fact that $\hat{\mathcal{A}}\pi\gamma_i = -\gamma_i$. In the second case, one has $\deg \gamma_1 \leq 0$ and $\deg \gamma_2 > 0$, but $\deg \gamma_1 + \deg \gamma_2 \leq 0$ so that $\pi\mathfrak{G}\tau = \mathfrak{G}\tau$. In this case, the only subgraph of $\mathfrak{G}\tau$ of negative degree is γ_1, so that

$$\hat{\mathcal{A}}\mathfrak{G}\tau = -\mathfrak{G}\tau + \tau \;,$$

thus yielding $\Pi_-^K \hat{\mathcal{A}}\mathfrak{G}\tau = 0$ as required.

We now write Γ for the graph associated to $\mathfrak{G}\tau$ and $\Gamma_i \subset \Gamma$ for the subgraphs associated to each of the factors γ_i. Writing $\mathcal{U}_\Gamma$ for the set of all non-empty proper subgraphs of Γ, we then have a natural bijection

$$\begin{aligned}\mathcal{U}_\Gamma = \mathcal{U}_{\Gamma_1} \sqcup \{\bar{\gamma}_1 \sqcup \Gamma_2 \,:\, \bar{\gamma}_1 \in \mathcal{U}_{\Gamma_1}\} \sqcup \mathcal{U}_{\Gamma_2} \sqcup \{\bar{\gamma}_2 \sqcup \Gamma_1 \,:\, \bar{\gamma}_2 \in \mathcal{U}_{\Gamma_2}\} \\ \sqcup \{\bar{\gamma}_1 \sqcup \bar{\gamma}_2 \,:\, \bar{\gamma}_1 \in \mathcal{U}_{\Gamma_1},\ \bar{\gamma}_2 \in \mathcal{U}_{\Gamma_2}\} \sqcup \{\Gamma_1, \Gamma_2\} \;. \qquad (54)\end{aligned}$$

Take now an element of the form $\bar{\gamma}_1 \sqcup \Gamma_2$ from the first set above. As before, there are no edges in Γ adjacent to Γ_1 other than those incident to $v_\star$. Furthermore, $\bar{\gamma}_1 \sqcup \Gamma_2$ has strictly less edges than Γ, so we can apply our induction hypothesis, yielding

$$\begin{aligned}\Pi_-^K \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1 \sqcup \Gamma_2}\mathfrak{G}\tau &= \Pi_-^K \mathcal{M}(\hat{\mathcal{A}}\pi\, \mathcal{M}_{\gamma_2} \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1}\gamma_1 \\ &= \Pi_-^K \big(\hat{\mathcal{A}}\pi\gamma_2 \cdot \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1}\gamma_1\big) \;,\end{aligned}$$

where $\mathcal{M}_{\gamma_2} \colon \gamma \mapsto \mathfrak{G}(\gamma \cdot \gamma_2)$. In a similar way, we obtain the identities

$$\begin{aligned}\Pi_-^K \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1}\mathfrak{G}\tau &= \Pi_-^K \big(\gamma_2 \cdot \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1}\gamma_1\big) \;, \\ \Pi_-^K \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1 \sqcup \bar{\gamma}_2}\mathfrak{G}\tau &= \Pi_-^K \big(\mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_1}\gamma_1 \cdot \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\bar{\gamma}_2}\gamma_2\big) \;, \\ \Pi_-^K \mathcal{M}(\hat{\mathcal{A}}\pi \otimes \mathrm{id})\mathfrak{X}_{\gamma_1}\mathfrak{G}\tau &= \Pi_-^K \big(\gamma_2 \cdot \hat{\mathcal{A}}\pi\gamma_1\big) \;,\end{aligned}$$

as well as the corresponding identities with 1 and 2 exchanged. Inserting these identities into (53) (with the sum broken up according to (54)), we obtain

$$\begin{aligned}\Pi_-^K \hat{\mathscr{A}} \mathfrak{G}\tau = &-\Pi_-^K \mathfrak{G}\tau - \sum_{\bar\gamma_1 \in \mathcal{U}_{\Gamma_1}} \Pi_-^K\big((\gamma_2 + \hat{\mathscr{A}}\pi\gamma_2)\cdot \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_1}\gamma_1\big)\\ &- \sum_{\bar\gamma_2 \in \mathcal{U}_{\Gamma_2}} \Pi_-^K\big((\gamma_1 + \hat{\mathscr{A}}\pi\gamma_1)\cdot \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_2}\gamma_2\big)\\ &- \sum_{\bar\gamma_1 \in \mathcal{U}_{\Gamma_1}}\sum_{\bar\gamma_2 \in \mathcal{U}_{\Gamma_2}} \Pi_-^K\big(\mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_1}\gamma_1\cdot \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_2}\gamma_2\big)\\ &- \Pi_-^K\big(\gamma_2\cdot \hat{\mathscr{A}}\pi\gamma_1\big) - \Pi_-^K\big(\gamma_1\cdot \hat{\mathscr{A}}\pi\gamma_2\big)\,.\end{aligned}$$

At this stage, we differentiate again between the case in which $\deg\gamma_i \le 0$ for both i and the case in which one of the two has positive degree. (The case in which both have positive degree is again trivial.) In the former case, $\pi\gamma_i = \gamma_i$ and one has

$$\gamma_2 + \hat{\mathscr{A}}\pi\gamma_2 = -\sum_{\bar\gamma_2 \in \mathcal{U}_{\Gamma_2}} \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_2}\gamma_2\,.$$

In particular, the second and third terms are the same as the fourth, but with opposite sign and one has

$$\begin{aligned}\Pi_-^K \hat{\mathscr{A}} \mathfrak{G}\tau = &\sum_{\bar\gamma_1 \in \mathcal{U}_{\Gamma_1}}\sum_{\bar\gamma_2 \in \mathcal{U}_{\Gamma_2}} \Pi_-^K\big(\mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_1}\gamma_1\cdot \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_2}\gamma_2\big)\\ &\quad - \Pi_-^K(\gamma_1\cdot\gamma_2) - \Pi_-^K\big(\gamma_2\cdot \hat{\mathscr{A}}\gamma_1\big) - \Pi_-^K\big(\gamma_1\cdot \hat{\mathscr{A}}\gamma_2\big)\\ = &\,\Pi_-^K\big((\hat{\mathscr{A}}\gamma_1 + \gamma_1)\cdot(\hat{\mathscr{A}}\gamma_2 + \gamma_2)\big)\\ &\quad - \Pi_-^K(\gamma_1\cdot\gamma_2) - \Pi_-^K\big(\gamma_2\cdot \hat{\mathscr{A}}\gamma_1\big) - \Pi_-^K\big(\gamma_1\cdot \hat{\mathscr{A}}\gamma_2\big)\\ = &\,\Pi_-^K\big(\hat{\mathscr{A}}\gamma_1\cdot \hat{\mathscr{A}}\gamma_2\big)\,,\end{aligned}$$

as claimed. Consider now the case $\deg\gamma_1 > 0$. Then, the two terms containing $\hat{\mathscr{A}}\pi\gamma_1$ vanish and we obtain similarly

$$\begin{aligned}\Pi_-^K \hat{\mathscr{A}} \mathfrak{G}\tau &= -\Pi_-^K \mathfrak{G}\tau - \sum_{\bar\gamma_1 \in \mathcal{U}_{\Gamma_1}} \Pi_-^K\big(\gamma_1\cdot \mathscr{M}(\hat{\mathscr{A}}\pi\otimes \mathrm{id})\mathfrak{X}_{\bar\gamma_2}\gamma_2\big) - \Pi_-^K\big(\gamma_1\cdot \hat{\mathscr{A}}\gamma_2\big)\\ &= -\Pi_-^K(\gamma_1\cdot\gamma_2) + \Pi_-^K\big(\gamma_1\cdot(\hat{\mathscr{A}}\gamma_2 + \gamma_2)\big) - \Pi_-^K\big(\gamma_1\cdot \hat{\mathscr{A}}\gamma_2\big) = 0\,,\end{aligned}$$

as claimed, thus concluding the proof. □

As a consequence of this result, we have the following. Recall that $\mathcal{S}_k^{(c)}$ is the space of translation invariant compactly supported (modulo translations) distributions in k variables. Given $x \in \mathbf{S}^k$, $y \in \mathbf{S}^\ell$, we also write $x \sqcup y = (x_1, \dots, x_k, y_1, \dots, y_\ell) \in \mathbf{S}^{k+\ell}$. For any $k, \ell \geq 1$, we then have a bilinear "convolution operator" $\star\colon \mathcal{S}_k^{(c)} \times \mathcal{S}_\ell^{(c)} \to \mathcal{S}_{k+\ell-2}^{(c)}$ obtained by setting

$$\big(\eta \star \zeta\big)(x \sqcup y) = \int_{\mathbf{S}} \eta(x \sqcup z)\zeta(z \sqcup y)\, dz\,, \quad x \in \mathbf{S}^{k-1}, y \in \mathbf{S}^{\ell-1}\,,$$

whenever η and ζ are represented by continuous functions. It is straightforward to see that this extends continuously to all of $\mathcal{S}_k^{(c)} \times \mathcal{S}_\ell^{(c)}$, and that it coincides with the usual convolution in the special case $k = \ell = 2$.

Similarly, we have a convolution operator $\star\colon \mathcal{H}_k \times \mathcal{H}_\ell \to \mathcal{H}_{k+\ell-2}$ obtained in the following way. Let $\Gamma \in \mathcal{T}_k$ and $\bar{\Gamma} \in \mathcal{T}_\ell$ be Feynman diagrams such that the label of the kth leg of Γ and the first leg of $\bar{\Gamma}$ are both given by δ. We then define $\Gamma \star \bar{\Gamma} \in \mathcal{T}_{k+\ell-2}$ to be the Feynman diagram with $k+\ell-2$ legs obtained by removing the kth leg of Γ as well as the first leg of $\bar{\Gamma}$, and identifying the two vertices these legs were connected to. (We also need to relabel the legs of $\bar{\Gamma}$ accordingly.) This operation extends to all of $\mathcal{H}_k \times \mathcal{H}_\ell$ by noting that given a Feynman diagram $\Gamma \in \mathcal{T}_k$, there always exists $\Gamma_n \in \mathcal{T}_k$ with $\Gamma_n = \Gamma$ in $\mathcal{H}_k$ which is a linear combination of diagrams with label δ on the nth leg: if the nth leg of Γ has label $\delta^{(m)}$ with $m \neq 0$, one obtains Γ_n by performing $|m|$ "integrations by parts" using (12). We then define in general $\Gamma \star \bar{\Gamma}$ by setting $\Gamma \star \bar{\Gamma} \stackrel{\text{def}}{=} \Gamma_k \star \bar{\Gamma}_0$ and we can check that this is indeed well-defined in $\mathcal{H}_{k+\ell-2}$. We then have the following consequence of Lemma 3.9.

Proposition 3.10 *The BPHZ valuation satisfies* $\Pi_{\text{BPHZ}}(\Gamma \star \bar{\Gamma}) = \Pi_{\text{BPHZ}}\Gamma \star \Pi_{\text{BPHZ}}\bar{\Gamma}$.

Proof Write $\mathcal{M}^\star\colon \mathcal{H} \otimes \mathcal{H} \to \mathcal{H}$ for the convolution operator introduced above and note that the canonical valuation Π (we suppress the dependence on K) does satisfy the property of the statement. It therefore suffices to show that one has the identity

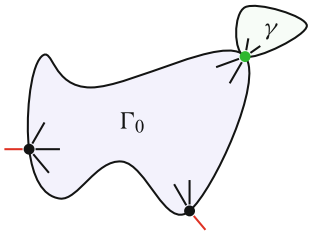
$$(\Pi_- \hat{\mathcal{A}} \otimes \mathrm{id})\Delta \mathcal{M}^\star = \mathcal{M}^\star\big((\Pi_- \hat{\mathcal{A}} \otimes \mathrm{id})\Delta \otimes (\Pi_- \hat{\mathcal{A}} \otimes \mathrm{id})\Delta\big) \tag{55}$$

between maps $\mathcal{H} \otimes \mathcal{H} \to \mathcal{H}$.

Suppose that $\Gamma \in \mathcal{T}_k$ and $\bar{\Gamma} \in \mathcal{T}_\ell$, write v for the vertex of Γ adjacent to the kth leg, and let $\bar{v}$ be the vertex of $\bar{\Gamma}$ adjacent to its first leg. Fix furthermore an arbitrary map $\sigma : (V_\star \sqcup \bar{V}_\star)/\{v, \bar{v}\} \to \mathbf{N}$ which is injective and such that $\sigma(v) = 0$. Since internal edges of $\Gamma \star \bar{\Gamma}$ are in bijection with the disjoint union of the internal edges of Γ and those of $\bar{\Gamma}$, we have an obvious bijection between subgraphs γ of $\Gamma \star \bar{\Gamma}$ and pairs (γ_1, γ_2) of subgraphs of Γ and $\bar{\Gamma}$. We also have a natural choice of distinguished vertex for each connected subgraph of Γ, $\bar{\Gamma}$ or $\Gamma \star \bar{\Gamma}$ by choosing the vertex with the lowest value of σ. If we then write $\hat{\Delta}\tau \in \hat{\mathcal{T}}_- \otimes \mathcal{H}$ for the right hand side of (19) with this choice of distinguished vertices, then we see that

$$(\mathfrak{G} \otimes \mathrm{id})\hat{\Delta}(\Gamma \star \bar{\Gamma}) = (\mathfrak{G}\mathcal{M} \otimes \mathcal{M}^\star)(\mathrm{id} \otimes \tau \otimes \mathrm{id})(\hat{\Delta}\Gamma \otimes \hat{\Delta}\bar{\Gamma})\,,$$

Fig. 4 Generalised self-loop



where $\tau : \hat{\mathcal{T}}_- \otimes \mathcal{H} \to \mathcal{H} \otimes \hat{\mathcal{T}}_-$ is the map that exchanges the two factors. Applying $\Pi_- \hat{\mathcal{A}} \pi$ to both sides and making use of Lemma 3.9, the required identity (55) follows at once.

Consider the situation of a Feynman diagram Γ containing a vertex v and a subgraph γ which is a "generalised self-loop at v" in the sense that

- The vertex v is the only vertex of γ that is adjacent to any edge not in γ.
- No leg of Γ is adjacent to any vertex of γ, except possibly for v.

We then obtain a new diagram Γ_0 by collapsing all of γ onto the vertex v, as illustrated in Fig. 4, where the vertex v is indicated in green and legs are drawn in red.

As a consequence of Proposition 3.10, we conclude that in such a situation there exists a constant $c_\gamma \in \mathbf{R}$ such that

$$\Pi_{\text{BPHZ}} \Gamma = c_\gamma \Pi_{\text{BPHZ}} \Gamma_0 \,,$$

and that furthermore $c_\gamma = 0$ as soon as $\deg \gamma \le 0$ as a consequence of Proposition 2.22. One particularly important special case is that of actual self-loops, where γ consists of a single edge connecting v to itself, thus showing that $\Pi_{\text{BPHZ}} \Gamma = 0$ for every Γ containing self-loops since the degree of a self-loop of type $\mathfrak{t}$ is given by $\deg \mathfrak{t}$, which is always negative.

Finally, it would also appear natural to restrict the sums in (19) and (24) to subgraphs $\bar{\Gamma}$ that are c-full in Γ (in the sense that each connected component of $\bar{\Gamma}$ is a full subgraph of Γ), especially in view of the proof of the BPHZ theorem where we saw that the "dangerous" connected subgraphs are always the full ones. We can then perform the exact same steps as before, including the construction of a corresponding twisted antipode and the verification of the forest formula. Writing $\hat{\mathcal{F}}_\Gamma^-$ for the subset of $\mathcal{F}_\Gamma^-$ consisting of forests $\mathcal{F}$ such that each $\gamma \in \mathcal{F}$ is a full subgraph of its parent $\mathcal{A}(\gamma)$ (as usual with the convention that the parent of the maximal elements is Γ itself), it is therefore natural in view of (33) to define a valuation

$$\Pi^{\text{full}}_{\text{BPHZ}} \Gamma = (\Pi_- \otimes \Pi) \sum_{\mathcal{F} \in \hat{\mathcal{F}}_\Gamma^-} (-1)^{|\mathcal{F}|} \mathcal{C}_{\mathcal{F}} \Gamma \,, \tag{56}$$

where Π and Π_- are the canonical valuations associated to some $K \in \mathcal{K}_\infty^-$. It turns out that, maybe not so surprisingly in view of Proposition 2.22, this actually yields the exact same valuation:

Proposition 3.11 *One has* $\Pi_{\text{BPHZ}}^{\text{full}} = \Pi_{\text{BPHZ}}$.

Proof In order to show that

$$(\Pi_- \otimes \Pi) \sum_{\mathcal{F} \in \mathcal{F}_\Gamma^- \setminus \hat{\mathcal{F}}_\Gamma^-} (-1)^{|\mathcal{F}|} \mathcal{C}_{\mathcal{F}} \Gamma = 0 \,,$$

we will partition $\mathcal{F}_\Gamma^- \setminus \hat{\mathcal{F}}_\Gamma^-$ into sets such that the above sum vanishes, when restricted to any of the sets in the partition. In order to formulate our construction, given $\gamma \in \mathcal{G}_\Gamma^-$, we write $\gamma^{\text{cl}} \in \mathcal{G}_\Gamma^-$ for the "closure" of γ in Γ, i.e. the full subgraph of Γ with the same vertex set as γ. For $\mathcal{F} \in \mathcal{F}_\Gamma^- \setminus \hat{\mathcal{F}}_\Gamma^-$, we then have a unique decomposition $\mathcal{F} = \mathcal{F}^{\text{full}} \cup \mathcal{F}^p$ such that each $\gamma \in \mathcal{F}^{\text{full}}$ is full in Γ, no element of $\mathcal{F}^{\text{full}}$ is contained in an element of $\mathcal{F}^p$, and no root of $\mathcal{F}^p$ is full in Γ.

Write $\mathcal{F}^p_{\text{max}}$ for the set of roots of $\mathcal{F}^p$ and set

$$\overline{\mathcal{F}^p} = \{\gamma^{\text{cl}} : \gamma \in \mathcal{F}^p_{\text{max}}\} \,.$$

In general, one may have $\mathcal{F}^{\text{full}} \cap \overline{\mathcal{F}^p} \neq \emptyset$, so we also set $\mathcal{F}_\circ^{\text{full}} = \mathcal{F}^{\text{full}} \setminus \overline{\mathcal{F}^p}$. If we write $\mathfrak{N} \colon \mathcal{F} \mapsto (\mathcal{F}^p, \mathcal{F}_\circ^{\text{full}})$, then we see that the preimage of $(\mathcal{F}^p, \mathcal{F}_\circ^{\text{full}})$ under $\mathfrak{N}$ consists of all forests of the form $\mathcal{F}^p \cup \mathcal{F}_\circ^{\text{full}} \cup \mathcal{B}$, where $\mathcal{B}$ is an arbitrary subset of $\overline{\mathcal{F}^p}$. Furthermore, $\mathcal{F}_\Gamma^- \setminus \hat{\mathcal{F}}_\Gamma^-$ consists precisely of those forests $\mathcal{F}$ such that $\mathcal{F}^p \neq \emptyset$. Since $\sum_{\mathcal{B} \subset \overline{\mathcal{F}^p}} (-1)^{|\mathcal{B}|} = 0$, it thus remains to show that the quantity

$$(\Pi_- \otimes \Pi) \mathcal{C}_{\mathcal{F}^p \cup \mathcal{F}_\circ^{\text{full}} \cup \mathcal{B}} \Gamma \tag{57}$$

is independent of $\mathcal{B} \subset \overline{\mathcal{F}^p}$.

To see that this is the case, consider the space $\hat{\mathcal{T}}_\Gamma$ and the operators $\hat{\mathcal{C}}_{\mathcal{F}}$ as in the proof of the BPHZ theorem and denote by $\hat{\Pi} \colon \hat{\mathcal{T}}_\Gamma \to \mathcal{S}$ the composition of $\Pi \colon \hat{\mathcal{T}} \to \mathcal{S}$ with the natural injection $\hat{\mathcal{T}}_\Gamma \hookrightarrow \hat{\mathcal{T}}$. One then has for every forest $\mathcal{G}$ the identity

$$(\Pi_- \otimes \Pi) \mathcal{C}_{\mathcal{G}} \Gamma = \hat{\Pi} \hat{\mathcal{C}}_{\mathcal{G}} \Gamma \,, \qquad \hat{\mathcal{C}}_{\mathcal{G}} \stackrel{\text{def}}{=} \prod_{\gamma \in \mathcal{G}} \hat{\mathcal{C}}_\gamma \,.$$

(As already pointed out before, the order of the operations does not matter here.) Let now $\gamma \in \mathcal{G}_\Gamma^-$ and consider the elements $\hat{\mathcal{C}}_\gamma \Gamma$ and $\hat{\mathcal{C}}_\gamma \hat{\mathcal{C}}_{\gamma^{\text{cl}}} \Gamma$. It follows from the definition of the operators $\hat{\mathcal{C}}_\gamma$ that all the terms appearing in both expressions consist of the same graph where edges in $\Gamma \setminus \gamma^{\text{cl}}$ adjacent to γ^{cl} are reconnected to the distinguished vertex $v_\star$ of γ and the edges in γ^{cl} that are not in γ are turned into self-loops for $v_\star$.

Regarding the edge and vertex-labels ℓ and $\mathfrak{n}$ generated by these operations, a straightforward application of the Chu-Vandermonde theorem shows that they yield the exact same terms in both cases. The only difference is that the function $\mathfrak{d}$ is equal to 1 on γ in the first case, while it equals 2 on γ and 1 on edges of γ^{cl} that are not in γ in the second case. This however would only make a difference if we were to compose this with an operator of the type $\hat{\mathcal{C}}_{\bar{\gamma}}$ for some $\bar{\gamma}$ with $\gamma \subset \bar{\gamma} \subset \gamma^{\mathrm{cl}}$. In our case however, we only use this in order to compare $\mathcal{C}_{\mathcal{F}^p \cup \mathcal{F}_\circ^{\mathrm{full}} \cup \mathcal{B}}$ to $\mathcal{C}_{\mathcal{F}^p \cup \mathcal{F}_\circ^{\mathrm{full}}}$, so that we consider the situation $\gamma \in \mathcal{F}^p_{\max}$. Since these graphs are all vertex-disjoint, it follows that $(\prod_{\gamma \in \mathcal{F}^p_{\max}} \hat{\mathcal{C}}_\gamma)\Gamma$ and $(\prod_{\gamma \in \mathcal{F}^p_{\max} \cup \mathcal{B}} \hat{\mathcal{C}}_\gamma)\Gamma$ only differ by the value of $\mathfrak{d}$ in the way described above.

Our construction of the sets $\mathcal{F}_\circ^{\mathrm{full}}$ and $\mathcal{F}^p$ then guarantees that this discrepancy is irrelevant when further applying $\hat{\mathcal{C}}_{\bar{\gamma}}$ for $\bar{\gamma} \in \mathcal{F}_\circ^{\mathrm{full}} \cup (\mathcal{F}^p \setminus \mathcal{F}^p_{\max})$, so that (57) is indeed independent of $\mathcal{B}$ as claimed.

4 Large-Scale Behaviour

We now consider the case of kernels $K_{\mathfrak{t}}$ that don't have compact support. In order to encode their behaviour at infinity, we assign to each label $\mathfrak{t} \in \mathfrak{L}$ a second degree $\deg_\infty \colon \mathfrak{L} \to \mathbf{R}_- \cup \{-\infty\}$ with $\deg_\infty \delta^{(k)} = -\infty$ and satisfying this time the consistency condition $\deg_\infty \mathfrak{t}^{(k)} = \deg_\infty \mathfrak{t}$.[1] We furthermore assume that we are given a collection of smooth kernels $R_{\mathfrak{t}} \colon \mathbf{R}^d \to \mathbf{R}$ for $\mathfrak{t} \in \mathfrak{L}_\star$ satisfying the bounds

$$|D^k R_{\mathfrak{t}}(x)| \lesssim (2 + |x|)^{\deg_\infty \mathfrak{t}} , \tag{58}$$

for all multiindices k, uniformly over all $x \in \mathbf{R}^d$, and such that

$$R_{\mathfrak{t}^{(k)}} = D^k R_{\mathfrak{t}} . \tag{59}$$

Similarly to before, we extend this to $\mathfrak{L}$ by using the convention $R_{\delta^{(m)}} \equiv 0$ and we write $\mathcal{K}_\infty^+$ for the set of all smooth *compactly supported* kernel assignments $\mathfrak{t} \mapsto R_{\mathfrak{t}}$, as well as $\mathcal{K}_0^+$ for its closure under the system of seminorms defined by (58).

Consider then the formal expression (5), but with each instance of $K_{\mathfrak{t}}$ replaced by $G_{\mathfrak{t}} = K_{\mathfrak{t}} + R_{\mathfrak{t}}$. The aim of this section is to exhibit a sufficient condition on Γ which guarantees that this expression can also be renormalised, using the same procedure as in the previous sections. The conditions we require in Theorem 4.3 below can be

[1] It would have looked more natural to impose the stronger condition $\deg_\infty \mathfrak{t}^{(k)} = \deg_\infty \mathfrak{t} - |k|$ as before. One may further think that in this case one would be able to extend Theorem 4.3 to all diagrams Γ, not just those in $\mathcal{H}_+$. This is wrong in general, although we expect it to be true after performing a suitable form of positive renormalisation as in [2, 3]. This is not performed here, and as a consequence we are unable to take advantage of the additional large-scale cancellations that the stronger condition $\deg_\infty \mathfrak{t}^{(k)} = \deg_\infty \mathfrak{t} - |k|$ would offer.

viewed as a large-scale analogue to the conditions of Weinberg's theorem. They are required because, unlike in [2, 3], we do not perform any "positive renormalisation" in the present article.

To formulate our main result, we introduce the following construction. Given a Feynman diagram Γ with at least one edge, consider a partition $\mathscr{P}_\Gamma$ of its inner vertex set, i.e. elements of $\mathscr{P}_\Gamma$ are non-empty subsets of $\mathscr{V}_\star$ and $\bigcup \mathscr{P}_\Gamma = \mathscr{V}_\star$. We always consider the case where the partition $\mathscr{P}_\Gamma$ consists of at least two subsets, in other words $|\mathscr{P}_\Gamma| \geq 2$. Given such a partition, we then set

$$\deg_\infty \mathscr{P}_\Gamma \stackrel{\text{def}}{=} \sum_{e \in \mathscr{E}(\mathscr{P}_\Gamma)} \mathfrak{t}(e) + d(|\mathscr{P}_\Gamma| - 1) \;,$$

where $\mathscr{E}(\mathscr{P}_\Gamma)$ consists of all internal edges $e \in \mathscr{E}_\star$ such that both ends e_+ and e_- are contained in different elements of $\mathscr{P}_\Gamma$. Note the strong similarity to (9), which is of course not a coincidence. We will call a partition $\mathscr{P}_\Gamma$ "tight" if there exists one single element $A \in \mathscr{P}_\Gamma$ containing all of the vertices $v_{i,\star} \in \mathscr{V}_\star$ that are connected to legs of Γ.

Given K and R in $\mathscr{K}_\infty^-$ and $\mathscr{K}_\infty^+$ respectively, we furthermore define a valuation $\Pi^{K,R}$ by setting as in (5)

$$\big(\Pi^{K,R}\Gamma\big)(\varphi) = \int_{\mathbf{S}^{\mathscr{V}_\star}} \prod_{e \in \mathscr{E}_\star} G_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \big(D_1^{\ell_1} \cdots D_k^{\ell_k} \varphi\big)(x_{v_1}, \dots, x_{v_k})\, dx \;, \quad (60)$$

where we used again the notation $G_\mathfrak{t} = K_\mathfrak{t} + R_\mathfrak{t}$. We then have the following result which is the analogue in this context of Proposition 2.4.

Proposition 4.1 *Let Γ be such that every tight partition $\mathscr{P}_\Gamma$ of its inner vertices satisfies $\deg_\infty \mathscr{P}_\Gamma < 0$. Then, the map $(K, R) \mapsto \Pi^{K,R}\Gamma$ extends continuously to all of $(K, R) \in \mathscr{K}_\infty^- \times \mathscr{K}_0^+$.*

Proof This is a corollary of Theorem 4.3 below: given (60) and given that we restrict ourselves to $K \in \mathscr{K}_\infty^-$, it suffices to note that $\Pi^{K,R} = \Pi^{0,K+R}_{\text{BPHZ}}$.

Remark 4.3 The reason why it is natural to restrict oneself to tight partitions can best be seen with the following very simple example. Consider the case

$$\Gamma = \quad \overset{0}{\bullet} \xrightarrow[v_1]{\mathfrak{t}_1} \bullet \xrightarrow[v_2]{\mathfrak{t}_2} \bullet \overset{0}{\underset{v_3}{}} \;.$$

Writing $G_i = K_{\mathfrak{t}_i} + R_{\mathfrak{t}_i}$ and identifying functions with distributions as usual, one then has $(\Pi^{K,R}\Gamma)(x, y) = (G_1 \star G_2)(y - x)$. If the G_i are smooth functions, then this is of course well-defined as soon as their combined decay at infinity is integrable, which naturally leads to the condition $\deg_\infty \mathfrak{t}_1 + \deg_\infty \mathfrak{t}_1 < -d$, which corresponds indeed to the condition $\deg_\infty \mathscr{P}_\Gamma < 0$ for $\mathscr{P}_\Gamma = \{\{v_1, v_3\}, \{v_2\}\}$, the only tight partition of the inner vertices of Γ. Considering instead *all* partitions would lead to the condition $\deg_\infty \mathfrak{t}_i < -d$ for $i = 1, 2$, which is much stronger than necessary.

Note now the following two facts.

- The condition of Proposition 4.1 is compatible with the definition of the space $\mathscr{H}$ in the sense that if it is satisfied for one of the summands in the left hand side of (12), then it is also satisfied for all the others, as an immediate consequence of the fact that $\deg_\infty \mathfrak{t}^{(k)} = \deg_\infty \mathfrak{t}$. In particular, we have a well-defined subspace $\mathscr{H}_+ \subset \mathscr{H}$ on which the condition of Proposition 4.1 holds and therefore $\Pi^{K,R}\Gamma$ is well-defined for $(K, R) \in \mathscr{K}_\infty^- \times \mathscr{K}_0^+$.
- If Γ satisfies the assumption of Proposition 4.1, then it is also satisfied for all of the Feynman diagrams appearing in the second factor of the summands of $\Delta\Gamma$, so that $\mathscr{H}_+$ is invariant under the action of $\mathscr{G}_-$ on $\mathscr{H}$.

This suggests that if we define a BPHZ renormalised valuation on $\mathscr{H}_+$ by

$$\Pi^{K,R}_{\text{BPHZ}} = \big(\Pi^K_- \hat{\mathscr{A}} \otimes \Pi^{K,R}\big)\Delta \,, \tag{61}$$

then it should be possible to extend it to kernel assignments exhibiting self-similar behaviour both at the origin and at infinity. This is indeed the case, as demonstrated by the main theorem of this section.

Theorem 4.3 *The map* $(K, R) \mapsto \Pi^{K,R}_{\text{BPHZ}}\Gamma$ *extends continuously to* $(K, R) \in \mathscr{K}_0^- \times \mathscr{K}_0^+$ *for all* $\Gamma \in \mathscr{H}_+$.

Proof Consider the space $\tilde{\mathcal{T}}$ defined as the vector space generated by the set of pairs $(\Gamma, \tilde{\mathcal{E}})$, where Γ is a Feynman diagram as before and $\tilde{\mathcal{E}} \subset \mathcal{E}_\star$ is a subset of its internal edges. We furthermore define a linear map $\mathfrak{X} : \mathcal{T} \to \tilde{\mathcal{T}}$ by $\mathfrak{X}\Gamma = \sum_{\tilde{\mathcal{E}} \subset \mathcal{E}_\star} (\Gamma, \tilde{\mathcal{E}})$, and we define a valuation on $\tilde{\mathcal{T}}$ by setting

$$\begin{aligned}\big(\tilde{\Pi}^{K,R}(\Gamma, \tilde{\mathcal{E}})\big)(\varphi) = \int_{\mathbf{S}^{\mathcal{V}_\star}} \prod_{e \in \mathcal{E}_\star \setminus \tilde{\mathcal{E}}} K_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \prod_{e \in \tilde{\mathcal{E}}} R_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-}) \\ \times \big(D_1^{\ell_1} \cdots D_k^{\ell_k} \varphi\big)(x_{v_1}, \dots, x_{v_k})\, dx \,,\end{aligned} \tag{62}$$

so that $\Pi^{K,R} = \tilde{\Pi}^{K,R}\mathfrak{X}$. Similarly to before, we define $\partial\tilde{\mathcal{T}}$ by the analogue of (12) and we set $\tilde{\mathscr{H}} = \tilde{\mathcal{T}}/\partial\tilde{\mathcal{T}}$, noting that $\tilde{\Pi}^{K,R}$ is well-defined on $\tilde{\mathscr{H}}$.

We also define a map $\tilde{\Delta} : \tilde{\mathscr{H}} \to \mathscr{H}_- \otimes \tilde{\mathscr{H}}$ in the same way as (19), but with the sum restricted to subgraphs γ whose edge sets are subsets of $\mathcal{E}_\star \setminus \tilde{\mathcal{E}}$. (This condition guarantees that $\tilde{\mathcal{E}}$ can naturally be identified with a subset of the quotient graph Γ/γ.) With this definition, one has the identity

$$\tilde{\Delta}\mathfrak{X} = (\mathrm{id} \otimes \mathfrak{X})\Delta \,,$$

as a consequence of the fact that the set of pairs $(\tilde{\mathcal{E}}, \gamma)$ such that $\tilde{\mathcal{E}} \subset \mathcal{E}_\star$ and γ is a subgraph of Γ containing only edges in $\mathcal{E}_\star \setminus \mathcal{E}$ is the same as the set of pairs such that γ is an arbitrary subgraph of Γ and $\tilde{\mathcal{E}}$ is a subset of the edges of Γ/γ. This in

turn implies that one has the identity

$$\Pi^{K,R}_{\text{BPHZ}}\Gamma = \big(\Pi^K_- \hat{\mathcal{A}} \otimes \Pi^{K,R}\big)\Delta\Gamma = \big(\Pi^K_- \hat{\mathcal{A}} \otimes \tilde{\Pi}^{K,R}\big)\tilde{\Delta}\mathfrak{X}\Gamma \,. \tag{63}$$

Let now $\tilde{\mathcal{T}}_+$ be the subspace of $\tilde{\mathcal{T}}$ consisting of pairs $(\Gamma, \tilde{\mathcal{E}})$ such that $\deg_\infty \mathcal{P} < 0$ for every tight partition $\mathcal{P}$ with $\mathcal{E}(\mathcal{P}) \subset \tilde{\mathcal{E}}$. Again, this defines a subspace $\tilde{\mathcal{H}}_+ \subset \tilde{\mathcal{H}}$ invariant under the action of $\mathcal{G}_-$ by $\tilde{\Delta}$ and $\mathfrak{X}$ maps $\mathcal{H}_+$ (defined as in the statement of the theorem) into $\tilde{\mathcal{H}}_+$, so that it remains to show that $\big(\Pi^K_- \hat{\mathcal{A}} \otimes \tilde{\Pi}^{K,R}\big)\tilde{\Delta}$ extends to kernels $(K, R) \in \mathcal{K}_0^- \times \mathcal{K}_0^+$ on all of $\tilde{\mathcal{H}}_+$.

For this, we now fix $\tau = (\Gamma, \tilde{\mathcal{E}}) \in \tilde{\mathcal{T}}_+$ and we remark that for $R \in \mathcal{K}_\infty^+$ we can interpret the factor $\prod_{e\in\tilde{\mathcal{E}}} R_{\mathfrak{t}(e)}(x_{e_+} - x_{e_-})$ in (62) as being part of the test function. More precisely, we set

$$\varphi \otimes_\tau R = \varphi(x_1, \dots, x_k) \prod_{e\in\tilde{\mathcal{E}}} R_{\mathfrak{t}(e)}(x_{[e]_+} - x_{[e]_-}) \,,$$

where $[\cdot]_\pm \colon \tilde{\mathcal{E}} \to \{k+1, \dots, k+2|\tilde{\mathcal{E}}|\}$ is an arbitrary but fixed numbering of the half-edges of $\tilde{\mathcal{E}}$. We then have $(\tilde{\Pi}^{K,R}\tau)(\varphi) = (\Pi^K \mathcal{U}\tau)(\varphi \otimes_\tau R)$, where $\mathcal{U}\tau \in \mathcal{T}_+$ is the Feynman diagram obtained by cutting each of the edges $e \in \tilde{\mathcal{E}}$ open, replacing them by two legs with label δ and numbers given by $[e]_\pm$. It is immediate from the definitions and the condition (59) that this is compatible with the actions of $\tilde{\Delta}$ and Δ in the sense that one has

$$((g \otimes \tilde{\Pi}^{K,R})\tilde{\Delta}\tau)(\varphi) = ((g \otimes \Pi^K)\Delta\mathcal{U}\tau)(\varphi \otimes_\tau R) \,, \qquad \forall g \in \mathcal{G}_- \,.$$

Inserting this into (63), we conclude that

$$\big(\Pi^{K,R}_{\text{BPHZ}}\Gamma\big)(\varphi) = \sum_{\tilde{\mathcal{E}}\subset\mathcal{E}_\star} \big(\Pi^K_{\text{BPHZ}}\mathcal{U}(\Gamma, \tilde{\mathcal{E}})\big)(\varphi \otimes_{(\Gamma,\tilde{\mathcal{E}})} R) \,,$$

so that it remains to bound separately each of the terms in this sum.

For this, we write $\mathbf{S}_d = \mathbf{Z}^d$ for the discrete analogue of our state space $\mathbf{S} = \mathbf{R}^d$, we set $N = k + 2|\tilde{\mathcal{E}}|$, and we write $1 = \sum_{x\in\mathbf{S}_d^N} \Psi_x$ for a partition of unity with the property that $\Psi_x(y) = \Psi_0(y - x)$ and that Ψ_0 is supported in a cube of sidelength 2 centred at the origin, so that it remains to show that

$$\sum_{x\in\mathbf{S}_d^N} S_x \,, \qquad S_x \stackrel{\text{def}}{=} \big(\Pi^K_{\text{BPHZ}}\mathcal{U}(\Gamma, \tilde{\mathcal{E}})\big)((\varphi \otimes_{(\Gamma,\tilde{\mathcal{E}})} R)\Psi_x) \,,$$

is absolutely summable. It then follows from Theorem 3.1 that the summand in the above expression is bounded by

$$|S_x| \lesssim \prod_{e\in\tilde{\mathcal{E}}} (1 + |x_{[e]_+} - x_{[e]_-}|)^{\deg_\infty \mathfrak{t}(e)} \,,$$

for all $(K, R) \in \mathcal{K}_0^- \times \mathcal{K}_0^+$. This expression is not summable in general, so we need to exploit the fact that there are many terms that vanish. For instance, since the test function φ is compactly supported, there exists C such that $S_x = 0$ as soon as $|x_i| \geq C$ for some $i \leq k$. Similarly, since the kernels $K_{\mathfrak{t}}$ are compactly supported, there exists C such that $S_x = 0$ as soon as there are two legs $[i]$ and $[j]$ of $\mathcal{U}_\tau$ attached to the same connected component and such that $|x_i - x_j| \geq C$.

Let now $\mathcal{P}$ be the finest tight partition for Γ with $\mathcal{E}(\mathcal{P}) \subset \tilde{\mathcal{E}}$ and let $L \in \mathcal{P}$ denote the (unique) set which contains all the vertices adjacent to the legs of Γ. We conclude from the above consideration that one has

$$\sum_{x \in \mathbf{S}_d^N} |S_x| \lesssim \sum_{y \in \mathbf{S}_d^{\mathcal{P}}} \mathbf{1}_{\{y_L = 0\}} \prod_{e \in \mathcal{E}(\mathcal{P})} (1 + |y_{[e_+]} - y_{[e_-]}|)^{\deg_\infty \mathfrak{t}(e)} , \tag{64}$$

where $[v] \in \mathcal{P}$ denotes the element of $\mathcal{P}$ containing the vertex v. At this stage, the proof is virtually identical to that of Weinberg's theorem, with the difference that we need to control the large-scale behaviour instead of the small-scale behaviour. We define Hepp sectors $D_{\mathbf{T}} \subset \mathbf{S}_d^{\mathcal{P}}$ for $\mathbf{T} = (T, \mathbf{n})$ in exactly the same way as before, the difference being that this time no two elements can be at distance *less* than 1, so that we can restrict ourselves to scale assignments with $\mathbf{n}_v \leq 0$ for every inner vertex of T. Also, in view of (64), the leaves of T are this time given by elements of $\mathcal{P}$. In the same way as before, the number of elements of $D_{\mathbf{T}}$ is of the order of $\prod_{u \in T} 2^{-d\mathbf{n}_u}$ so that one has again a bound of the type

$$\sum_{x \in \mathbf{S}_d^N} |S_x| \lesssim \sum_{\mathbf{T}} \prod_{u \in T} 2^{-\mathbf{n}_u \eta_u} , \quad \eta_u = d + \sum_{e \in \mathcal{E}(\mathcal{P})} \mathbf{1}_{e^\uparrow} \deg_\infty \mathfrak{t}(e) , \tag{65}$$

where $e^\uparrow$ denotes the common ancestor in T of the two elements of $\mathcal{P}$ containing the two endpoints of e. Our assumption on Γ now implies that for every initial segment T_i of T,[2] one has $\sum_{u \in T_i} \eta_u < 0$. This is because one has $\sum_{u \in T_i} \eta_u = \deg \mathcal{P}_{T_i}$, where $\mathcal{P}_{T_i}$ is the coarsest coarsening of $\mathcal{P}$ such that for every edge $e \in \mathcal{E}(\mathcal{P}_{T_i})$, one has $e^\uparrow \notin \mathcal{P}_{T_i}$.

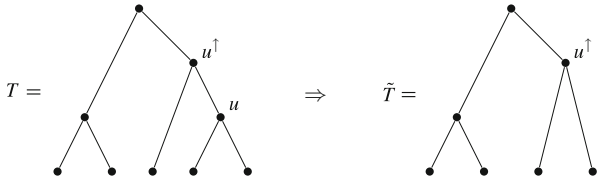
We claim that any such η satisfies

$$S(T, \eta) \stackrel{\text{def}}{=} \sum_{\mathbf{n}} \prod_{u \in T} 2^{-\mathbf{n}_u \eta_u} < \infty ,$$

where again the sum is restricted to negative $\mathbf{n}$ that are monotone on T. This can be shown by induction over the number of leaves of T. If T has only two leaves, then this is a converging geometric series and the claim is trivial. Let now T be a tree with $m \geq 3$ leaves and assume that the claim holds for all trees with $m - 1$ leaves. Pick an inner vertex u of T which has exactly two descendants (such a vertex always

[2] I.e. T_i is such that if $u \in T_i$ and $v \leq u$, then $v \in T_i$.

exists since T is binary) and write $\tilde{T}$ for the new tree obtained from T by deleting u and coalescing its two descendants into one single leaf. Write furthermore $u^\uparrow$ for the parent of u in T, which exists since T has at least three leaves. The following example illustrates this construction:



Since the condition on η is open and since S_η increases when increasing η_u, we can assume without loss of generality that $\eta_u \neq 0$. There are then two cases:

- If $\eta_u < 0$, we have $\sum_{\mathbf{n}_u > \mathbf{n}_{u^\uparrow}} 2^{-\mathbf{n}_u \eta_u} \approx 1$, so that

$$S(T, \eta) \approx S(\tilde{T}, \tilde{\eta}) \,, \tag{66}$$

  where $\tilde{\eta}_v$ is just the restriction of η_v to the tree $\tilde{T}$. Since initial segments of $\tilde{T}$ are also initial segments of T and since $\tilde{\eta} = \eta$ on them, we can make use of the induction hypothesis to conclude.
- If $\eta_u > 0$, we have $\sum_{\mathbf{n}_u > \mathbf{n}_{u^\uparrow}} 2^{-\mathbf{n}_u \eta_u} \approx 2^{-\mathbf{n}_{u^\uparrow} \eta_u}$, so that (66) holds again, but this time $\tilde{\eta}_{u^\uparrow} = \eta_{u^\uparrow} + \eta_u$ and $\tilde{\eta}_v = \eta_v$ otherwise. We conclude in the same way as before since the only "dangerous" case is that of initial segments $\tilde{T}_i$ containing $u^\uparrow$, but these are in bijection with the initial segment $T_i = \tilde{T}_i \cup \{u\}$ of T such that $\sum_{v \in \tilde{T}_i} \tilde{\eta}_v = \sum_{v \in T_i} \eta_v$, so that the induction hypothesis still holds.

Applying this to (65) completes the proof of the theorem.

Remark 4.4 While the definition of $\Pi^{K,R}_{\text{BPHZ}}$ is rather canonical, given kernel assignments K and R, the decomposition $G = K + R$ is *not*. Using the fact that $\mathcal{G}_-$ is a group, it is however not difficult to see that, for any two choices $(K, R), (\bar{K}, \bar{R}) \in \mathcal{K}_0^- \times \mathcal{K}_0^+$ such that

$$K_\mathfrak{t} + R_\mathfrak{t} = \bar{K}_\mathfrak{t} + \bar{R}_\mathfrak{t} \,, \qquad \forall \mathfrak{t} \in \mathfrak{L}_\star \,,$$

there exists an element $g \in \mathcal{G}_-$ such that $\Pi^{\bar{K},\bar{R}}_{\text{BPHZ}} = (g \otimes \Pi^{K,R}_{\text{BPHZ}})\Delta$.

Acknowledgements The author would like to thank Ajay Chandra and Philipp Schönbauer for several useful discussions during the preparation of this article. Financial support through ERC consolidator grant 615897 and a Leverhulme leadership award is gratefully acknowledged.

References

1. Bogoliubow, N.N., Parasiuk, O.S.: Über die Multiplikation der Kausalfunktionen in der Quantentheorie der Felder. Acta Math. **97**, 227–266 (1957). https://doi.org/10.1007/BF02392399
2. Bruned, Y., Hairer, M., Zambotti, L.: Algebraic renormalisation of regularity structures (2016). ArXiv e-prints. http://arxiv.org/abs/1610.08468
3. Chandra, A., Hairer, M.: An analytic BPHZ theorem for regularity structures (2016). ArXiv e-prints. http://arxiv.org/abs/1612.08138
4. Connes, A., Kreimer, D.: Hopf algebras, renormalization and noncommutative geometry. Commun. Math. Phys. **199**(1), 203–242 (1998). http://arxiv.org/abs/hep-th/9808042, https://doi.org/10.1007/s002200050499
5. Connes, A., Kreimer, D.: Renormalization in quantum field theory and the Riemann-Hilbert problem. I. The Hopf algebra structure of graphs and the main theorem. Commun. Math. Phys. **210**(1), 249–273 (2000). http://arxiv.org/abs/hep-th/9912092, https://doi.org/10.1007/s002200050779
6. Connes, A., Kreimer, D.: Renormalization in quantum field theory and the Riemann-Hilbert problem. II. The β-function, diffeomorphisms and the renormalization group. Commun. Math. Phys. **216**(1), 215–241 (2001). http://arxiv.org/abs/hep-th/0003188, https://doi.org/10.1007/PL00005547
7. de Calan, C., Rivasseau, V.: Local existence of the Borel transform in Euclidean Φ_4^4. Commun. Math. Phys. **82**(1), 69–100 (1981/1982)
8. Dyson, F.J.: The radiation theories of Tomonaga, Schwinger, and Feynman. Phys. Rev. (2) **75**, 486–502 (1949). https://doi.org/10.1103/PhysRev.75.486
9. Dyson, F.J.: The S matrix in quantum electrodynamics. Phys. Rev. (2) **75**, 1736–1755 (1949). https://doi.org/10.1103/PhysRev.75.1736
10. Epstein, H., Glaser, V.: The role of locality in perturbation theory. Ann. Inst. H. Poincaré Sect. A (N.S.) **19**, 211–295 (1973/1974)
11. Feldman, J., Magnen, J., Rivasseau, V., Sénéor, R.: Bounds on renormalized Feynman graphs. Commun. Math. Phys. **100**(1), 23–55 (1985). https://doi.org/10.1007/BF01212686
12. Gallavotti, G., Nicolò, F.: Renormalization theory in four-dimensional scalar fields. I. Commun. Math. Phys. **100**(4), 545–590 (1985)
13. Gallavotti, G., Nicolò, F.: Renormalization theory in four-dimensional scalar fields. II. Commun. Math. Phys. **101**(2), 247–282 (1985)
14. Hairer, M.: A theory of regularity structures. Invent. Math. **198**(2), 269–504 (2014). http://arxiv.org/abs/1303.5113, https://doi.org/10.1007/s00222-014-0505-4
15. Hairer, M.: The motion of a random string (2016). ArXiv e-prints. http://arxiv.org/abs/1605.02192
16. Hairer, M., Quastel, J.: A class of growth models rescaling to KPZ (2015). ArXiv e-prints. http://arxiv.org/abs/1512.07845
17. Hepp, K.: Proof of the Bogoliubov-Parasiuk theorem on renormalization. Commun. Math. Phys. **2**(4), 301–326 (1966). https://doi.org/10.1007/BF01773358
18. Salam, A.: Divergent integrals in renormalizable field theories. Phys. Rev. (2) **84**, 426–431 (1951). https://doi.org/10.1103/PhysRev.84.426
19. Salam, A.: Overlapping divergences and the S-matrix. Phys. Rev. (2) **82**, 217–227 (1951). https://doi.org/10.1103/PhysRev.82.217
20. Weinberg, S.: High-energy behavior in quantum field-theory. Phys. Rev. (2) **118**, 838–849 (1960). https://doi.org/10.1103/PhysRev.118.838
21. Zimmermann, W.: Convergence of Bogoliubov's method of renormalization in momentum space. Commun. Math. Phys. **15**, 208–234 (1969). https://doi.org/10.1007/BF01645676

Parabolic Anderson Model with Rough Dependence in Space

Yaozhong Hu, Jingyu Huang, Khoa Lê, David Nualart, and Samy Tindel

Abstract This paper studies the one-dimensional parabolic Anderson model driven by a Gaussian noise which is white in time and has the covariance of a fractional Brownian motion with Hurst parameter $H \in (\frac{1}{4}, \frac{1}{2})$ in the space variable. We derive the Wiener chaos expansion of the solution and a Feynman-Kac formula for the moments of the solution. These results allow us to establish sharp lower and upper asymptotic bounds for the nth moment of the solution.

Y. Hu is supported by an NSERC discovery grant.
D. Nualart is supported by the NSF grant DMS1512891 and the ARO grant FED0070445.
S. Tindel is supported by the NSF grant DMS1613163.

Y. Hu (✉)
Department of Mathematical and Statistical Sciences, University of Alberta at Edmonton, Edmonton, AB, Canada
e-mail: yaozhong@ualberta.ca

J. Huang
School of Mathematics, University of Birmingham, Birmingham, UK
e-mail: j.huang@bham.ac.uk

K. Lê
Department of Mathematics, Imperial College London, London, UK
e-mail: k.le@imperial.ac.uk

D. Nualart
Department of Mathematics, University of Kansas, Lawrence, KS, USA

S. Tindel
Department of Mathematics, Purdue University, West Lafayette, IN, USA
e-mail: stindel@purdue.edu

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_17

1 Introduction

A recent paper [9] studies the stochastic heat equation for $(t, x) \in (0, \infty) \times \mathbb{R}$

$$\frac{\partial u}{\partial t} = \frac{\kappa}{2} \frac{\partial^2 u}{\partial x^2} + \sigma(u)\, \dot{W}\,, \tag{1}$$

where $\dot{W}$ is a centered Gaussian noise which is white in time and behaves as fractional Brownian motion with Hurst parameter $1/4 < H < 1/2$ in space, and σ may be a nonlinear function with some smoothness.

However, the specific case $\sigma(u) = u$, i.e.

$$\frac{\partial u}{\partial t} = \frac{\kappa}{2} \frac{\partial^2 u}{\partial x^2} + u\, \dot{W} \tag{2}$$

deserves some specific treatment due to its simplicity. Indeed, this linear equation turns out to be a continuous version of the parabolic Anderson model, and is related to challenging systems in random environment like KPZ equation [3, 6] or polymers [1, 4]. The localization and intermittency properties of (2) have thus been thoroughly studied for equations driven by a space-time white noise (see [13] for a nice survey), while a recent trend consists in extending this kind of result to equations driven by very general Gaussian noises [5, 8, 10, 11]. However, the rough noise $\dot{W}$ presented in this work is not covered by the aforementioned references.

To fill this gap, we first tackle the existence and uniqueness problem. Although the existence and uniqueness of the solution in the general nonlinear case (1) has been established in [9], in this linear case (2), one can implement a rather simple procedure involving Fourier transforms. Since this point of view is interesting in its own right and is short enough, we develop it in Sect. 3.1. In Sect. 3.2, we study the random field solution using chaos expansion. Following the approach introduced in [8, 10], we obtain an explicit formula for the kernels of the Wiener chaos expansion and we show its convergence, and thus obtain the existence and uniqueness of the solution. It is worth noting these methods treat different classes of initial data which are more general than in [9] and different from [2].

We then move to a Feynman-Kac type representation for the moments of the solution. In fact, we cannot expect a Feynman-Kac formula for the solution, because the covariance is rougher than the space-time white noise case, and this type of formula requires smoother covariance structures (see, for instance, [11]). However, by means of Fourier analysis techniques as in [8, 10], we are able to obtain a Feynman-Kac formula for the moments that involves a fractional derivative of the Brownian local time.

Finally, the previous considerations allow us to handle, in the last section of the paper, the intermittency properties of the solution. More precisely, we show sharp lower bounds for the moments of the solution of the form $\mathbf{E}[u(t, x)^n] \geq \exp(Cn^{1+\frac{1}{H}}t)$, for all $t \geq 0$, $x \in \mathbb{R}$ and $n \geq 2$, where C is independent of $t \geq 0$,

$x \in \mathbb{R}$ and n. These bounds entail the intermittency phenomenon and match the corresponding estimates for the case $H > \frac{1}{2}$ obtained in [10]. After the completion of this work, three of the authors have studied the parabolic Anderson model in more details in [12]. Existence and uniqueness results are extended to wider class of initial data. In particular, exact long term asymptotics for the moments of the solution of the form $\limsup \frac{1}{t} \sup_{|x|>\alpha t} \log \mathbb{E}(|u(t,x)|^p)$ are obtained.

2 Preliminaries

Let us start by introducing our basic notation on Fourier transforms of functions. The space of Schwartz functions is denoted by $\mathcal{S}$. Its dual, the space of tempered distributions, is $\mathcal{S}'$. The Fourier transform of a function $u \in \mathcal{S}$ is defined with the normalization

$$\mathcal{F}u(\xi) = \int_{\mathbb{R}} e^{-i\xi x} u(x) dx,$$

so that the inverse Fourier transform is given by $\mathcal{F}^{-1}u(\xi) = (2\pi)^{-1}\mathcal{F}u(-\xi)$. The Fourier transform of a tempered distribution can also be defined (see [18]).

Let $\mathcal{D}((0,\infty) \times \mathbb{R})$ denote the space of real-valued infinitely differentiable functions with compact support on $(0,\infty) \times \mathbb{R}$. Taking into account the spectral representation of the covariance function of the fractional Brownian motion in the case $H < \frac{1}{2}$ proved in [17, Theorem 3.1], we represent our noise W by a zero-mean Gaussian family $\{W(\varphi),\ \varphi \in \mathcal{D}((0,\infty) \times \mathbb{R})\}$ defined on a complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$, whose covariance structure is given by

$$\mathbf{E}\left[W(\varphi)\, W(\psi)\right] = c_{1,H} \int_{\mathbb{R}_+ \times \mathbb{R}} \mathcal{F}\varphi(s,\xi)\, \overline{\mathcal{F}\psi(s,\xi)}\, |\xi|^{1-2H}\, ds d\xi, \tag{3}$$

where the Fourier transforms $\mathcal{F}\varphi, \mathcal{F}\psi$ are understood as Fourier transforms in space only and

$$c_{1,H} = \frac{1}{2\pi} \Gamma(2H+1) \sin(\pi H)\,. \tag{4}$$

We denote by $\mathfrak{H}$ the Hilbert space obtained by completion of $\mathcal{D}((0,\infty) \times \mathbb{R})$ with respect to the inner product

$$\langle \varphi, \psi \rangle_{\mathfrak{H}} = c_{1,H} \int_{\mathbb{R}_+ \times \mathbb{R}} \mathcal{F}\varphi(s,\xi) \overline{\mathcal{F}\psi(s,\xi)} |\xi|^{1-2H} d\xi ds\,. \tag{5}$$

The next proposition is from Theorem 3.1 and Proposition 3.4 in [17].

Proposition 2.1 *If $\mathfrak{H}_0$ denotes the class of functions $\varphi \in L^2(\mathbb{R}_+ \times \mathbb{R})$ such that*

$$\int_{\mathbb{R}_+ \times \mathbb{R}} |\mathcal{F}\varphi(s,\xi)|^2 |\xi|^{1-2H} d\xi ds < \infty ,$$

then $\mathfrak{H}_0$ is not complete and the inclusion $\mathfrak{H}_0 \subset \mathfrak{H}$ is strict.

We recall that the Gaussian family W can be extended to $\mathfrak{H}$ and this produces an isonormal Gaussian process, for which Malliavin calculus can be applied. We refer to [16] and [7] for a detailed account of the Malliavin calculus with respect to a Gaussian process. On our Gaussian space, the smooth and cylindrical random variables F are of the form

$$F = f(W(\phi_1), \ldots, W(\phi_n)) ,$$

with $\phi_i \in \mathfrak{H}$, $f \in C_p^\infty(\mathbb{R}^n)$ (namely f and all its partial derivatives have polynomial growth). For this kind of random variable, the derivative operator D in the sense of Malliavin calculus is the $\mathfrak{H}$-valued random variable defined by

$$DF = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(W(\phi_1), \ldots, W(\phi_n))\phi_j .$$

The operator D is closable from $L^2(\Omega)$ into $L^2(\Omega; \mathfrak{H})$ and we define the Sobolev space $\mathbb{D}^{1,2}$ as the closure of the space of smooth and cylindrical random variables under the norm

$$\|DF\|_{1,2} = \sqrt{\mathbf{E}[F^2] + \mathbf{E}[\|DF\|_{\mathfrak{H}}^2]} .$$

We denote by δ the adjoint of the derivative operator (called divergence operator) given by the duality formula

$$\mathbf{E}\left[\delta(u)F\right] = \mathbf{E}\left[\langle DF, u\rangle_{\mathfrak{H}}\right], \tag{6}$$

for any $F \in \mathbb{D}^{1,2}$ and any element $u \in L^2(\Omega; \mathfrak{H})$ in the domain of δ.

For any integer $n \geq 0$ we denote by $\mathbf{H}_n$ the nth Wiener chaos of W. We recall that $\mathbf{H}_0$ is simply $\mathbb{R}$ and for $n \geq 1$, $\mathbf{H}_n$ is the closed linear subspace of $L^2(\Omega)$ generated by the random variables $\{H_n(W(\phi)), \phi \in \mathfrak{H}, \|\phi\|_{\mathfrak{H}} = 1\}$, where H_n is the nth Hermite polynomial. For any $n \geq 1$, we denote by $\mathfrak{H}^{\otimes n}$ (resp. $\mathfrak{H}^{\odot n}$) the nth tensor product (resp. the nth symmetric tensor product) of $\mathfrak{H}$. Then, the mapping $I_n(\phi^{\otimes n}) = H_n(W(\phi))$ can be extended to a linear isometry between $\mathfrak{H}^{\odot n}$ (equipped with the modified norm $\sqrt{n!}\|\cdot\|_{\mathfrak{H}^{\otimes n}}$) and $\mathbf{H}_n$.

Consider now a random variable $F \in L^2(\Omega)$ which is measurable with respect to the σ-field $\mathcal{F}$ generated by W. This random variable can be expressed as

$$F = \mathbf{E}\big[F\big] + \sum_{n=1}^{\infty} I_n(f_n), \tag{7}$$

where the series converges in $L^2(\Omega)$, and the elements $f_n \in \mathfrak{H}^{\odot n}$, $n \geq 1$, are determined by F. This identity is called the Wiener chaos expansion of F.

The Skorohod integral (or divergence) of a random field u can be computed by using the Wiener chaos expansion. More precisely, suppose that $u = \{u(t,x), (t,x) \in \mathbb{R}_+ \times \mathbb{R}\}$ is a random field such that for each (t,x), $u(t,x)$ is an $\mathcal{F}$-measurable and square-integrable random variable. Then, for each (t,x) we have a Wiener chaos expansion of the form

$$u(t,x) = \mathbf{E}\big[u(t,x)\big] + \sum_{n=1}^{\infty} I_n(f_n(\cdot,t,x)). \tag{8}$$

Suppose that $\mathbf{E}[\|u\|^2_{\mathfrak{H}}]$ is finite. Then, we can interpret u as a square-integrable random function with values in $\mathfrak{H}$ and the kernels f_n in the expansion (8) are functions in $\mathfrak{H}^{\otimes(n+1)}$ which are symmetric in the first n variables. In this situation, u belongs to the domain of the divergence operator (that is, u is Skorohod integrable with respect to W) if and only if the following series converges in $L^2(\Omega)$

$$\delta(u) = \int_0^{\infty} \int_{\mathbb{R}^d} u(t,x)\, \delta W(t,x) = W(\mathbf{E}[u]) + \sum_{n=1}^{\infty} I_{n+1}(\widetilde{f}_n), \tag{9}$$

where $\widetilde{f}_n$ denotes the symmetrization of f_n in all its $n+1$ variables.

For each $t \geq 0$, let $\mathcal{F}_t$ be the σ-field generated by W up to time t. Define the predictable σ-field as the σ-field of subsets of $\Omega \times \mathbb{R}_+ \times \mathbb{R}$ generated by the collection of sets $\{A \times (s,t] \times B$, where $0 \leq s < t$, $A \in \mathcal{F}_s$ and B is a Borel set in $\mathbb{R}$. Denote by Λ_H the space of predictable processes g defined on $\mathbb{R}_+ \times \mathbb{R}$ such that almost surely $g \in \mathfrak{H}$ and $\mathbf{E}[\|g\|^2_{\mathfrak{H}}] < \infty$. Then, if $g \in \Lambda_H$, the Skorohod integral of g with respect to W coincides with the Itô integral defined in [9] and we have the isometry

$$\mathbf{E}\left[\left(\int_{\mathbb{R}_+}\int_{\mathbb{R}} g(s,x) W(ds,dx)\right)^2\right] = \mathbb{E}\|g\|^2_{\mathfrak{H}}\,. \tag{10}$$

Now we are ready to state the definition of the solution to Eq. (2). Denote by $p_t(x)$ the heat kernel on the real line related to $\frac{\kappa}{2}\Delta$. We denote by $*$ the convolution operation.

Definition 2.2 Let $u = \{u(t,x), 0 \le t \le T, x \in \mathbb{R}\}$ be a real-valued predictable stochastic process such that for all $t \in [0,T]$ and $x \in \mathbb{R}$ the process $\{p_{t-s}(x-y)u(s,y)\,\mathbf{1}_{[0,t]}(s), s \ge 0, y \in \mathbb{R}\}$ belongs to Λ_H. We say that u is a mild solution of (2) if for all $t \in [0,T]$ and $x \in \mathbb{R}$ we have

$$u(t,x) = p_t * u_0(x) + \int_0^t \int_{\mathbb{R}} p_{t-s}(x-y)u(s,y)W(ds,dy) \quad a.s., \tag{11}$$

where and in what follows the stochastic integral is always understood in the sense of Itô and coincides with the Skorohod integral defined by (6).

3 Existence and Uniqueness

In this section we prove the existence and uniqueness result for the solution to Eq. (2) by means of two different methods: one is via Fourier transform and the other is via chaos expansion.

3.1 Existence and Uniqueness via Fourier Transform

In this subsection we discuss the existence and uniqueness of Eq. (2) using techniques of Fourier analysis.

Let $\dot{H}_0^{\frac{1}{2}-H}$ be the set of functions $f \in L^2(\mathbb{R})$ such that $\int_{\mathbb{R}} |\mathcal{F}f(\xi)|^2 |\xi|^{1-2H} d\xi < \infty$. This space is the time independent analogue to the space $\mathfrak{H}_0$ introduced in Proposition 2.1. We know that $\dot{H}_0^{\frac{1}{2}-H}$ is not complete with the seminorm $\left[\int_{\mathbb{R}} |\mathcal{F}f(\xi)|^2 |\xi|^{1-2H} d\xi\right]^{\frac{1}{2}}$ (see [17]). However, it is not difficult to check that the space $\dot{H}_0^{\frac{1}{2}-H}$ is complete for the seminorm $\|f\|^2_{\mathcal{V}(H)} := \int_{\mathbb{R}} |\mathcal{F}f(\xi)|^2 (1 + |\xi|^{1-2H}) d\xi$.

In the next theorem we show the existence and uniqueness result assuming that the initial condition belongs to $\dot{H}_0^{\frac{1}{2}-H}$ and using estimates based on the Fourier transform in the space variable. To this purpose, we introduce the space $\mathcal{V}_T(H)$ as the completion of the set of elementary $\dot{H}_0^{\frac{1}{2}-H}$-valued stochastic processes

$$u(t) = \sum_{i=0}^{n-1} \mathbf{1}_{(t_i, t_{i+1}]}(t) u_i \,, \quad t \in [0,T] \,,$$

where $0 = t_0 < t_1 < \cdots < t_n = T$ is a partition of $[0, T]$ and $u_i \in \dot{H}_0^{\frac{1}{2}-H}$, with respect to the seminorm

$$\|u\|^2_{\mathcal{V}_T(H)} := \sup_{t\in[0,T]} \mathbf{E}\|u(t,\cdot)\|^2_{\mathcal{V}(H)}. \tag{12}$$

We now state a convolution lemma.

Proposition 3.1 *Consider a function $u_0 \in \dot{H}_0^{\frac{1}{2}-H}$ and $\frac{1}{4} < H < \frac{1}{2}$. For any $v \in \mathcal{V}_T(H)$ we set $\Gamma(v) = V$ in the following way:*

$$\Gamma(v) := V(t,x) = p_t * u_0(x) + \int_0^t \int_{\mathbb{R}} p_{t-s}(x-y)v(s,y)W(ds,dy), \quad t \in [0,T],\ x \in \mathbb{R}.$$

Then Γ is well-defined as a map from $\mathcal{V}_T(H)$ to $\mathcal{V}_T(H)$. Furthermore, there exist two positive constants c_1, c_2 such that the following estimate holds true on $[0, T]$:

$$\|V(t,\cdot)\|^2_{\mathcal{V}(H)} \le c_1 \|u_0\|^2_{\mathcal{V}(H)} + c_2 \int_0^t (t-s)^{2H-3/2}\|v(s,\cdot)\|^2_{\mathcal{V}(H)}\, ds\,. \tag{13}$$

Proof Let v be a process in $\mathcal{V}_T(H)$ and set $V = \Gamma(v)$. The stochastic integral appearing in the definition of $\Gamma(v)$ exists as an Itô (or Skorohod) integral, because the process $\{p_{t-s}(x-y)v(s,y), \mathbf{1}_{[0,t]}(s), s \ge 0, y \in \mathbb{R}\}$ is predictable and square integrable. We focus on the bound (13) for V.

Notice that the Fourier transform of V can be computed easily. Indeed, setting $v_0(t,x) = p_t * u_0(x)$ and invoking a stochastic version of Fubini's theorem, which can be easily proved in our framework, we get

$$\mathcal{F}V(t,\xi) = \mathcal{F}v_0(t,\xi) + \int_0^t \int_{\mathbb{R}} \left(\int_{\mathbb{R}} e^{ix\xi}\, p_{t-s}(x-y)\, dx \right) v(s,y)W(ds,dy)\,.$$

According to the expression of $\mathcal{F}p_t$, we obtain

$$\mathcal{F}V(t,\xi) = \mathcal{F}v_0(t,\xi) + \int_0^t \int_{\mathbb{R}} e^{-i\xi y} e^{-\frac{\kappa}{2}(t-s)\xi^2} v(s,y)W(ds,dy)\,.$$

We now evaluate the quantity $\mathbf{E}[\int_{\mathbb{R}} |\mathcal{F}V(t,\xi)|^2 |\xi|^{1-2H} d\xi]$ in the definition of $\|V\|_{\mathcal{V}_T(H)}$ given by (12). We thus write

$$\mathbf{E}\left[\int_{\mathbb{R}} |\mathcal{F}V(t,\xi)|^2 |\xi|^{1-2H} d\xi\right] \le 2\int_{\mathbb{R}} |\mathcal{F}v_0(t,\xi)|^2 |\xi|^{1-2H} d\xi$$
$$+ 2\int_{\mathbb{R}} \mathbf{E}\left[\left|\int_0^t \int_{\mathbb{R}} e^{-i\xi y} e^{-\frac{\kappa}{2}(t-s)\xi^2} v(s,y)W(ds,dy)\right|^2\right] |\xi|^{1-2H} d\xi := 2(I_1 + I_2)\,,$$

and we handle the terms I_1 and I_2 separately.

The term I_1 can be easily bounded by using that $u_0 \in \dot{H}_0^{\frac{1}{2}-H}$ and recalling $v_0 = p_t * u_0$. That is,

$$I_1 = \int_{\mathbb{R}} |\mathcal{F}u_0(\xi)|^2 e^{-\kappa t|\xi|^2} |\xi|^{1-2H} d\xi \leq C \|u_0\|^2_{\mathcal{V}(H)}.$$

We thus focus on the estimation of I_2, and we set $f_\xi(s,\eta) = e^{-i\xi\eta} e^{-\frac{\kappa}{2}(t-s)\xi^2} v(s,\eta)$. Applying the isometry property (10) we have:

$$\mathbf{E}\left[\Big| \int_0^t \int_{\mathbb{R}} e^{-i\xi y} e^{-\frac{\kappa}{2}(t-s)\xi^2} v(s,y) W(ds,dy)\Big|^2\right] = c_{1,H} \int_0^t \int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}_\eta f_\xi(s,\eta)|^2\right] |\eta|^{1-2H} ds d\eta,$$

where $\mathcal{F}_\eta$ is the Fourier transform with respect to η. It is obvious that the Fourier transform of $e^{-i\xi y} V(y)$ is $\mathcal{F}V(\eta+\xi)$. Thus we have

$$\begin{aligned} I_2 &= C \int_0^t \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\kappa(t-s)\xi^2} \mathbf{E}\left[|\mathcal{F}v(s,\eta+\xi)|^2\right] |\eta|^{1-2H} |\xi|^{1-2H} \, d\eta d\xi ds \\ &= C \int_0^t \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\kappa(t-s)\xi^2} \mathbf{E}\left[|\mathcal{F}v(s,\eta)|^2\right] |\eta-\xi|^{1-2H} |\xi|^{1-2H} \, d\eta d\xi ds \,. \end{aligned}$$

We now bound $|\eta-\xi|^{1-2H}$ by $|\eta|^{1-2H} + |\xi|^{1-2H}$, which yields $I_2 \leq I_{21} + I_{22}$ with:

$$\begin{aligned} I_{21} &= C \int_0^t \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\kappa(t-s)\xi^2} \mathbf{E}\left[|\mathcal{F}v(s,\eta)|^2\right] |\eta|^{1-2H} |\xi|^{1-2H} \, d\eta d\xi ds \\ I_{22} &= C \int_0^t \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\kappa(t-s)\xi^2} \mathbf{E}\left[|\mathcal{F}v(s,\eta)|^2\right] |\xi|^{2-4H} \, d\eta d\xi ds \,. \end{aligned}$$

Performing the change of variable $\xi \to (t-s)^{-1/2}\xi$ and then trivially bounding the integrals of the form $\int_{\mathbb{R}} |\xi|^\beta e^{-\kappa\xi^2} d\xi$ by constants, we end up with

$$\begin{aligned} I_{21} &\leq C \int_0^t (t-s)^{H-1} \int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}v(s,\eta)|^2\right] |\eta|^{1-2H} \, d\eta \, ds \\ I_{22} &\leq C \int_0^t (t-s)^{2H-3/2} \int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}v(s,\eta)|^2\right] d\eta \, ds. \end{aligned}$$

Observe that for $H \in (\frac{1}{4}, \frac{1}{2})$ the term $(t-s)^{2H-3/2}$ is more singular than $(t-s)^{H-1}$, but we still have $2H - \frac{3}{2} > -1$ (this is where we need to impose $H > 1/4$).

Summarizing our consideration up to now, we have thus obtained

$$\begin{aligned}&\int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}V(t,\xi)|^2\right]|\xi|^{1-2H}d\xi\\ &\leq C_{1,T}\,\|u_0\|^2_{\mathcal{V}(H)} + C_{2,T}\int_0^t (t-s)^{2H-3/2}\int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}v(s,\xi)|^2\right](1+|\xi|^{1-2H})\,d\xi\,ds,\end{aligned} \tag{14}$$

for two strictly positive constants $C_{1,T}, C_{2,T}$.

The term $\mathbf{E}[\int_{\mathbb{R}} |\mathcal{F}V(t,\xi)|^2 d\xi]$ in the definition of $\|V\|_{\mathcal{V}_T(H)}$ can be bounded with the same computations as above, and we find

$$\begin{aligned}&\int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}V(t,\xi)|^2\right]d\xi\\ &\leq C_{1,T}\,\|u_0\|^2_{\mathcal{V}(H)} + C_{2,T}\int_0^t (t-s)^{H-1}\int_{\mathbb{R}} \mathbf{E}\left[|\mathcal{F}v(s,\xi)|^2\right](1+|\xi|^{1-2H})\,d\eta\,ds\,.\end{aligned} \tag{15}$$

Hence, gathering our estimates (14) and (15), our bound (13) is easily obtained, which finishes the proof. □

As in the forthcoming general case, Proposition 3.1 is the key to the existence and uniqueness result for Eq. (2).

Theorem 3.2 *Suppose that u_0 is an element of $\dot{H}_0^{\frac{1}{2}-H}$ and $\frac{1}{4} < H < \frac{1}{2}$. Fix $T > 0$. Then there is a unique process u in the space $\mathcal{V}_T(H)$ such that for all $t \in [0,T]$,*

$$u(t,\cdot) = p_t * u_0 + \int_0^t \int_{\mathbb{R}} p_{t-s}(\cdot - y)u(s,y)W(ds,dy). \tag{16}$$

Proof The proof follows from the standard Picard iteration scheme, where we just set $u_{n+1} = \Gamma(u_n)$. Details are left to the reader for the sake of conciseness. □

3.2 Existence and Uniqueness via Chaos Expansions

Next, we provide another way to prove the existence and uniqueness of the solution to Eq. (2), by means of chaos expansions. This will enable us to obtain moment estimates. Before stating our main theorem in this direction, let us label an elementary lemma borrowed from [10] for further use.

Lemma 3.3 *For $m \geq 1$ let $\alpha \in (-1+\varepsilon, 1)^m$ with $\varepsilon > 0$ and set $|\alpha| = \sum_{i=1}^m \alpha_i$. For $t \in [0,T]$, the m-th dimensional simplex over $[0,t]$ is denoted by $T_m(t) =$*

$\{(r_1, r_2, \ldots, r_m) \in \mathbb{R}^m : 0 < r_1 < \cdots < r_m < t\}$. *Then there is a constant* $c > 0$ *such that*

$$J_m(t, \alpha) := \int_{T_m(t)} \prod_{i=1}^{m} (r_i - r_{i-1})^{\alpha_i} dr \leq \frac{c^m t^{|\alpha|+m}}{\Gamma(|\alpha| + m + 1)},$$

where by convention, $r_0 = 0$.

Let us now state a new existence and uniqueness theorem for our equation of interest (2).

Theorem 3.4 *Suppose that* $\frac{1}{4} < H < \frac{1}{2}$ *and that the initial condition* u_0 *satisfies*

$$\int_{\mathbb{R}} (1 + |\xi|^{\frac{1}{2}-H}) |\mathcal{F}u_0(\xi)| d\xi < \infty . \tag{17}$$

Then there exists a unique solution to Eq. (2)*, that is, there is a unique process* u *such that the Itô (or Skorohod) integral of the process* $\{p_{t-s}(x - y)u(s, y)\mathbf{1}_{[0,t]}(s), s \geq 0, y \in \mathbb{R}\}$ *exists for any* $(t, x) \in [0, T] \times \mathbb{R}$ *and relation* (11) *holds true.*

Remark 3.5

(i) The formulation of Theorem 3.4 yields the definition of our solution u for all $(t, x) \in [0, T] \times \mathbb{R}$. This is in contrast with Theorem 3.2 which gives a solution sitting in $\dot{H}_0^{\frac{1}{2}-H}$ for every value of t, and thus defined a.e. in x only.
(ii) Obviously a constant can be considered as a tempered distribution. Condition (17) is satisfied by constant functions.

Remark 3.6 In the later paper [12], the existence and uniqueness in Theorem 3.4 is obtained under a more general initial condition. Since the proof of Theorem 3.4 for condition (17) is easier and shorter, we present the proof as follows.

Proof of Theorem 3.4 Suppose that $u = \{u(t, x), t \geq 0, x \in \mathbb{R}^d\}$ is a solution to Eq. (11) in Λ_H. Then according to (7), for any fixed (t, x) the random variable $u(t, x)$ admits the following Wiener chaos expansion

$$u(t, x) = \sum_{n=0}^{\infty} I_n(f_n(\cdot, t, x)) , \tag{18}$$

where for each (t, x), $f_n(\cdot, t, x)$ is a symmetric element in $\mathfrak{H}^{\otimes n}$. Hence, thanks to (9) and using an iteration procedure, one can find an explicit formula for the kernels f_n for $n \geq 1$. Indeed, we have:

$$\begin{aligned} &f_n(s_1, x_1, \ldots, s_n, x_n, t, x) \\ &= \frac{1}{n!} p_{t-s_{\sigma(n)}}(x - x_{\sigma(n)}) \cdots p_{s_{\sigma(2)}-s_{\sigma(1)}}(x_{\sigma(2)} - x_{\sigma(1)}) p_{s_{\sigma(1)}} u_0(x_{\sigma(1)}) , \end{aligned} \tag{19}$$

where σ denotes the permutation of $\{1, 2, \dots, n\}$ such that $0 < s_{\sigma(1)} < \cdots < s_{\sigma(n)} < t$ (see, for instance, formula (4.4) in [8] or formula (3.3) in [10]). Then, to show the existence and uniqueness of the solution it suffices to prove that for all (t, x) we have

$$\sum_{n=0}^{\infty} n! \| f_n(\cdot, t, x) \|^2_{\mathfrak{H}^{\otimes n}} < \infty . \tag{20}$$

The remainder of the proof is devoted to prove relation (20).

Starting from relation (19), some elementary Fourier computations show that

$$\mathcal{F} f_n(s_1, \xi_1, \dots, s_n, \xi_n, t, x) = \frac{c_H^n}{n!} \int_{\mathbb{R}} \prod_{i=1}^{n} e^{-\frac{\kappa}{2}(s_{\sigma(i+1)} - s_{\sigma(i)})|\xi_{\sigma(i)} + \cdots + \xi_{\sigma(1)} - \zeta|^2}$$
$$\times e^{-ix(\xi_{\sigma(n)} + \cdots + \xi_{\sigma(1)} - \zeta)} \mathcal{F} u_0(\zeta) e^{-\frac{\kappa s_{\sigma(1)} |\zeta|^2}{2}} d\zeta,$$

where we have set $s_{\sigma(n+1)} = t$. Hence, owing to formula (5) for the norm in $\mathfrak{H}$ (in its Fourier mode version), we have

$$n! \| f_n(\cdot, t, x) \|^2_{\mathfrak{H}^{\otimes n}} = \frac{c_H^{2n}}{n!} \int_{[0,t]^n} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}} \prod_{i=1}^{n} e^{-\frac{\kappa}{2}(s_{\sigma(i+1)} - s_{\sigma(i)})|\xi_i + \cdots + \xi_1 - \zeta|^2} e^{-ix(\xi_{\sigma(n)} + \cdots + \xi_{\sigma(1)} - \zeta)} \right.$$
$$\left. \mathcal{F} u_0(\zeta) e^{-\frac{\kappa s_{\sigma(1)} |\zeta|^2}{2}} d\zeta \right|^2 \times \prod_{i=1}^{n} |\xi_i|^{1-2H} d\xi ds , \tag{21}$$

where $d\xi$ denotes $d\xi_1 \cdots d\xi_n$ and similarly for ds. Then using the change of variable $\xi_i + \cdots + \xi_1 = \eta_i$, for all $i = 1, 2, \dots, n$ and a linearization of the above expression, we obtain

$$n! \| f_n(\cdot, t, x) \|^2_{\mathfrak{H}^{\otimes n}} = \frac{c_H^{2n}}{n!} \int_{[0,t]^n} \int_{\mathbb{R}^n} \int_{\mathbb{R}^2} \prod_{i=1}^{n} e^{-\frac{\kappa}{2}(s_{\sigma(i+1)} - s_{\sigma(i)})(|\eta_i - \zeta|^2 + |\eta_i - \zeta'|^2)} \mathcal{F} u_0(\zeta) \overline{\mathcal{F} u_0(\zeta')}$$
$$\times e^{ix(\zeta - \zeta')} e^{-\frac{\kappa s_{\sigma(1)}(|\zeta|^2 + |\zeta'|^2)}{2}} \prod_{i=1}^{n} |\eta_i - \eta_{i-1}|^{1-2H} d\zeta d\zeta' d\eta ds ,$$

where we have set $\eta_0 = 0$. Then we use Cauchy-Schwarz inequality and bound the term $\exp(-\kappa s_{\sigma(1)}(|\zeta|^2 + |\zeta'|^2)/2)$ by 1 to get

$$n! \| f_n(\cdot, t, x) \|^2_{\mathfrak{H}^{\otimes n}} \le \frac{c_H^{2n}}{n!} \int_{\mathbb{R}^2} \left(\int_{[0,t]^n} \int_{\mathbb{R}^n} \prod_{i=1}^{n} e^{-\kappa(s_{\sigma(i+1)} - s_{\sigma(i)})|\eta_i - \zeta|^2} \prod_{i=1}^{n} |\eta_i - \eta_{i-1}|^{1-2H} d\eta ds \right)^{\frac{1}{2}}$$
$$\times \left(\int_{[0,t]^n} \int_{\mathbb{R}^n} \prod_{i=1}^{n} e^{-\kappa(s_{\sigma(i+1)} - s_{\sigma(i)})|\eta_i - \zeta'|^2} \prod_{i=1}^{n} |\eta_i - \eta_{i-1}|^{1-2H} d\eta ds \right)^{\frac{1}{2}} \left| \mathcal{F} u_0(\zeta) \right| \left| \mathcal{F} u_0(\zeta') \right| d\zeta d\zeta' .$$

Arranging the integrals again, performing the change of variables $\eta_i := \eta_i - \zeta$ and invoking the trivial bound $|\eta_i - \eta_{i-1}|^{1-2H} \le |\eta_{i-1}|^{1-2H} + |\eta_i|^{1-2H}$, this yields

$$n!\|f_n(\cdot,t,x)\|^2_{\mathfrak{H}^{\otimes n}} \le \frac{c_H^{2n}}{n!}\left(\int_{\mathbb{R}} L_{n,t}^{\frac{1}{2}}(\zeta)\left|\mathcal{F}u_0(\zeta)\right|d\zeta\right)^2, \tag{22}$$

where $L_{n,t}(\zeta)$ is

$$\int_{[0,t]^n}\int_{\mathbb{R}^n}\prod_{i=1}^n e^{-\kappa(s_{\sigma(i+1)}-s_{\sigma(i)})|\eta_i|^2}(|\zeta|^{1-2H}+|\eta_1|^{1-2H})\times\prod_{i=2}^n(|\eta_i|^{1-2H}+|\eta_{i-1}|^{1-2H})d\eta ds.$$

Let us expand the product $\prod_{i=2}^n(|\eta_i|^{1-2H}+|\eta_{i-1}|^{1-2H})$ in the integral defining $L_{n,t}(\zeta)$. We obtain an expression of the form $\sum_{\alpha\in D_n}\prod_{i=1}^n|\eta_i|^{\alpha_i}$, where D_n is a subset of multi-indices of length $n-1$. The complete description of D_n is omitted for the sake of conciseness, and we will just use the following facts: $\mathrm{Card}(D_n) = 2^{n-1}$ and for any $\alpha \in D_n$ we have

$$|\alpha| \equiv \sum_{i=1}^n \alpha_i = (n-1)(1-2H), \quad \text{and} \quad \alpha_i \in \{0, 1-2H, 2(1-2H)\}, \quad i = 1,\dots,n.$$

This simple expansion yields the following bound

$$\begin{aligned} L_{n,t}(\zeta) \le |\zeta|^{1-2H}&\sum_{\alpha\in D_n}\int_{[0,t]^n}\int_{\mathbb{R}^n}\prod_{i=1}^n e^{-\kappa(s_{\sigma(i+1)}-s_{\sigma(i)})|\eta_i|^2}\prod_{i=1}^n|\eta_i|^{\alpha_i}d\eta ds \\ &+\sum_{\alpha\in D_n}\int_{[0,t]^n}\int_{\mathbb{R}^n}\prod_{i=1}^n e^{-\kappa(s_{\sigma(i+1)}-s_{\sigma(i)})|\eta_i|^2}|\eta_1|^{1-2H}\prod_{i=1}^n|\eta_i|^{\alpha_i}d\eta ds\,. \end{aligned}$$

Perform the change of variable $\xi_i = (\kappa(s_{\sigma(i+1)} - s_{\sigma(i)}))^{1/2}\eta_i$ in the above integral, and notice that $\int_{\mathbb{R}} e^{-\xi^2}|\xi|^{\alpha_i}d\xi$ is bounded by a constant for $\alpha_i > -1$. Changing the integral over $[0,t]^n$ into an integral over the simplex, we get

$$\begin{aligned} L_{n,t}(\zeta) \le C|\zeta|^{1-2H}&n!c_H^n\sum_{\alpha\in D_n}\int_{T_n(t)}\prod_{i=1}^n(\kappa(s_{i+1}-s_i))^{-\frac{1}{2}(1+\alpha_i)}ds. \\ &+Cn!c_H^n\sum_{\alpha\in D_n}\int_{T_n(t)}(\kappa(s_2-s_1))^{-\frac{2-2H+\alpha_1}{2}}\prod_{i=2}^n(\kappa(s_{i+1}-s_i))^{-\frac{1}{2}(1+\alpha_i)}ds. \end{aligned}$$

We observe that whenever $\frac{1}{4} < H < \frac{1}{2}$, we have $\frac{1}{2}(1+\alpha_i) < 1$ for all $i = 2,\dots n$, and it is easy to see that α_1 is at most $1-2H$ so $\frac{1}{2}(2-2H+\alpha_1) < 1$. Condition $H > 1/4$ comes from the requirement that when $\alpha_1 = 1-2H$, we

need $\frac{1}{2}(2-2H+\alpha_1)=\frac{1}{2}(3-4H)<1$. Thanks to Lemma 3.3 and recalling that $\sum_{i=1}^n \alpha_i=(n-1)(1-2H)$ for all $\alpha\in D_n$, we thus conclude that

$$L_{n,t}(\zeta)\le \frac{C(1+t^{\frac{1}{2}-H}\kappa^{\frac{1}{2}-H}|\zeta|^{1-2H})n!c^n c_H^n t^{nH}\kappa^{nH-n}}{\Gamma(nH+1)}.$$

Plugging this expression into (22), we end up with

$$n!\|f_n(\cdot,t,x)\|^2_{\mathfrak{H}^{\otimes n}}\le \frac{Cc_H^n c^n t^{nH}\kappa^{nH-n}}{\Gamma(nH+1)}\left(\int_{\mathbb{R}}(1+t^{\frac{1}{2}-H}\kappa^{\frac{1}{2}-H}|\zeta|^{\frac{1}{2}-H})\left|\mathcal{F}u_0(\zeta)\right|d\zeta\right)^2. \tag{23}$$

The proof of (20) is now easily completed thanks to the asymptotic behavior of the Gamma function and our assumption of u_0. This finishes the existence and uniqueness proof. □

4 Moment Formula and Bounds

In this section we derive the Feynman-Kac formula for the moments of the solution to Eq. (2) and the upper and lower bounds for the moments of the solution to Eq. (2) which allow us to conclude on the intermittency of the solution. We proceed by first getting an approximation result for u, and then deriving the upper and lower bounds for the approximation.

4.1 Approximation of the Solution

The approximation of the solution we consider is based on the following approximation of the noise W. For each $\varepsilon>0$ and $\varphi\in\mathfrak{H}$, we define

$$W_\varepsilon(\varphi)=\int_0^\infty\int_{\mathbb{R}}[\rho_\varepsilon*\varphi](s,x)W(ds,dy)=\int_0^\infty\int_{\mathbb{R}}\int_{\mathbb{R}}\varphi(s,x)\rho_\varepsilon(x-y)W(ds,dy)dx, \tag{24}$$

where $\rho_\varepsilon(x)=(2\pi\varepsilon)^{-\frac{1}{2}}e^{-x^2/(2\varepsilon)}$. Notice that the covariance of W_ε can be read (either in Fourier or direct coordinates) as:

$$\begin{aligned}\mathbf{E}\left[W_\varepsilon(\varphi)W_\varepsilon(\psi)\right]&=c_{1,H}\int_0^\infty\int_{\mathbb{R}}\mathcal{F}\varphi(s,\xi)\,\overline{\mathcal{F}\psi(s,\xi)}\,e^{-\varepsilon|\xi|^2}|\xi|^{1-2H}d\xi ds \qquad (25)\\ &=c_{1,H}\int_0^\infty\int_{\mathbb{R}}\int_{\mathbb{R}}\varphi(s,x)f_\varepsilon(x-y)\psi(s,y)\,dxdyds,\end{aligned}$$

for every φ, ψ in $\mathfrak{H}$, where f_ε is given by $f_\varepsilon(x) = \mathcal{F}^{-1}(e^{-\varepsilon|\xi|^2}|\xi|^{1-2H})$. In other words, W_ε is still a white noise in time but its space covariance is now given by f_ε. Note that f_ε is a real positive-definite function, but is not necessarily positive. The noise W_ε induces an approximation to the mild formulation of Eq. (2), namely

$$u_\varepsilon(t,x) = p_t * u_0(x) + \int_0^t \int_{\mathbb{R}} p_{t-s}(x-y) u_\varepsilon(s,y)\, W_\varepsilon(ds,dy), \tag{26}$$

where the integral is understood (as in Sect. 3.1) in the Itô sense. We will start by a formula for the moments of u_ε.

Proposition 4.1 *Let W_ε be the noise defined by (24), and assume $\frac{1}{4} < H < \frac{1}{2}$. Assume u_0 is such that $\int_{\mathbb{R}}(1+|\xi|^{\frac{1}{2}-H})|\mathcal{F}u_0(\xi)|d\xi < \infty$. Then*

(i) *Equation (26) admits a unique solution.*
(ii) *For any integer $n \geq 2$ and $(t,x) \in [0,T] \times \mathbb{R}$, we have*

$$\boldsymbol{E}\left[u_\varepsilon^n(t,x)\right] = \boldsymbol{E}_B\left[\prod_{j=1}^n u_0(x + B_{\kappa t}^j) \exp\left(c_{1,H} \sum_{1\leq j\neq k\leq n} V_{t,x}^{\varepsilon,j,k}\right)\right], \tag{27}$$

with

$$V_{t,x}^{\varepsilon,j,k} = \int_0^t f_\varepsilon(B_{\kappa r}^j - B_{\kappa r}^k)dr = \int_0^t \int_{\mathbb{R}} e^{-\varepsilon|\xi|^2}|\xi|^{1-2H} e^{i\xi(B_{\kappa r}^j - B_{\kappa r}^k)}\, d\xi dr. \tag{28}$$

In formula (28), $\{B^j; j = 1,\ldots,n\}$ is a family of n independent standard Brownian motions which are also independent of W and $\boldsymbol{E}_B$ denotes the expected value with respect to the randomness in B only.

(iii) *The quantity $\boldsymbol{E}[u_\varepsilon^n(t,x)]$ is uniformly bounded in ε. More generally, for any $a > 0$ we have*

$$\sup_{\varepsilon>0} \boldsymbol{E}_B\left[\exp\left(a \sum_{1\leq j\neq k\leq n} V_{t,x}^{\varepsilon,j,k}\right)\right] \equiv c_a < \infty.$$

Proof The proof of item (i) is almost identical to the proof of Theorem 3.4, and is omitted for the sake of conciseness. Moreover, in the proof of (ii) and (iii), we may take $u_0(x) \equiv 1$ for simplicity.

In order to check item (ii), set

$$A_{t,x}^\varepsilon(r,y) = \rho_\varepsilon(B_{\kappa(t-r)}^x - y), \quad \text{and} \quad \alpha_{t,x}^\varepsilon = \|A_{t,x}^\varepsilon\|_{\mathfrak{H}}^2. \tag{29}$$

Then one can prove, similarly to Proposition 5.2 in [8], that u_ε admits a Feynman-Kac representation of the form

$$u_\varepsilon(t,x) = \mathbf{E}_B\left[\exp\left(W(A^\varepsilon_{t,x}) - \frac{1}{2}\alpha^\varepsilon_{t,x}\right)\right]. \tag{30}$$

Now fix an integer $n \geq 2$. According to (30) we have

$$\mathbf{E}\left[u^n_\varepsilon(t,x)\right] = \mathbf{E}_W\left[\prod_{j=1}^n \mathbf{E}_B\left[\exp\left(W(A^{\varepsilon,B^j}_{t,x}) - \frac{1}{2}\alpha^{\varepsilon,B^j}_{t,x}\right)\right]\right],$$

where for any $j = 1,\dots,n$, $A^{\varepsilon,B^j}_{t,x}$ and $\alpha^{\varepsilon,B^j}_{t,x}$ are evaluations of (29) using the Brownian motion B^j. Therefore, since $W(A^{\varepsilon,B^j}_{t,x})$ is a Gaussian random variable conditionally on B, we obtain

$$\begin{aligned}
\mathbf{E}\left[u^n_\varepsilon(t,x)\right] &= \mathbf{E}_B\left[\exp\left(\frac{1}{2}\|\sum_{j=1}^n A^{\varepsilon,B^j}_{t,x}\|^2_{\mathfrak{H}} - \frac{1}{2}\sum_{j=1}^n \alpha^{\varepsilon,B^j}_{t,x}\right)\right] \\
&= \mathbf{E}_B\left[\exp\left(\frac{1}{2}\|\sum_{j=1}^n A^{\varepsilon,B^j}_{t,x}\|^2_{\mathfrak{H}} - \frac{1}{2}\sum_{j=1}^n \|A^{\varepsilon,B^j}_{t,x}\|^2_{\mathfrak{H}}\right)\right] \\
&= \mathbf{E}_B\left[\exp\left(\sum_{1\leq i<j\leq n}\langle A^{\varepsilon,B^i}_{t,x}, A^{\varepsilon,B^j}_{t,x}\rangle_{\mathfrak{H}}\right)\right].
\end{aligned}$$

The evaluation of $\langle A^{\varepsilon,B^i}_{t,x}, A^{\varepsilon,B^j}_{t,x}\rangle_{\mathfrak{H}}$ easily yields our claim (27), the last details being left to the patient reader.

Let us now prove item (iii), namely

$$\sup_{\varepsilon>0}\sup_{t\in[0,T],x\in\mathbb{R}} \mathbf{E}\left[u^n_\varepsilon(t,x)\right] < \infty. \tag{31}$$

To this aim, notice first from the expression (27) that $\mathbf{E}\left[u^n_\varepsilon(t,x)\right]$ does not depend on $x \in \mathbb{R}$ (since $u_0(x) \equiv 1$) so that the $\sup_{t\in[0,T],x\in\mathbb{R}}$ in (31) can be reduced to a sup in t only. Next, still resorting to formula (27), it is readily seen that it suffices to show that for two independent Brownian motions B and $\tilde{B}$, we have

$$\sup_{\varepsilon>0,t\in[0,T]} \mathbf{E}_B\left[\exp\left(c\,F^\varepsilon_t\right)\right] < \infty, \quad \text{with} \quad F^\varepsilon_t \equiv \int_0^t\int_{\mathbb{R}} e^{-\varepsilon|\xi|^2}|\xi|^{1-2H} e^{i\xi(B_{\kappa r}-\tilde{B}_{\kappa r})}d\xi dr, \tag{32}$$

for any positive constant c. In order to prove (32), we expand the exponential and write:

$$\mathbf{E}_B\left[\exp(c\,F_t^\varepsilon)\right] = \sum_{l=0}^{\infty} \frac{\mathbf{E}_B\left[(c\,F_t^\varepsilon)^l\right]}{l!}. \tag{33}$$

Next, we have

$$\mathbf{E}_B\left[\left(F_t^\varepsilon\right)^l\right] = \mathbf{E}_B\left[\int_{[0,t]^l}\int_{\mathbb{R}^l}\prod_{j=1}^{l} e^{-i\xi_j(B_{\kappa r_j}-\tilde{B}_{\kappa r_j})-\varepsilon|\xi_j|^2}|\xi_j|^{1-2H}d\xi dr\right]$$

$$\leq \int_{[0,t]^l}\int_{\mathbb{R}^l}\prod_{j=1}^{l} e^{-\kappa(t-r_{\sigma(j)})|\xi_j+\cdots+\xi_1|^2}\,|\xi_j|^{1-2H}\,d\xi dr\,,$$

where σ is the permutation on $\{1,2,\ldots,l\}$ such that $t \geq r_{\sigma(l)} \geq \cdots \geq r_{\sigma(1)}$. We have thus gone back to an expression which is very similar to (21). We now proceed as in the proof of Theorem 3.4 to show that (31) holds true from Eq. (33). □

Starting from Proposition 4.1, let us take limits in order to get the moment formula for the solution u to Eq. (2).

Theorem 4.2 *Assume* $\frac{1}{4} < H < \frac{1}{2}$ *and consider* $n \geq 1$, $j,k \in \{1,\ldots,n\}$ *with* $j \neq k$. *For* $(t,x) \in [0,T]\times\mathbb{R}$, *denote by* $V_{t,x}^{j,k}$ *the limit in* $L^2(\Omega)$ *as* $\varepsilon \to 0$ *of*

$$V_{t,x}^{\varepsilon,j,k} = \int_0^t\int_{\mathbb{R}} e^{-\varepsilon|\xi|^2}|\xi|^{1-2H}e^{i\xi(B_{\kappa r}^j - B_{\kappa r}^k)}d\xi dr.$$

Then $\boldsymbol{E}\left[u_\varepsilon^n(t,x)\right]$ *converges as* $\varepsilon \to 0$ *to* $\boldsymbol{E}[u^n(t,x)]$, *which is given by*

$$\boldsymbol{E}[u^n(t,x)] = \boldsymbol{E}_B\left[\prod_{j=1}^{n} u_0(B_{\kappa t}^j + x)\exp\left(c_{1,H}\sum_{1\leq j\neq k\leq n} V_{t,x}^{j,k}\right)\right]. \tag{34}$$

We note that in a recent paper [12], the moment formula for general covariance function is obtained. However we present the proof here for the sake of completeness.

Proof As in Proposition 4.1, we will prove the theorem for $u_0 \equiv 1$ for simplicity. For any $p \geq 1$ and $1 \leq j < k \leq n$, we can easily prove that $V_{t,x}^{\varepsilon,j,k}$ converges in $L^p(\Omega)$ to $V_{t,x}^{j,k}$ defined by

$$V_{t,x}^{j,k} = \int_0^t\int_{\mathbb{R}} |\xi|^{1-2H}e^{i\xi(B_{\kappa r}^j - B_{\kappa r}^k)}d\xi dr. \tag{35}$$

Indeed, this is due to the fact that $e^{-\varepsilon|\xi|^2}|\xi|^{1-2H}e^{i\xi(B^j_{kr}-B^k_{kr})}$ converges to $|\xi|^{1-2H}e^{i\xi(B^j_{kr}-B^k_{kr})}$ in the $d\xi\otimes dr\otimes d\mathbf{P}$ sense, plus standard uniform integrability arguments. Now, taking into account relation (27), Proposition 4.1 (iii), the fact that $V^{\varepsilon,j,k}_{t,x}$ converges to $V^{j,k}_{t,x}$ in $L^2(\Omega)$ as $\varepsilon\to 0$, and the inequality $|e^x-e^y|\le(e^x+e^y)|x-y|$, we obtain

$$\begin{aligned}
&\mathbf{E}_B\left|\exp\left(c_{1,H}\sum_{1\le j\ne k\le n}V^{\epsilon,j,k}_{t,x}\right)-\exp\left(c_{1,H}\sum_{1\le j\ne k\le n}V^{j,k}_{t,x}\right)\right|\\
&\le\sup_{\epsilon>0}2\left(\mathbf{E}_B\left|\exp\left(2c_{1,H}\sum_{1\le j\ne k\le n}V^{\epsilon,j,k}_{t,x}\right)+\exp\left(2c_{1,H}\sum_{1\le j\ne k\le n}V^{j,k}_{t,x}\right)\right|^2\right)^{\frac12}\\
&\quad\times\left(\mathbf{E}_B\left|c_{1,H}\sum_{1\le j\ne k\le n}V^{\epsilon,j,k}_{t,x}-c_{1,H}\sum_{1\le j\ne k\le n}V^{j,k}_{t,x}\right|^2\right)^{\frac12},
\end{aligned}$$

which implies

$$\begin{aligned}
\lim_{\varepsilon\to0}\mathbf{E}\left[u^n_\varepsilon(t,x)\right]&=\lim_{\varepsilon\to0}\mathbf{E}_B\left[\exp\left(c_{1,H}\sum_{1\le j\ne k\le n}V^{\varepsilon,j,k}_{t,x}\right)\right]\\
&=\mathbf{E}_B\left[\exp\left(c_{1,H}\sum_{1\le j\ne k\le n}V^{j,k}_{t,x}\right)\right].
\end{aligned}\tag{36}$$

To end the proof, let us now identify the right hand side of (36) with $\mathbf{E}[u^n(t,x)]$, where u is the solution to Eq. (2). For $\varepsilon,\varepsilon'>0$ we write

$$\mathbf{E}\left[u_\varepsilon(t,x)\,u_{\varepsilon'}(t,x)\right]=\mathbf{E}_B\left[\exp\left(\langle A^{\varepsilon,B^1}_{t,x},A^{\varepsilon',B^2}_{t,x}\rangle_{\mathfrak{H}}\right)\right],$$

where we recall that $A^{\varepsilon,B}_{t,x}$ is defined by relation (29). As for (36) we can show that the above $\mathbf{E}\left[u_\varepsilon(t,x)\,u_{\varepsilon'}(t,x)\right]$ converges as ε,ε' tend to zero. So, $u_\varepsilon(t,x)$ converges in L^2 to some limit $v(t,x)$. For any positive integer k notice the identity

$$\mathbf{E}|u_\varepsilon(t,x)-u_{\varepsilon'}(t,x)|^{2k}=\sum_{j=0}^{2k}\frac{(-1)^j(2k)!}{j!(2k-j)!}\mathbb{E}\left[u_\varepsilon(t,x)^{2k-j}u_{\varepsilon'}(t,x)^j\right].\tag{37}$$

We can find the limit of $\mathbb{E}\left[u_\varepsilon(t,x)^{2k-j}u_{\varepsilon'}(t,x)^j\right]$ and then show that (37) converges to 0 as ε and ε' tend to 0. It is now clear that $u_\varepsilon(t,x)$ converges to $v(t,x)$ in L^p for all $p \geq 1$. Moreover, $\mathbf{E}[v^n(t,x)]$ is equal to the right hand side of (36). Finally, for any smooth random variable F which is a linear combination of $W(\mathbf{1}_{[a,b]}(s)\varphi(x))$, where φ is a C^∞ function with compact support, using the duality relation (6), we have

$$\mathbf{E}\left[Fu_\varepsilon(t,x)\right] = \mathbf{E}\left[F\right] + \mathbf{E}\left[\langle Y^\varepsilon, DF\rangle_{\mathfrak{H}}\right], \tag{38}$$

where

$$Y^{t,x}(s,z) = \left(\int_{\mathbb{R}} p_{t-s}(x-y)\, p_\varepsilon(y-z)u_\varepsilon(s,y)\,dy\right)\mathbf{1}_{[0,t]}(s).$$

Letting ε tend to zero in Eq. (38), after some easy calculation we get

$$\mathbf{E}[Fv_{t,x}] = \mathbf{E}[F] + \mathbf{E}\left[\langle DF, vp_{t-\cdot}(x-\cdot)\rangle_{\mathfrak{H}}\right].$$

This equation is valid for any $F \in \mathbb{D}^{1,2}$ by approximation. So the above equation implies that the process v is the solution of Eq. (2), and by the uniqueness of the solution we have $v = u$. □

4.2 Intermittency Estimates

In this subsection we prove some upper and lower bounds on the moments of the solution which entail the intermittency phenomenon.

Theorem 4.3 *Let $\frac{1}{4} < H < \frac{1}{2}$, and consider the solution u to Eq. (2). For simplicity we assume that the initial condition is $u_0(x) \equiv 1$. Let $n \geq 2$ be an integer, $x \in \mathbb{R}$ and $t \geq 0$. Then there exist some positive constants c_1, c_2, c_3 independent of n, t and κ with $0 < c_1 < c_2 < \infty$ satisfying*

$$\exp(c_1 n^{1+\frac{1}{H}}\kappa^{1-\frac{1}{H}}t) \leq \boldsymbol{E}\left[u^n(t,x)\right] \leq c_3 \exp\left(c_2 n^{1+\frac{1}{H}}\kappa^{1-\frac{1}{H}}t\right). \tag{39}$$

Remark 4.4 It is interesting to point out from the above inequalities that when $\kappa \downarrow 0$, the moments of u go to infinity. This is because the equation $\frac{\partial u}{\partial t} = u\,\dot{W}$ has no classical solution due to the singularity of the noise $\dot{W}$ in spatial variable x.

Proof of Theorem 4.3 We divide this proof into upper and lower bound estimates.

Step 1: Upper bound. Recall from Eq. (18) that for $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$, $u(t,x)$ can be written as: $u(t,x) = \sum_{m=0}^{\infty} I_m(f_m(\cdot,t,x))$. Moreover, as a consequence of the hypercontractivity property on a fixed chaos we have (see [16, p. 62])

$$\|I_m(f_m(\cdot,t,x))\|_{L^n(\Omega)} \le (n-1)^{\frac{m}{2}} \|I_m(f_m(\cdot,t,x))\|_{L^2(\Omega)},$$

and substituting the above right hand side by the bound (23), we end up with

$$\|I_m(f_m(\cdot,t,x))\|_{L^n(\Omega)} \le n^{\frac{m}{2}} \|I_m(f_m(\cdot,t,x))\|_{L^2(\Omega)} \le \frac{c^{\frac{n}{2}} n^{\frac{m}{2}} t^{\frac{mH}{2}} \kappa^{\frac{Hm-m}{2}}}{\Gamma(mH/2+1)}.$$

Therefore from by the asymptotic bound of Mittag-Leffler function $\sum_{n\ge 0} x^n/\Gamma(\alpha n+1) \le c_1 \exp(c_2 x^{1/a})$ (see [14], Formula (1.8.10)), we get:

$$\|u(t,x)\|_{L^n(\Omega)} \le \sum_{m=0}^{\infty} \|J_m(t,x)\|_{L^n(\Omega)} \le \sum_{m=0}^{\infty} \frac{c^{\frac{m}{2}} n^{\frac{m}{2}} t^{\frac{mH}{2}} \kappa^{\frac{Hm-m}{2}}}{\left(\Gamma(mH+1)\right)^{\frac{1}{2}}} \le c_1 \exp\left(c_2 t n^{\frac{1}{H}} \kappa^{\frac{H-1}{H}}\right),$$

from which the upper bound in our theorem is easily deduced.
Step 2: Lower bound for u_ε. For the lower bound, we start from the moment formula (27) for the approximate solution, and write

$$\mathbf{E}\left[u_\varepsilon^n(t,x)\right] = \mathbf{E}_B\left[\exp\left(c_{1,H}\left[\int_0^t \int_{\mathbb{R}} e^{-\varepsilon|\xi|^2} \left|\sum_{j=1}^n e^{-iB_{kr}^j \xi}\right|^2 |\xi|^{1-2H} d\xi dr - nt \int_{\mathbb{R}} e^{-\varepsilon|\xi|^2} |\xi|^{1-2H} d\xi\right]\right)\right].$$

In order to estimate the expression above, notice first that the obvious change of variable $\lambda = \varepsilon^{1/2}\xi$ yields $\int_{\mathbb{R}} e^{-\varepsilon|\xi|^2} |\xi|^{1-2H} d\xi = C\varepsilon^{-(1-H)}$ for some constant C. Now for an additional arbitrary parameter $\eta > 0$, consider the set

$$A_\eta = \left\{\omega;\ \sup_{1\le j\le n} \sup_{0\le r\le t} |B_{kr}^j(\omega)| \le \frac{\pi}{3\eta}\right\}.$$

Observe that classical small balls inequalities for a Brownian motion (see (1.3) in [15]) yield $\mathbf{P}(A_\eta) \ge c_1 e^{-c_2 \eta^2 n\kappa t}$ for a large enough η. In addition, if we assume that A_η is realized and $|\xi| \le \eta$, some elementary trigonometric identities show that

the following deterministic bound hold true: $|\sum_{j=1}^n e^{-iB_{kr}^j\xi}| \geq \frac{n}{2}$. Gathering those considerations, we thus get

$$\mathbf{E}\left[u_\varepsilon^n(t,x)\right] \geq \exp\left(c_1 n^2 \int_0^t \int_0^\eta e^{-\varepsilon|\xi|^2}|\xi|^{1-2H} d\xi dr - c_2 nt\varepsilon^{H-1}\right) \mathbf{P}\left(A_\eta\right)$$
$$\geq C \exp\left(c_1 n^2 t\varepsilon^{-(1-H)} \int_0^{\varepsilon^{1/2}\eta} e^{-|\xi|^2}|\xi|^{1-2H} d\xi - c_2 nt\varepsilon^{-(1-H)} - c_3 n\kappa t\eta^2\right).$$

We now choose the parameter η such that $\kappa\eta^2 = \varepsilon^{-(1-H)}$, which means in particular that $\eta \to \infty$ as $\varepsilon \to 0$. It is then easily seen that $\int_0^{\varepsilon^{1/2}\eta} e^{-|\xi|^2}|\xi|^{1-2H} d\xi$ is of order $\varepsilon^{H(1-H)}$ in this regime, and some elementary algebraic manipulations entail

$$\mathbf{E}\left[u_\varepsilon^n(t,x)\right] \geq C \exp\left(c_1 n^2 t\kappa^{H-1}\varepsilon^{-(1-H)^2} - c_2 nt\varepsilon^{-(1-H)}\right) \geq C \exp\left(c_3 t\kappa^{1-\frac{1}{H}} n^{1+\frac{1}{H}}\right),$$

where the last inequality is obtained by choosing $\varepsilon^{-(1-H)} = c\,\kappa^{\frac{H-1}{H}} n^{\frac{1}{H}}$ in order to optimize the second expression. We have thus reached the desired lower bound in (39) for the approximation u^ε in the regime $\varepsilon = c\,\kappa^{\frac{1}{H}} n^{-\frac{1}{H(1-H)}}$.

Step 3: Lower bound for u. To complete the proof, we need to show that for all sufficiently small ε, $\mathbf{E}\left[u_\varepsilon^n(t,x)\right] \leq \mathbf{E}[u^n(t,x)]$. We thus start from Eq. (27) and use the series expansion of the exponential function as in (33). We get

$$\mathbf{E}\left[u_\varepsilon^n(t,x)\right] = \sum_{m=0}^\infty \frac{c_{1,H}^m}{m!} \mathbf{E}_B\left[\left(\sum_{1\leq j\neq k\leq n} V_{t,x}^{\varepsilon,j,k}\right)^m\right], \tag{40}$$

where we recall that $V_{t,x}^{\varepsilon,j,k}$ is defined by (28). Furthermore, expanding the mth power above, we have

$$\mathbf{E}_B\left[\left(\sum_{1\leq j\neq k\leq n} V_{t,x}^{\varepsilon,j,k}\right)^m\right] = \sum_{\alpha\in K_{n,m}} \int_{[0,t]^m} \int_{\mathbb{R}^m} e^{-\varepsilon\sum_{l=1}^m |\xi_l|^2} \mathbf{E}_B\left[e^{iB^\alpha(\xi)}\right] \prod_{l=1}^m |\xi_l|^{1-2H} d\xi dr\,,$$

where $K_{n,m}$ is a set of multi-indices defined by

$$K_{n,m} = \left\{\alpha = (j_1,\dots,j_m,k_1,\dots,k_m) \in \{1,\dots,n\}^{2m};\ j_l < k_l \text{ for all } l = 1,\dots,n\right\},$$

and $B^\alpha(\xi)$ is a shorthand for the linear combination $\sum_{l=1}^m \xi_l(B^{j_l}_{\kappa r_l} - B^{k_l}_{\kappa r_l})$. The important point here is that $E_B e^{iB^\alpha(\xi)}$ is positive for any $\alpha \in K_{n,m}$. We thus get the following inequality, valid for all $m \geq 1$

$$\mathbf{E}_B\left[\left(\sum_{1\leq j\neq k\leq n} V^{\varepsilon,j,k}_{t,x}\right)^m\right] \leq \sum_{\alpha\in K_{n,m}} \int_{[0,t]^m}\int_{\mathbb{R}^m} \mathbf{E}_B\left[e^{iB^\alpha(\xi)}\right]\prod_{l=1}^m |\xi_l|^{1-2H}\, d\xi\, dr$$
$$= \mathbf{E}_B\left[\left(\sum_{1\leq j\neq k\leq n} V^{j,k}_{t,x}\right)^m\right],$$

where $V^{j,k}_{t,x}$ is defined by (35). Plugging this inequality back into (40) and recalling expression (34) for $\mathbf{E}[u^n(t,x)]$, we easily deduce that $\mathbf{E}[u^n_\varepsilon(t,x)] \leq \mathbf{E}[u^n(t,x)]$, which finishes the proof. □

Acknowledgements We thank the referees for their useful comments which improved the presentation of the paper.

References

1. Alberts, T., Khanin, K., Quastel, J.: The continuum directed random polymer. J. Stat. Phys. **154**(1–2), 305–326 (2014).
2. Balan, R., Jolis, M., Quer-Sardanyons, L.: SPDEs with fractional noise in space with index $H < 1/2$. Electron. J. Probab. **20**(54), 36 (2015)
3. Bertini, L., Cancrini, N.: The stochastic heat equation: Feynman- Kac formula and intermittence. J. Stat. Phys. **78**(5–6), 1377–1401 (1995)
4. Bezerra, S., Tindel, S., Viens, F.: Superdiffusivity for a Brownian polymer in a continuous Gaussian environment. Ann. Probab. **36**(5), 1642–1675 (2008)
5. Chen, X.: Spatial asymptotics for the parabolic Anderson models with generalized time-space Gaussian noise. Ann. Probab. **44**(2), 1535–1598 (2016)
6. Hairer, M.: Solving the KPZ equation. Ann. Math. **178**(2), 559–664 (2013)
7. Hu, Y.: Analysis on Gaussian Space. World Scientific, Singapore (2017)
8. Hu, Y., Nualart, D.: Stochastic heat equation driven by fractional noise and local time. Probab. Theory Related Fields **143**(1–2), 285–328 (2009)
9. Hu, Y., Huang, J., Lê, K., Nualart, D., Tindel, S.: Stochastic heat equation with rough dependence in space. Ann. Probab. **45**, 4561–4616 (2017)
10. Hu, Y., Huang, J., Nualart, D., Tindel, S.: Stochastic heat equations with general multiplicative Gaussian noises: Hölder continuity and intermittency. Electron. J. Probab. **20**(55), 50 (2015)
11. Hu, Y., Nualart, D., Song, J.: Feynman-Kac formula for heat equation driven by fractional white noise. Ann. Probab. **30**, 291–326 (2011)
12. Huang, J., Lê, K., Nualart, D.: Large time asymptotics for the parabolic Anderson model driven by spatially correlated noise. Ann. Inst. H. Poincaré **53**, 1305–1340 (2017)
13. Khoshnevisan, D.: Analysis of Stochastic Partial Differential Equations. CBMS Regional Conference Series in Mathematics, vol. 119, pp. viii+116. Published for the Conference

Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence (2014)
14. Kilbas, A.A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractional Differential Equations. North-Holland Mathematics Studies, vol. 204. Elsevier Science B.V., Amsterdam (2006)
15. Li, W.V., Shao, Q.-M.: Gaussian processes: inequalities, small ball probabilities and applications. In: Stochastic Processes: Theory Methods, Handbook of Statistics, vol. 19. North-Holland, Amsterdam (2001)
16. Nualart, D.: The Malliavin Calculus and Related Topics. 2nd edn. Probability and its Applications (New York), pp. xiv+382. Springer, Berlin (2006)
17. Pipiras, V., Taqqu, M.: Integration questions related to fractional Brownian motion. Probab. Theory Related Fields **118**(2), 251–291 (2000)
18. Strichartz, R.S.: A guide to distribution theory and Fourier transforms. World Scientific Publishing Co., Inc., River Edge (2003)

Perturbation of Conservation Laws and Averaging on Manifolds

Xue-Mei Li

Abstract We prove a stochastic averaging theorem for stochastic differential equations in which the slow and the fast variables interact. The approximate Markov fast motion is a family of Markov process with generator $\mathscr{L}_x$ for which we obtain a quantitative locally uniform law of large numbers and obtain the continuous dependence of their invariant measures on the parameter x. These results are obtained under the assumption that $\mathscr{L}_x$ satisfies Hörmander's bracket conditions, or more generally $\mathscr{L}_x$ is a family of Fredholm operators with sub-elliptic estimates. For stochastic systems in which the slow and the fast variable are not separate, conservation laws are essential ingredients for separating the scales in singular perturbation problems we demonstrate this by a number of motivating examples, from mathematical physics and from geometry, where conservation laws taking values in non-linear spaces are used to deduce slow-fast systems of stochastic differential equations.

1 Introduction

A deterministic or random system with a conservation law is often used to approximate the motion of an object that is also subjected to many other smaller deterministic or random influences. The latter is a perturbation of the former. To describe the evolution of the dynamical system, we begin with these conservation laws. A conservation law is a quantity which does not change with time, for us it is an equi-variant map on a manifold, i.e. a map which is invariant under an action of a group. They describe the orbit of the action. Quantities describing the perturbed systems have their natural scales, the conservations laws can be used to determine the different components of the system which evolve at different speeds. Some components may move at a much faster speed than some others, in

X.-M. Li (✉)
Department of Mathematics, Imperial College London, London, UK
e-mail: xue-mei.li@imperial.ac.uk

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_18

which case we either ignore the slow components, in other words we approximate the perturbed system by the unperturbed one, or ignore the fast components and describe the slow components for which the key ingredient is ergodic averaging. It is a standard assumption that the fast variable moves so fast that its influence averaged over any time interval, of the size comparable to the natural scale of our observables, is effectively that of an averaged vector field. The averaging is with respect to a probability measure on the state space of the fast variable. Depending on the object of the study, we will need to neglect either the small perturbations or quantities too large (infinities) to fit into the natural scale of things. To study singularly perturbation operators, we must discard the infinities and at the same time retain the relevant information on the natural scale. In Hamiltonian formulation, for example, the time evolution of an object, e.g. the movements of celestial bodies, is governed by a Hamiltonian function. If the magnitude of the Hamiltonian is set to be of order '1', the magnitude of the perturbation (the collective negligible influences) is of order ϵ, then the perturbation is negligible on an interval of any fixed length. This ratio in magnitudes translates into time scales. If the original system is on scale 1, we work on a time interval of length $\frac{1}{\epsilon}$ to see the deviation of the perturbed trajectories. Viewed on the time interval $[0, 1]$ the perturbation is not observable. On $[0, \frac{1}{\epsilon}]$ the perturbation is observable, the natural object to study is the evolution of the energies while the dynamics of the Hamiltonian dynamics becomes too large. See [3, 18, 33].

If the state space of our dynamical system has an action by a group, the orbit manifold is a fundamental object. We use the projection to the orbit manifold as a conservation law and use it to separate the slow and the fast variables in the system. The slow variables lie naturally on a quotient manifold. In many examples we can further reduce this system of slow-fast stochastic differential equations (SDEs) to a product manifold $N \times G$, which we describe later by examples. From here we proceed to prove an averaging principle for the family of SDEs with a parameter ϵ. In these SDEs the slow and the fast variables are already separate, but they interact with each other.

This can then be applied to a local product space such as a principal bundle. In [64–66], the slow variables in the reduced system are random ODEs, where we study the system on the scale of $[1, \frac{1}{\epsilon^2}]$ to obtain results of the nature of diffusion creation. In these studies we bypassed stochastic averaging and went straight for the diffusion creation. In [38, 49, 62] stochastic averaging are studied, but they are computed in local coordinates. Here the slow variables solve a genuine SDE with a stochastic integral and the computations are global. We first prove an averaging theorem for these SDEs and then study some examples where we deduce a slow-fast system of SDEs from non-linear conservation laws, to which our main theorems apply.

Throughout the article $(\Omega, \mathscr{F}, \mathscr{F}_t, P)$ is a probability space satisfying the usual assumptions. Let (B_t, W_t) be a Brownian motion on $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ where $m_1, m_2 \in \mathscr{N}$. We write $B_t = (B_t^1, \ldots, B_t^{m_1})$ and $W_t = (W_t^1, \ldots, W_t^{m_1})$. Let N and G be two complete connected smooth Riemannian manifolds, let $x_0 \in N$ and $y_0 \in G$. Let ϵ denote a small positive number and let m_1 and m_2 be two natural numbers. Let $X : N \times G \times \mathbb{R}^{m_1} \to TN$ and $Y : N \times G \times \mathbb{R}^{m_2} \to TG$ be C^3 smooth maps

linear in the last variable. Let X_0 and Y_0 be C^2 smooth vector fields on N and on G respectively, with a parameter taking its values in the other manifold. Let us consider the SDEs,

$$\begin{cases} dx_t^\epsilon = X(x_t^\epsilon, y_t^\epsilon) \circ dB_t + X_0(x_t^\epsilon, y_t^\epsilon)\, dt, & x_0^\epsilon = x_0, \\ dy_t^\epsilon = \dfrac{1}{\sqrt{\epsilon}} Y(x_t^\epsilon, y_t^\epsilon) \circ dW_t + \dfrac{1}{\epsilon} Y_0(x_t^\epsilon, y_t^\epsilon)\, dt, & y_0^\epsilon = y_0. \end{cases} \tag{1}$$

The symbol $\circ$ is used to denote Stratonovich integrals. By choosing an orthonormal basis $\{e_i\}$ of $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$, we obtain a family of vector fields $\{X_1, \ldots, X_{m_1}, Y_1, \ldots, Y_{m_2}\}$ as following: $X_i(x) = X(x)(e_i)$ for $1 \le i \le m_1$ and $Y_i(x) = Y(x)(e_i)$ for $i = m_1 + 1, \ldots, m_1 + m_2$. Then the system of SDEs (1) is equivalent to the following

$$\begin{cases} dx_t^\epsilon = \displaystyle\sum_{k=1}^{m_1} X_k(x_t^\epsilon, y_t^\epsilon) \circ dB_t^k + X_0(x_t^\epsilon, y_t^\epsilon)\, dt, & x_0^\epsilon = x_0, \\ dy_t^\epsilon = \dfrac{1}{\sqrt{\epsilon}} \displaystyle\sum_{k=1}^{m_2} Y_k(x_t^\epsilon, y_t^\epsilon) \circ dW_t^k + \dfrac{1}{\epsilon} Y_0(x_t^\epsilon, y_t^\epsilon)\, dt, & y_0^\epsilon = y_0. \end{cases}$$

If V is a vector field, by Vf we mean $df(V)$ or $L_V f$, the Lie differential of f in the direction of V. Then $(x_t^\epsilon, y_t^\epsilon)$ is a sample continuous Markov process with generator $\mathscr{L}^\epsilon := \frac{1}{\epsilon}\mathscr{L} + \mathscr{L}^{(1)}$ where

$$\mathscr{L} = \frac{1}{2} \sum_{k=1}^{m_2} Y_i^2 + Y_0, \quad \mathscr{L}^{(1)} = \frac{1}{2} \sum_{k=1}^{m_1} X_i^2 + X_0.$$

In other words if $f : N \times G \to \mathbb{R}$ is a smooth function then

$$\mathscr{L}^\epsilon f(x, y) := \frac{1}{\epsilon} \mathscr{L}_x (f(\cdot, y))(x) + \mathscr{L}_y^{(1)} (f(x, \cdot))(y),$$

where

$$\mathscr{L}_x f(x, \cdot) = \left(\frac{1}{2} \sum_{k=1}^{m_2} Y_i^2(x, \cdot) + Y_0(x, \cdot) \right) f(x, \cdot),$$

$$\mathscr{L} y^{(1)} f(\cdot, y) = \left(\frac{1}{2} \sum_{k=1}^{m_1} X_i^2(\cdot, y) + X_0(\cdot, y) \right) f(\cdot, y).$$

The result we seek is the weak convergence of the slow variables x_t^ϵ to a Markov process $\bar{x}_t$ whose Markov generator $\bar{\mathscr{L}}$ is to be described.

Let T be a positive number and let $C([0,T];N)$ denote the family continuous functions from $[0,T]$ to N, the topology on $C([0,T];N)$ is given by the uniform distance. A family of continuous stochastic processes x_t^ϵ on N is said to converge to a continuous process $\bar{x}_t$ if for every bounded continuous function $F : C([0,T];N) \to \mathbb{R}$, as ϵ converges to zero,

$$\mathbf{E}[F(x_\cdot^\epsilon)] \to \mathbf{E}[F(\bar{x}_\cdot)].$$

In particular, if $u^\epsilon(t,x,y)$ is a bounded regular solution to the Cauchy problem for the PDE (for example C^3 in space and C^1 in time) $\frac{\partial u^\epsilon}{\partial t} = \mathscr{L}^\epsilon u$ with the initial value f in L_∞, then $u^\epsilon(t,x_0,y_0) = \mathbf{E}[f(x_t^\epsilon, y_t^\epsilon)]$. Suppose that the initial value function f is independent of the second variable so $f : N \to \mathbb{R}$. Then the weak convergence will imply that

$$\lim_{\epsilon \to 0} u^\epsilon(t,x_0,y_0) = u(t,x_0)$$

where $u(t,x)$ is the bounded regular solution to the Cauchy problem

$$\frac{\partial u}{\partial t} = \bar{\mathscr{L}} u, \quad u(0,x) = f(x).$$

Stochastic averaging is a procedure of equating time averages with space averages using a form of Birkhoff's ergodic theorem or a law of large numbers. Birkhoff's pointwise ergodic theorem states that if $T : E \to E$ is a measurable transformation preserving a probability measure μ on the metric space E then for any $F \in L^1(\mu)$,

$$\frac{1}{n}\sum_{k=1}^{n} F(T^k x) \to \mathbf{E}(F|\mathscr{I})$$

for almost surely all x, as $n \to \infty$, and where $\mathscr{I}$ is the invariant σ-algebra of T. Suppose that (z_t) is a sample continuous ergodic stochastic process with values in E, stationary on the space of paths $C([0,1];E)$. Denote by μ its one time probability distribution. Then for any real valued function $f \in L^1(\mu)$,

$$\frac{1}{t}\int_0^t f(z_s)ds \to \int f(z)\mu(dz).$$

This is simply Birkhoff's theorem applied to the shift operator and to the function $F(\omega) = \int_0^1 f(z_s(\omega))ds$. If z_t is not stationary, but a Markov process with initial value a point, conditions are needed to ensure the convergence of the Markov process to equilibrium with sufficient speed.

We explain below stochastic averaging for a random field whose randomness is introduced by a fast diffusion. Let $(x_t^\epsilon, y_t^\epsilon)$ be solution to the SDE on $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$:

$$dx_t^\epsilon = \sum_{k=1}^{m_1} \sigma_k(x_t^\epsilon, y_t^\epsilon)\, dB_t^k + b(x_t^\epsilon, y_t^\epsilon)\, dt, \quad dy_t^\epsilon = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} \theta_k(x_t^\epsilon, y_t^\epsilon) dW_t^k + \frac{1}{\epsilon} b(x_t^\epsilon, y_t^\epsilon)\, dt.$$

with initial values $x_0^\epsilon = x_0$, and $y_0^\epsilon = y_0$. Here the stochastic integrations are Itô integrals. A sample averaging theorem is as following. Let z_t^x denote the solution to the SDE

$$dz_t^x = \sum_{k=1}^{m_2} \theta_k(x, z_t^x) dW_t^k + b(x, z_t^x)\, dt$$

with initial value z_0^x. Suppose that the coefficients are globally Lipschitz continuous and bounded. Suppose that $\sup_{t\in[0,T]} \sup_{\epsilon\in(0,1]} \mathbf{E}|y_t^\epsilon|^2$ and $\sup_x \sup_{t\in[0,T]} \mathbf{E}|z_t^x|^2$ are finite. Also suppose that there exist functions $\bar{a}_{i,j}$ and $\bar{b}$ such that

$$\begin{aligned} \left| \frac{1}{t} \mathbf{E} \int_0^t b(x, z_s^x) ds - \bar{b}(x) \right| &\le C(t)(|x|^2 + |z|^2 + 1), \\ \left| \frac{1}{t} \mathbf{E} \int_0^t \sum_k \sigma_k^i \sigma_k^j (x, z_s^x) ds - \bar{a}_{i,j}(x) \right| &\le C(t)(|x|^2 + |z|^2 + 1). \end{aligned} \tag{2}$$

Then the stochastic processes x_t^ϵ converge weakly to a Markov process with generator $\frac{1}{2}\bar{a}_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} + \bar{b}_k \frac{\partial^2}{\partial x_k}$, see [17, 46, 80, 81, 89]. See also [47, 53, 60, 73, 75]. See [8, 18, 20, 25, 35–37, 40, 41, 43, 51, 67, 77, 91] for a range of more recent related work. We also refer to the following books [12, 55, 58, 76, 79]

Averaging of stochastic differential equations on manifolds has been studied in the following articles [52, 62, 63], and [38]. In these studies either one restricts to local coordinates, or has a set of convenient coordinates, or one works directly with local coordinates. We will be using a global approach.

We will first deduce a locally uniform Birkhoff's ergodic theorem for $\mathscr{L}_x$, then prove an averaging theorem for (1). Finally we study a number of examples of singular perturbation problems.

The main assumptions on $\mathscr{L}_x$ is a Hörmander's (bracket) condition.

Definition 1 Let $X_0, X_1, \ldots, X_k$ be smooth vector fields.

1. The differential operator $\sum_{k=1}^m (X_i)^2 + X_0$ is said to satisfy *Hörmander's condition* if $\{X_k, k = 0, 1, \ldots, m\}$ and their iterated Lie brackets generate the tangent space at each point.

2. The differential operator $\sum_{k=1}^{m}(X_i)^2 + X_0$, is said to satisfy *strong Hörmander's condition* if $\{X_k, k = 1, \ldots, m\}$ and their iterated Lie brackets generate the tangent space at each point.

Outline of the paper In Sect. 3 we study the regularity of invariant probability measures μ^x of $\mathscr{L}_x$ with respect to the parameter x and prove the local uniform law of large numbers with rate. We may assume that each $\mathscr{L}_x$ satisfies Hörmander's condition. What we really need is that $\mathscr{L}_X$ is a family of Fredholm operators satisfying the sub-elliptic estimates and with zero Fredholm index. In Sect. 4 we give estimates for SDEs on manifolds. It is worth noticing that we do not assume that the transition probabilities have densities. We use an approximating family of distance functions to overcome the problem that the distance function is not smooth. These estimates lead easily to the tightness of the slow variables. In Sect. 5 we prove the convergence of the slow variables, for which we first prove a theorem on time averaging of path integrals of the slow variables. This is proved under a law of large numbers with any uniform rate. In Sect. 2 we study some examples of singular perturbation problems. Finally, we pose a number of open questions, one of which is presented in the next section, the others are presented in Sect. 2.

1.1 Description of Results

The following law of large numbers with a locally uniform rate is proved is Sect. 3.

Theorem 1 (Quantitative Locally Uniform Law of Large Numbers) *Let G be a compact manifold. Suppose that Y_i are bounded, C^∞ with bounded derivatives. Suppose that each*

$$\mathscr{L}_x = \frac{1}{2}\sum_{i=1}^{m} Y_i^2(x, \cdot) + Y_0(x, \cdot)$$

satisfies Hörmander's condition (Definition 1), and has a unique invariant probability measure μ_x. Then the following statements hold for μ_x.

(a) *$x \mapsto \mu_x$ is locally Lipschitz continuous in the total variation norm.*
(b) *For every $s > 1 + \frac{\dim(G)}{2}$ there exists a positive constant $C(x)$, depending continuously in x, such that for every smooth function $f : G \to \mathbb{R}$,*

$$\left| \frac{1}{T}\int_t^{t+T} f(z_r^x)\, dr - \int_G f(y)\mu_x(dy) \right|_{L_2(\Omega)} \leq C(x)\|f\|_s \frac{1}{\sqrt{T}}, \tag{3}$$

where z_r denotes an $\mathscr{L}_x$-diffusion.

Remark 1 Let $P^x(t, y, \cdot)$ denote the transition probability of $\mathscr{L}_x$. Suppose $\mathscr{L}_x$ satisfies Doeblin's condition. Then $\mathscr{L}_x$ has a unique invariant probability measure.

This holds in particular if $\mathscr{L}_x$ satisfies the strong Hörmander's condition and G is compact. The uniqueness follows from the fact that it has a smooth strictly positive density. (Hörmander's condition ensures that any invariant measure has a smooth kernel and the kernel of its L_2 adjoint $\mathscr{L}^*$ contains a non- negative function. The density is however not necessarily positive.) Suppose that each $\mathscr{L}_x$ satisfies the strong Hörmander's condition (cf. Definition 1) and G is compact. It is well know that the transition probability measures $P^x(t, y_0, \cdot)$, with any initial value y_0, converges to the unique invariant probability measure μ^x with an exponential rate which we denote by $C(x)e^{\gamma(x)t}$. If x takes values also in a compact space N, the exponential rate and the constant in front of the exponential rate can be taken to be independent of x. When N is non-compact, we obviously need to make further assumptions on $\mathscr{L}_x$ for a uniform estimate. There have been work on ergodicity of this type. We refer to : [10, 24, 69, 90]

Set

$$\tilde{X}_0(\cdot, y) = \frac{1}{2}\sum_{i=1}^{m_1} \nabla X_i(X_i)(y, \cdot) + X_0(y, \cdot),$$

$$\tilde{Y}_0(x, \cdot) = \frac{1}{2}\sum_{i=1}^{m_2} \nabla Y_i(Y_i)(x, \cdot) + Y_0(x, \cdot).$$

Let O be a reference point in N and ρ is the Riemannian distance from O.

Assumption 1 (Assumptions on X_i and N) *Suppose that $\tilde{X}_0$ and X_i are C^1, where $i = 1, \ldots, m$. Suppose that **one** of the following two statements holds.*

(i) *The sectional curvature of N is bounded. There exists a constant K such that*

$$\sum_{i=1}^{m} |X_i(x, y)|^2 \leq K(1 + \rho(x)), \quad \left|X_0(x, y)\right| \leq K(1 + \rho(x)), \quad \forall x \in N, \forall y \in G.$$

(ii) *Suppose that the square of the distance function on N is smooth. Suppose that*

$$\frac{1}{2}\sum_{i=1}^{m} \nabla d\rho^2(X_i(\cdot, y), X_i(\cdot, y)) + d\rho^2(\tilde{X}_0(\cdot, y)) \leq K + K\rho^2(\cdot), \quad \forall y \in G.$$

Assumption 2 (Assumptions on Y_i and G) *We suppose that G has bounded sectional curvature. Suppose that $\tilde{Y}_0$ and Y_j are C^2 and bounded with bounded first order derivatives.*

The following is extracted from Theorem 5.6.

Theorem 2 (Averaging Theorem) *Suppose that there exists a family of invariant probability measure μ_x on G that satisfies the conclusions of Theorem 1. Suppose the assumptions on X_i, N, Y_i and G hold (Assumptions 1 and 2). Then as $\epsilon \to 0$,*

the stochastic processes x_t^ϵ converges weakly on $C([0,T],N)$ to a Markov process with generator $\bar{\mathscr{L}}$.

Remark 2

(i) If f is a smooth function on N with compact support then

$$\bar{\mathscr{L}}f(x) = \int_G \left(\frac{1}{2} \sum_{i=1}^{m_1} X_i^2(\cdot, y) f + X_0(\cdot, y) f \right)(x)\, \mu_x(dy). \tag{4}$$

See the Appendix in Sect. 5 for a sum of squares of vector fields decomposition of $\bar{\mathscr{L}}$.

(ii) Under Assumptions 1 there exists a unique global solution x_t^ϵ for each initial value (x, y). We also have uniform estimates on the distance $\rho(x_s^\epsilon, x_t^\epsilon)$ which leads to the conclusion that the family $\{x_\cdot^\epsilon, \epsilon > 0\}$ is tight. Also we may conclude that the moments of the solutions are bounded uniformly on any compact time interval and in ϵ for $\epsilon \in (0, 1]$. Such estimates are given in Sect. 4.

(iii) Under Assumptions 1 and 2 we may approximate the fast motion, on sub-intervals $[t_i, t_{i+1}]$, by freezing the slow variables and obtain a family of Markov processes with generator $\mathscr{L}_x$. The size of the sub-intervals must be of size $o(\epsilon)$ for the error of the approximation to converge to zero as $\epsilon \to 0$, and large on the scale of $\frac{1}{\epsilon}$ for the ergodic average to take effect.

Problem 1 Suppose that $\mathscr{L}_x$ satisfies Hörmander's condition. Then the kernel of $\mathscr{L}_x^*$ is finite dimensional. Without assuming the uniqueness of the invariant probability measures, it is possible to define a projection to the kernel of $\mathscr{L}_x$, by pairing up a basis $\{u_i(x)\}$ of $\ker(\mathscr{L}_x)$ with a dual basis $\pi^i(x)$ of $\ker(\mathscr{L}_x^*)$ and this leads to a family of projection operators $\Pi(x)$. To obtain a locally uniform version of this, we should consider the continuity of Π with respect to x. Let us consider the simple case of a family of Fredhom operators $T(x)$ from a Hilbert space E to a Hilbert space F. It is well known that the dimension of their kernels may not be a continuous function of x, but the Fredholm index of $T(x)$ is a continuous function if x in the space of bounded linear operators [5]. See also [87, 88] for non-elliptic operators. Given that the projection $\pi(x)$ involves both the kernel and the co-kernel, it is reasonable to expect that $\Pi(x)$ is continuous in x. The question is whether this is true and more importantly whether in this situation there is a local uniform Law of large numbers.

2 Examples

We describe some motivating examples, the first being dynamical descriptions for Brownian motions, the second being the convergence of metric spaces. The overarching question concerning the second is: given a family of metric spaces

converging to another in measured Gromov Hausdorff topology, can we give a good dynamical description for their convergence? What can one say about the associated singular operators? These will considered in terms of stochastic dynamics. See [50, 72] and [63] concerning collapsing of Riemannian manifolds. The third example is a model on the principal bundle. These singular perturbation models were introduced in [62–65], where the perturbations were chosen carefully for diffusion creation. The reduced systems are random ODEs for which a set of limit theorems are available, and the perturbations are chosen so that one could bypass the stochastic averaging procedure and work directly on the faster scale for diffusion creation $[0, \frac{1}{\epsilon}]$. Theorem 5.6 allows us to revisit these models to include more general perturbations, in which the effective limits on $[0, 1]$ are not trivial. It also highlights from a different angle the choice of the perturbation vector in the models which we explain below.

2.1 A Dynamical Description for Brownian Motions

In 1905, Einstein, while working on his atomic theory, proposed the diffusion model for describing the density of the probability for finding a particle at time t in a position x. A similar model was proposed by Smoluchowski with a force field. Some years later Langevin [61] and Ornstein-Uhlenbeck [85] proposed a dynamical model for Brownian motion for time larger than the relaxation time $\frac{1}{\beta}$:

$$\begin{cases} \dot{x}(t) = v(t) \\ \dot{v}(t) = -\beta v(t)dt + \sqrt{D}\beta dB_t + \beta b(x(t))dt \end{cases}$$

where B_t a one dimensional Brownian motion and b a vector field. This equation is stated for $\mathbb{R}$ with β, D constants and was studied by Kramers [57] and Nelson [71]. The model is on the real line, there exists only one direction for the velocity field. The magnitude of $v(t)$ together with the sign changes rapidly.

The second order differential equations for unit speed geodesics, on a manifold M, are equivalent to first order ODEs on the space of orthonormal frames of M, this space will be denoted by OM. Suppose that we rotate the direction of the geodesic uniformly, according to the probability distribution of a Brownian motion on $SO(n)$, while keeping its magnitude fixed to be 1, and suppose that the rotation is at the scale of $\frac{1}{\epsilon}$ then the projections to M of the solutions of the equations on OM converge to a fixed point as $\epsilon \to 0$. But if we further tune up the speed of the rotation, these motions converge to a scaled Brownian motion, whose scale is given by an eigenvalue of the fast motion on $SO(n)$. See [64]. An extension to manifolds was first studied [26] followed by [16]. That in [26] is different from that in Li-geodesic, which is also followed up in [1] where the authors iremoved the geometric curvature restrictions in [64]. See also [14] for a local coordinate approach and more recently

[27]. Assume the dimension of M is greater than 1. The equations, [64], describing this are as following:

$$\begin{cases} du_t^\epsilon = H_{u_t^\epsilon}(e_0)\,dt + \dfrac{1}{\sqrt{\epsilon}}\displaystyle\sum_{k=1}^{N} A_k^*(u_t^\epsilon)\circ dW_t^k + A_0^*(u_t^\epsilon)\,dt, \\ u_0^\epsilon = u_0. \end{cases} \tag{5}$$

where $\{A_1, \ldots, A_N\}$ is an o.n.b. of $\mathfrak{so}(n)$, and $A_0 \in \mathfrak{so}(n)$. The star sign denotes the corresponding vertical fundamental vector fields and $H(u)(e_0)$ is the horizontal vector field corresponding to a unit vector e_0 in $\mathbb{R}^n$. This following theorem is taken from [64].

Theorem 2A *The position part of $u^\epsilon_{\frac{t}{\epsilon}}$, which we denote by $(x^\epsilon_{\frac{t}{\epsilon}})$, converges to a Brownian motion on M with generator $\frac{4}{n(n-1)}\Delta$. Furthermore the parallel translations along these smooth paths of $(x^\epsilon_{\frac{t}{\epsilon}})$ converge to stochastic parallel translations along the Hörder continuous sample paths of the effective scaled Brownian motion.*

The conservation law in this case is the projection π, taking a frame to its base point, using which we obtain the following reduced system of slow-fast SDEs:

$$\begin{cases} \dfrac{d}{dt}\tilde{x}_t^\epsilon = H_{\tilde{x}_t^\epsilon}(g_{\frac{t}{\epsilon}}e_0), \quad \tilde{x}_0^\epsilon = u_0, \\ dg_t = \displaystyle\sum_{k=1}^{m} g_t A_k \circ dw_t^k + g_t A_0\,dt, \quad g_0 = Id. \end{cases}$$

The slow variable does not have a stochastic part, the averaging equation is given by the average vector field $\int_{SO(n)} H(ug)(e)dg$, where dg is the Haar measure, and vanishes. Hence we may observe the slow variable on a faster scale and consider $x^\epsilon_{\frac{t}{\epsilon}}$.

In Sect. 6.1 we use the general results obtained later to study two generalised models.

2.2 Collapsing of Manifolds

Our overarching question is how the stochastic dynamics describe the convergence of metric spaces. Let us consider a simple example: $SU(2)$ which can be identified with the sphere S^3. The Lie algebra of $SU(2)$ is given by the Pauli matrices

$$X_1 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

By declaring $\{\frac{1}{\sqrt{\epsilon}}X_1, X_2, X_3\}$ an orthonormal frame we define Berger's metrics g^ϵ. Thus (S^3, g^ϵ) converges to S^2. They are the first known family of manifolds which collapse to a lower dimensional one, while keeping the sectional curvatures uniformly bounded (J. Cheeger). Then all the operators in the sum

$$\Delta^\epsilon_{S^3} = \frac{1}{\epsilon}(X_1)^2 + (X_2)^2 + (X_3)^2 = \frac{1}{\epsilon}\Delta_{S^1} + \Delta_H$$

commute, the eigenvalues satisfy the relation $\lambda_3(\Delta^\epsilon_{S^3}) = \frac{1}{\epsilon}\lambda_1(\Delta_{S^1}) + \lambda_2(\Delta_H)$. The non-zero eigenvalues of Δ_{S^1} flies away and the eigenfunctions of $\lambda_1 = 0$ are function on the sphere $S^2(\frac{1}{2})$ of radius $\frac{1}{2}$, the convergence of the spectrum of $\Delta^\epsilon_{S^3}$ follows. See [11, 84, 86] for discussions on the spectrum of Laplacians on spheres, on homogeneous Riemannian manifolds and on Riemannian submersions with totally geodesic fibres.

We study

$$\mathscr{L}^\epsilon := \frac{1}{\epsilon}\Delta_{S^1} + Y_0$$

in which Δ_{S^1} and Y_0 do not commute. Take for example, $Y_0 = aX_2 + bX_3$ where $|Y_0| = 1$. Let $\pi(z, w) = \frac{1}{2}(|w|^2 - |z|^2, z\bar{w})$ be the Hopf map. Let u^ϵ_t be an $\mathscr{L}^\epsilon$-diffusion with the initial value u_0. Then $\pi(u^\epsilon_{\frac{t}{\epsilon}})$ converges to a BM on $S^2(\frac{1}{2})$, scaled by $\frac{1}{2}$. See [65]. See also [72] for related studies. It is perhaps interesting to observe that $\mathscr{L}^\epsilon$ satisfies Hörmander's condition for any $Y_0 \neq 0$. Later we see that this fact is not an essential feature of the problem. The model on S^3 in [63] is a variation of this one.

2.3 *Inhomogeneous Scaling of Riemannian Metrics*

If a manifold is given a family of Riemannian metrics depending on a small parameter $\epsilon > 0$, the Laplacian operators Δ^ϵ is a family of singularly perturbed operators. We might ask the question whether their spectra converge. More generally let us consider a family of second order differential operators $\mathscr{L}^\epsilon = \frac{1}{\epsilon}\mathscr{L}_0 + \mathscr{L}_1$, each in the form of a finite sum of squares of smooth vector fields with possibly a first order term. As $\epsilon \to 0$, the corresponding Markov process does not converge in general. In the spirit of Noether's theorem, to see a convergent slow component we expect to see some symmetries for the system $\mathscr{L}_0$. On the other hand, by a theorem of S. B. Myers and N. E. Steenrod [70], the set of all isometries of a Riemannian manifold M is a Lie group under composition of maps, and furthermore the isotropy subgroup $\mathrm{Iso}_o(M)$ is compact. See also S. Kobayashi and K. Nomizu [54]. We are led to study homogeneous manifolds G/H, where G is a smooth Lie group and H is a compact sub-group of G. We denote by $\mathfrak{g}$ and $\mathfrak{h}$ their respective Lie algebras.

Let $\mathfrak{g}$ be endowed an $\mathrm{Ad}(H)$-invariant inner product and take $\mathfrak{m} = \mathfrak{h}^{\perp}$. Then G/H is a reductive homogeneous manifold, in the sense of Nomizu, by which we mean $\mathrm{Ad}(H)(\mathfrak{m}) \subset \mathfrak{m}$. This is a different from the concept of a reductive Lie group, where the adjoint representation of the Lie group G is completely reducible. (Bismut studied a natural deformation of the standard Laplacian on a compact Lie group G into a hypoelliptic operator on TG see [15].) We assume that the real Lie group G is smooth, connected, not necessarily compact, of dimension n and H a closed connected proper subgroup of dimension at least one. We identify elements of the Lie algebra with left invariant vector fields.

We generate a family of Riemannian metrics on G by scaling the $\mathfrak{h}$ directions by ϵ. Let $\{A_1, \ldots, A_p, Y_{p+1}, \ldots, X_N\}$ be an orthonormal basis of $\mathfrak{g}$ for an inner product extending an orthonormal basis $\{A_1, \ldots, A_p\}$ of $\mathfrak{h}$ with the remaining vectors from $\mathfrak{m}$. By declaring

$$\left\{ \frac{1}{\sqrt{\epsilon}} A_1, \ldots, \frac{1}{\sqrt{\epsilon}} A_p, Y_{p+1}, \ldots Y_N \right\}$$

an orthonormal frame, we obtain a family of left invariant Riemannian metrics. Let us consider the following second order differential operator, related to the re-scaled metric:

$$\mathscr{L}^{\epsilon} = \frac{1}{2\epsilon} \sum_{k=1}^{m_2} (A_k)^2 + \frac{1}{\epsilon} A_0 + Y_0,$$

where $A_k \subset \mathfrak{h}$ and $Y_0 \in \mathfrak{m}$ is a unit vector. This leads to the following family of equations, where $\epsilon \in (0, 1]$,

$$dg_t^{\epsilon} = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} A_k(g_t^{\epsilon}) \circ dB_t^k + \frac{1}{\epsilon} A_0(g_0^{\epsilon}) dt + Y_0(g_t^{\epsilon}) dt, \quad g_0^{\epsilon} = g_0.$$

These SDEs belong to the following family of equations

$$dg_t = \sum_{k=1}^{m_2} \gamma A_k(g_t) \circ dB_t^k + \gamma A_0(g_0)\, dt + \delta Y_0(g_t) dt.$$

The solutions of the latter family of equations, with parameters γ and δ real numbers, interpolate between translates of a one parameter subgroups of G and diffusions on H. Our study of $\mathscr{L}^{\epsilon}$ is related to the concept o 'taking the adiabatic limit' [13, 68].

Let (g_t^{ϵ}) be a Markov processes with Markov generator $\mathscr{L}^{\epsilon}$, and set $x_t^{\epsilon} = \pi(g_t^{\epsilon})$ where π is the map taking an element of G to the coset gH. Then $\mathscr{L}^{\epsilon} = \frac{1}{\epsilon} \mathscr{L}_0 +$

Y_0 where $\mathscr{L}_0 = \frac{1}{2}\sum_{k=1}^{m_2}(A_k)^2 + A_0$. We will assume that $\{A_k\} \subset \mathfrak{h}$ are bracket generating. Scaled by $1/\epsilon$, the Markov generator of $(g^\epsilon_{\frac{t}{\epsilon}})$ is precisely $\frac{1}{\epsilon}\mathscr{L}^\epsilon$.

The operators $\mathscr{L}^\epsilon$ are not necessarily hypo-elliptic in G, and they will not expected to converge in the standard sense. Our first task is to understand the nature of the perturbation and to extract from them a family of first order random differential operators, $\tilde{\mathscr{L}}^\epsilon$, which converge and which have the same orbits as $\mathscr{L}^\epsilon$, the 'slow motions'. The reduced operators, $\frac{1}{\epsilon}\tilde{\mathscr{L}}^\epsilon$, describe the motion of the orbits under 'perturbation'.

Their effective limit is either a one parameter sub-groups of G in which case our study terminate, or a fixed point in which case we study the fluctuation dynamics on the time scale $[0, \frac{1}{\epsilon}]$. On the Riemmanian homogeneous manifold, if G is compact, the effect limit on G is a geodesic at level one and a fixed point at level two. On the scale of $[0, \frac{1}{\epsilon}]$ we would consider $\frac{1}{\epsilon}\mathscr{L}_0$ as perturbation. It is counter intuitive to consider the dominate part as the perturbation. But the perturbation, although very large in magnitude, is fast oscillating. The large oscillating motion get averaged out, leaving an effective motion corresponding to a second order differential operator on G.

This problems breaks into three parts: separate the slow and the fast variable, which depends on the principal bundle structure of the homogeneous space, and determine the natural scales; the convergence of the solutions of the reduced equations which is a family of random ODEs; finally the buck of the interesting study is to determine the effective limit, answering the question whether it solves an autonomous equation.

It is fairly easy to see that x^ϵ_t moves relatively slowly. The speed at which x^ϵ_t crosses M is expected to depend on the specific vector Y_0, however in the case of $\{A_1, \ldots, A_p\}$ is an o.n.b. of $\mathfrak{h}$ and $A_0 = 0$, they depend only on the $\mathrm{Ad}(H)$-invariant component of Y_0.

The separation of slow and fast variables are achieved by first projecting the motion down to G/H and then horizontally lift the paths back (a non-Markovian procedure), exposing the action in the fibre directions. The horizontal process thus obtained is the 'slow part' of g^ϵ_t and will be denoted by u^ϵ_t. It is easy to see that the reduced dynamic is given by

$$\dot{u}^\epsilon_t = \mathrm{Ad}(h_{\frac{t}{\epsilon}})(Y_0)(u^\epsilon_t).$$

where h_t has generator $\frac{1}{2}\sum(A_i)^2 + A_0$.

If $\{A_0, A_1, \ldots, A_m\}$ generates the vector space $\mathfrak{h}$, the differential operator $\frac{1}{2}\sum(A_i)^2 + A_0$ satisfies Hörmander's condition in which case the invariant probability measure is the normalised Haar measure. Then u^ϵ_t converges to the solution of the ODE:

$$\frac{d}{dt}\bar{u}_t = \int_H \mathrm{Ad}(h)(Y_0)dh.$$

Let us take an $\mathrm{Ad}(H)$ invariant decomposition of $\mathfrak{m}$, $\mathfrak{m} = \mathfrak{m}_0 + \mathfrak{m}_1$ where $\mathfrak{m}_0$ is the vector space of invariant vectors and $\mathfrak{m}'$ is its orthogonal complement. Then

$$\int_H \mathrm{Ad}(h)(Y_0)dh = {Y_0}^{\mathfrak{m}_0}$$

where the superscript $\mathfrak{m}_0$ denote the $\mathfrak{m}_0$ component of Y. This means that the dynamics is a fixed point if and only if $Y_0^{\mathfrak{m}_0} = 0$.

In [65] we take $Y_0^{\mathfrak{m}_0} = 0$ and answered this question by a multi-scale analysis and studied directly the question concerning $Y_0 \in \tilde{\mathfrak{m}}$, without having to go through stochastic averaging. Theorem 1 makes this procedure easier to understand. Then we consider the dynamics on $[0, \frac{1}{\epsilon}]$. The reduced first order random differential operators give rise to second order differential operators by the action of the Lie bracket.

2.4 Perturbed Dynamical Systems on Principal Bundles

In the examples described earlier, we have a perturbed dynamical system on a manifold P. On P there is an action by a Lie group G, and the projection to $M = P/G$ is a conservation law. We then study the convergence of the slow motion, the projection to M, and their horizontal lifts. More precisely we have a principal bundle with fibre the Lie group G. To describe these motions we consider the kernels of the differential of the projection π: they are called the vertical tangent spaces and will be denoted by VT_uP. Any vector field taking values in the vertical tangent space is called a vertical vector field, the Lie-bracket of any two vertical vector fields is vertical. A smooth choice of the complements of the vertical spaces, that are right invariant, determines a connection. These complements are called the horizontal spaces. The ensemble is denoted by HT_uP and called the horizontal bundle. From now on we assume that we have chosen such a horizontal space. A vector field taking values in the horizon tangent spaces is said to be a horizontal vector field. Right invariant horizontal vector fields are specially interesting, they are precisely the horizontal lifts of vector fields on M.

Let $\pi : P \to M$ denote the canonical projection taking an element of the total space P to the corresponding element of the base manifold. Also let $R_g : P \to P$ denote the right action by g, for simplicity we also write ug, where $u \in P$, for $R_g u$. A connection on a principal bundle P is a splitting of the tangent bundle $T_uP = HT_uP + VT_uP$ where VT_uP is the kernel of the differential of $\mathrm{t}\pi$. Let $\mathfrak{g}$ denote the Lie algebra of G. For any $A \in \mathfrak{g}$ we define

$$A^*(u) = \lim_{t \to 0} R_{\exp(tA)}u.$$

The splitting mentioned earlier is in one to one correspondence with a connection 1-form, by which we mean a map $\varpi : T_u P \to \mathfrak{g}$ with the following properties:

$$(R_g)^* \varpi = \mathrm{ad}(g^{-1})\varpi, \quad \varpi(A^*) \equiv A.$$

This splitting also determines a horizontal lifting map $\mathfrak{h}_u$ at $u \in P$ and a family of horizontal vector fields H_i. If $\{e_1, \ldots e_n\}$ is an orthonormal basis of $\mathbb{R}^n$, where $n = \dim(M)$, we set $H_i(u) = \mathfrak{h}_u(ue_i)$. If $\{A_1, \ldots, A_N\}$ is an orthonormal basis of the Lie algebra $\mathfrak{g}$, then at every point u, $\{H_1(u), \ldots, H_n(u), A_1^*(u), \ldots, A_N^*(u)\}$ is a basis of $T_u P$. We give P the Riemannian metric so that the basis is orthonormal.

Any stochastic differential equation (SDE) on P are of the following form, where β and γ are two real positive numbers and σ_j^k and θ_j^k are BC^3 functions on P.

$$\begin{aligned} du_t =& \beta \sum_{k=1}^{m_1} \left(\sum_{i=1}^{n} \sigma_k^i(u_t) H_i(u_t) \right) \circ dB_t^k + \beta^2 \sum_{i=1}^{n} \sigma_0^i(u_t) H_i(u_t) dt \\ &+ \gamma^2 \sum_{k=1}^{m_2} \left(\sum_{j=1}^{N} \theta_k^j(u_t) A_j(u_t) \right) \circ dW_t^k + \gamma \sum_{j=1}^{N} \theta_0^j(u_t) A_j(u_t) dt. \end{aligned}$$

Set $X_k = \sum_{i=1}^n \sigma_k^i \, H_i$, and $Y_k = \sum_{j=1}^N \theta_k^j \, A_j$. Then the equation is of the form

$$du_t = \beta \sum_{k=1}^{m_1} X_k(u_t) \circ dB_t^k + \beta^2 \, X_0(u_t) \, dt + \gamma \sum_{k=1}^{m_2} Y_k \circ dW_t^k + \gamma^2 \, Y_0(u_t) dt.$$

The solutions are Markov processes with Markov generator

$$\beta^2 \left(\sum_{k=1}^{m_1} (X_k)^2 + X_0 \right) + \gamma^2 \left(\sum_{k=1}^{m_2} (Y_k)^2 + Y_0 \right).$$

We observe that the projection of the second factor vanishes, so if $\beta = 0$, then $\pi(u_t) = \pi(u_0)$ and π is a conservation law. The equation with small β is a stochastic dynamic whose orbits deviate slightly from that of the initial value u_0. If on the other hand, X_i are vector fields invariant under the action of the group, and $\gamma = 0$ then the projection $\pi(u_t)$ is an autonomous SDE on the manifold M.

Let us take $\beta = 1$ and $\gamma = \frac{1}{\sqrt{\epsilon}}$.

$$\begin{cases} du_t^\epsilon = \sum_{k=1}^{m_1} \left(\sum_{i=1}^{n} \sigma_k^i(u_t^\epsilon) H_i(u_t^\epsilon) \right) \circ dB_t^k + \sum_{i=1}^{n} \sigma_0^i(u_t^\epsilon) H_i(u_t^\epsilon)\, dt \\ \qquad + \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} \left(\sum_{j=1}^{N} \theta_k^j(u_t^\epsilon) A_j^*(u_t^\epsilon) \right) \circ dW_t^k + \frac{1}{\epsilon} \sum_{j=1}^{N} \theta_0^j(u_t^\epsilon) A_j^*(u_t^\epsilon) dt, \\ u_0^\epsilon = u_0. \end{cases}$$

We proceed to compute the equations for the slow and for the fast variables. Let $x_t^\epsilon = \pi(u_t^\epsilon)$. Then x_t^ϵ has a horizontal lift, see e.g. [4, 30, 31]. See also [29] and [23]. Let TR_g denote the differential of R_g. For $k = 0, 1, \ldots, m_1$, set

$$X_k(ug) = \sum_{i=1}^{p} \sigma_k^i(ug)\, TR_{g^{-1}} H_i(ug).$$

Below we deduce an equation for x_t^ϵ which is typically not autonomous.

Lemma 2.1 *The horizontal lift processes satisfy the following system of slow-fast SDE's:*

$$\begin{aligned} d\tilde{x}_t^\epsilon &= \sum_{k=1}^{m_1} X_k\left(\tilde{x}_t^\epsilon g_t^\epsilon\right) \circ dB_t^k + X_0\left(\tilde{x}_t^\epsilon g_t^\epsilon\right) dt, \quad \tilde{x}_0^\epsilon = g_0 \\ dg_t^\epsilon &= \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} \left(\sum_{j=1}^{N} \theta_k^j(\tilde{x}_t^\epsilon g_t^\epsilon) A_j^*(g_t^\epsilon) \right) \circ dW_t^k + \frac{1}{\epsilon} \sum_{j=1}^{N} \theta_0^j(\tilde{x}_t^\epsilon g_t^\epsilon) A_j^*(g_t^\epsilon)\, dt, \ g_0^\epsilon = id. \end{aligned} \tag{6}$$

Proof Since $\tilde{x}_t^\epsilon$ and u_t^ϵ belong to the same fibre we may define $g_t^\epsilon \in G$ by $u_t^\epsilon = \tilde{x}_t^\epsilon g_t^\epsilon$. If a_t is a C^1 curve in the lie group G

$$\frac{d}{dt}|_t u a_t = \frac{d}{dr}_{|_{r=0}} u a_t a_t^{-1} a_{r+t} = (a_t^{-1}\dot{a}_t)^*(u a_t).$$

It follows that

$$du_t^\epsilon = TR_{g_t^\epsilon} d\tilde{x}_t^\epsilon + (TL_{(g_t^\epsilon)^{-1}} dg_t^\epsilon)^*(u_t^\epsilon).$$

Since right translations of horizontal vectors are horizontal,

$$TL_{(g_t^\epsilon)^{-1}} dg_t^\epsilon = \varpi(du_t^\epsilon) = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_1} \left(\sum_{j=1}^{N} \theta_k^j(u_t^\epsilon) A_j \right) \circ dW_t^k + \frac{1}{\epsilon} \sum_{j=1}^{N} \theta_0^j(u_t^\epsilon) A_j\, dt$$

Hence, denoting by A^* also the left invariant vector fields on G, we have an equation for g_t^ϵ:

$$dg_t^\epsilon = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} \left(\sum_{j=1}^{N} \theta_k^j(u_t^\epsilon) A_j^*(g_t^\epsilon) \right) \circ dW_t^k + \frac{1}{\epsilon} \sum_{j=1}^{N} \theta_0^j(u_t^\epsilon) A_j^*(g_t^\epsilon)\, dt.$$

Since $\pi_*(A_j) = 0$ and by the definition of H_i we also have,

$$dx_t^\epsilon = \sum_{k=1}^{m_1} \left(\sum_{i=1}^{p} \sigma_k^i(u_t^\epsilon)\, (u_t^\epsilon e_i) \right) \circ dB_t^k + \sum_{i=1}^{n} \sigma_0^i(u_t^\epsilon)(u_t^\epsilon e_i)\, dt.$$

Its horizontal lift is given by $d\tilde{x}_t = \mathfrak{h}_{\tilde{x}_t}(\circ dx_t^\epsilon)$ and so we have the following SDE

$$d\tilde{x}_t^\epsilon = \sum_{k=1}^{m_1} \left(\sum_{i=1}^{p} \sigma_k^i(u_t^\epsilon)\, \mathfrak{h}_{\tilde{x}_t^\epsilon}(u_t^\epsilon e_i) \right) \circ dB_t^k + \sum_{i=1}^{n} \sigma_0^i(u_t^\epsilon) \mathfrak{h}_{\tilde{x}_t^\epsilon}(u_t^\epsilon e_i)\, dt.$$

Since $\mathfrak{h}_u(uge_i) = TR_{g^{-1}} \mathfrak{h}_{ug}(uge_i) = TR_{g^{-1}} H_i(ug)$, we may rewrite the above equation in the following more convenient form:

$$d\tilde{x}_t^\epsilon = \sum_{k=1}^{m_1} \left(\sum_{i=1}^{p} \sigma_k^i(\tilde{x}_t^\epsilon g_t^\epsilon)\, TR_{g_t^\epsilon}^{-1} H_i(\tilde{x}_t^\epsilon g_t^\epsilon) \right) \circ dB_t^k + \sum_{i=1}^{n} \sigma_0^i(\tilde{x}_t^\epsilon g_t^\epsilon) TR_{g_t^\epsilon}^{-1} H_i(\tilde{x}_t^\epsilon g_t^\epsilon)\, dt. \tag{7}$$

Finally we also rewrite the equation for the fast variable in terms of the fast and slow splitting:

$$dg_t^\epsilon = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} \left(\sum_{j=1}^{N} \theta_k^j(\tilde{x}_t^\epsilon g_t^\epsilon) A_j^*(g_t^\epsilon) \right) \circ dW_t^k + \frac{1}{\epsilon} \sum_{j=1}^{N} \theta_0^j(\tilde{x}_t^\epsilon g_t^\epsilon) A_j^*(g_t^\epsilon)\, dt. \tag{8}$$

This completes the proof.

If θ_k^j are lifts of functions from M, i.e. equi-variant functions, then the system of SDEs for g_t^ϵ do not depend on the slow variables. Define

$$\mathscr{L}_u f(g) = \frac{1}{2} \sum_{k=1}^{m_1} \left(\theta_k^j(ug) A_j^*(g) \right)^2 f(g) + \sum_{j=1}^{N} \theta_0^j(ug) A_j^*(g) f(g).$$

The matrix with entries $\Theta_{i,j} = \sum_{k=1}^{m_1} \theta_k^j \theta_k^i$ measures the ellipticity of the system.

In Sect. 6.3 we state an averaging principle for this system of slow-fast equations.

2.5 Completely Integrable Stochastic Hamiltonian Systems

In [62] a completely integrable Hamiltonian system (CISHS) in an $2n$ dimensional symplectic manifolds is introduced, which has n Poisson commuting Hamiltonian functions. After some preparation this reduces to a slow-fast system in the action angle components.

We begin comparing this model with the very well studied random perturbation problem $dx_t = (\nabla H)^{\perp}(x_t)dt + \epsilon dB_t$ where B is a real valued Brownian motion, $H : \mathbb{R}^2 \to \mathbb{R}$, and $(\nabla H)^{\perp}$ is the skew gradient of H. In the more recent CISHS model, the energy function is assumed to be random and of the form $\dot{B}_t$ so we have the equation $dx_t = (\nabla H)^{\perp}(x_t) \circ dB_t$. In both cases $H(x_t) = H(x_0)$ for all time, so H is a conserved quantity for the stochastic system. Suppose that the CISHS system is perturbed by a small vector field, we have the family of equations

$$dx_t^{\epsilon} = (\nabla H)^{\perp}(x_t^{\epsilon}) \circ dB_t + \epsilon V(x_t)dt.$$

Given a perturbation transversal to the energy surface of the Hamiltonians, one can show that the energies converge on $[1, \frac{1}{\epsilon}]$ to the solution of a system of ODEs. If moreover the perturbation is Hamiltonian, the limit is a constant and one may rescale time and find an effective Markov process on the scale $1/\epsilon^2$. The averaging theorem was obtained from studying a reduced system of slow and fast variables. The CISHS reduces to a system of equations in (H, θ), the action angle coordinates, where $H \in \mathbb{R}^n$ is the slow variable and $\theta \in S^n$ is the fast variables.

$$\frac{d}{dt}H_t^i = \epsilon f(H_t^{\epsilon}, \theta_t^{\epsilon}),$$

$$d\theta_t^i = \sum_{i=1}^{n} X_i(H_t^{\epsilon}, \theta_t^{\epsilon}) \circ dW_t^i + \epsilon X_0(H_t^{\epsilon}, \theta_t^{\epsilon})dt.$$

This slow-fast system falls, essentially, into the scope of the article.

3 Ergodic Theorem for Fredholm Operators Depending on a Parameter

Birkhoff's theorem for a sample continuous Markov process is directly associated to the solvability of the elliptic differential equation $\mathscr{L}u = v$ where $\mathscr{L}$ is the diffusion operator (i.e. the Markov generator) of the Markov process and v is a given function. A function v for which $\mathscr{L}u = v$ is solvable should satisfy a number of independent constraints. The index of the operator $\mathscr{L}$ is the dimension of the solutions for the homogeneous problem minus the dimension of the independent constraints.

Definition 3.1 A linear operator $T : E \to F$, where E and F are Hilbert spaces, is said to be a Fredholm operator if both the dimensions of the kernel of T and the dimension of its cokernel $F/\text{Range}(T)$ are finite dimensional. The Fredholm index of a Fredholm operator T is defined to be

$$\text{index}(T) = \dim(\ker(T)) - \dim(\text{cokernel}(T)).$$

A Fredholm operator T has also closed range and $E_2/\text{Range}(T) = \ker(T^*)$.

A smooth elliptic diffusion operator on a compact space is Fredholm. It also has a unique invariant probability measure. The Poisson equation $\mathscr{L}u = v$ is solvable for a function $v \in L^2$ if and only if v has null average with respect to the invariant measure, the latter is the centre condition used in diffusion creations.

If we have a family of operators $\{\mathscr{L}_x : x \in N\}$ satisfying Hörmander's condition where x is a parameter taking values in a manifold N, the parameter space is typically the state space for the slow variable, we will need a continuity theorem on the projection operator $f \mapsto \bar{f}$. We give a theorem on this in case each $\mathscr{L}_x$ has a unique invariant probability measure. It is clear that for each bounded measurable function f, $\int f(z)d\mu_x(z)$ is a function of x. We study its smooth dependence on x.

For the remaining of the section, for $i = 0, 1, \ldots, m$, let $Y_i : N \times G \to TG$ be smooth vector fields and let $\mathscr{L}_x = \frac{1}{2}\sum_{i=1}^{m} Y_i^2(x, \cdot) + Y_0(x, \cdot)$.

Definition 3.2 If $\mathscr{L}_x$ satisfies Hörmander's condition, let $r(x, y)$ denote the minimum number for the vector fields and their iterated Lie brackets up to order $r(x, y)$ to span T_yG. Let $r(x) = \inf_{y\in G} r(x, y)$. If G is compact, $r(x)$ is a finite number and will be called the rank of $\mathscr{L}_x$.

Let $s \geq 0$, let dx denote the volume measure of a Riemannian manifold G and let Δ denote the Laplacian. If f is a C^∞ function we define its Sobolev norm to be

$$\|f\|_s = \left(\int_M f(x)(I + \Delta)^{s/2} f(x)\, dx\right)^{\frac{1}{2}}$$

and we let H_s denote the closure of C^∞ functions in this norm. This can also be defined without using a Riemannian structure. If $\{\lambda_i\}$ is a partition of unity subordinated to a system of coordinates $\{\phi_i, u_i\}$, then the above Sobolev norm is equivalent to the norm $\sum_i \|(\lambda_i f) \circ \phi_i\|_s$. For a compact manifold, the Sobolev spaces are independent of the choice of the Riemannian metric. Let us denote by $|T|$ the operator norm of a linear map T.

Suppose that $\mathscr{L}_x$ satisfies Hörmander's condition. Let us re-name the vector fields Y_i and their iterated Lie brackets up to order $r(x)$ as $\{Z_k\}$. Let us define the quadratic form

$$Q^x(y)(df, df) = \sum_i |df(Z_i(x, y))|^2.$$

Then $Q^x(y)$ measures the sub-ellipticity of the operator. Let

$$\gamma(x) = \inf_{|\xi|=1} Q^x(y)(\xi, \xi).$$

Then $\gamma(x)$ is locally bounded from below by a positive number.

We summarise the properties of Hörmander type operators in the proposition below. Let $\mathscr{L}_x^*$ denote the L_2 adjoint of $\mathscr{L}_x$. An invariant probability measure for $\mathscr{L}_x$ is a probability measure such that $\int_G \mathscr{L}_x f(y)\mu_x(dy) = 0$ for any f in the domain of the generator.

Proposition 3.3 *Suppose that each $\mathscr{L}_x$ satisfies Hörmander's condition and that G is compact. Then the following statements hold.*

(1) *There exists a positive number $\delta(x)$ such that for every $s \in \mathbb{R}$ there exists a constant $C(x)$ such that for all $u \in C^\infty(G;\mathbb{R})$ the following sub-elliptic estimates hold,*

$$\|u\|_{s+\delta} \le C(\|\mathscr{L}_x u\|_s + |u|_{L_2}), \quad \|u\|_{s+\delta} \le C(\|\mathscr{L}_x^* u\|_s + |u|_{L_2}).$$

We may and will choose $C(x)$ to be continuous and $\delta(x)$ to be locally bounded from below. If r is bounded there exists $\delta_0 > 0$ such that $\delta(x) \ge \delta_0$.

(2) *$\mathscr{L}_x$ and $\mathscr{L}_x^*$ are hypo-elliptic.*

(3) *$\mathscr{L}_x$ and $\mathscr{L}_x^*$ are Fredholm and* $\operatorname{index}(\mathscr{L}_x) = 0$.

(4) *If the dimension of* $\ker(\mathscr{L}_x)$ *is* 1, *then* $\ker(\mathscr{L}_x)$ *consists of constants.*

Proof It is clear that Hörmander's condition still holds if we change the sign of the drift Y_0, or add a zero order term, or add a first order term which can be written as a linear combination of $\{Y_0, Y_1, \ldots, Y_m\}$. Since

$$\mathscr{L}_x^* = \frac{1}{2}\sum_{i=1}^m (Y_i)^2 - Y_0 - \sum_i \operatorname{div}(Y_i) Y_i + \operatorname{div}(Y_0) - \frac{1}{2}\sum_i L_{Y_i}\operatorname{div}(Y_i) + \frac{1}{2}\sum_i [\operatorname{div}(Y_i)]^2,$$

$\mathscr{L}_x$ satisfies also Hörmander's condition.

By a theorem of Hörmander in [48], there exists a positive number $\delta(x)$, such that for every $s \in \mathbb{R}$ and all $u \in C^\infty(G;\mathbb{R})$,

$$\|u\|_{s+\delta(x)} \le C(x)(\|\mathscr{L}_x u\|_s + |u|_{L_2}).$$

The constant $C(x)$ may depend on s, the L_∞ bounds on the vector fields and their derivatives, and on the rank $r(x)$, and the sub-ellipticity constant $\gamma(x)$. The constant $\delta(x)$ in the sub-elliptic estimates depend only on how many number of brackets are needed for obtaining a basis of the tangent spaces, we can for example take $\delta(x)$ to be $\frac{1}{r(x)}$. The number of brackets needed to obtain a basis at $T_y G$ is upper semi-continuous in y and is bounded for a compact manifold. Since $\mathscr{L}_x$ varies smoothly in x, then for $x \in D$ there is a uniform upper bound on the number of brackets

needed. Also as indicated in Hörmander's proof [48], the constant $C(x)$ depends smoothly on the vector fields. If there exists a number k_0 such that $r(x) \leq k_0$ for all x, then we can choose a positive δ that is independent of x. This proves the estimates in part (1) for both $\mathscr{L}_x$ and $\mathscr{L}_x^*$. The hypo-ellipticity of $\mathscr{L}_x$ and $\mathscr{L}_x^*$ is the celebrated theorem of Hörmander and follows from his sub-elliptic estimates, this is part (2).

For part (3) we only need to work with $\mathscr{L}_x$. We sketch a proof for $\mathscr{L}_x$ to be Fredholm as a bounded operator from its domain with the graph norm to L_2. From the sub-elliptic estimates it is easy to see that $\mathscr{L}_x$ has compact resolvents and that $\ker(\mathscr{L}_x)$ and $\ker(\mathscr{L}_x^*)$ are finite dimensional. Then a standard argument shows that $\mathscr{L}_x$ has closed range: If $\mathscr{L}_x f_n$ converges in L_2, then either the sequence $\{f_n\}$ is bounded in which case they are also bounded in H_δ the latter is compactly embedded in L_2, and therefore has a convergent sub-sequence. Let us denote g a limit point. Then since $\mathscr{L}_x$ if closed, g satisfies that $\mathscr{L}g = \lim_{n\to\infty} \mathscr{L}_x f_n$. If $\{f_n\}$ is not L_2 bounded, we can find another sequence $\{g_n\}$ in the kernel of $\mathscr{L}$ such that $f_n - g_n$ is bounded to which the previous argument produces a convergent sub-sequence. The dimension of the cokernel is the dimension of the kernel of $\mathscr{L}_x^*$, proving the Fredholm property. That it has zero index is another consequence of the sub-elliptic estimates and can be proved from the definition and is an elementary (using properties of the eigenvalues of the resolvents and their duals), see [92]. Part (4) is clear as constants are always in the kernel of $\mathscr{L}_x$. □

If μ_1 and μ_2 are two probability measures on a metric space M we denote by $|\mu - \nu|_{TV} = \sup_{A\in\mathscr{F}} |\mu(A) - \nu(A)|$ their total variation norm and W_1 their Wasserstein distance:

$$W_1(\mu_1, \mu_2) = \inf_{\nu} \int_{M\times M} \rho(x, y)\nu(x, y)$$

where ρ is the distance function and the infimum is taken over all couplings of μ_1 and μ_2. Suppose that $\mathscr{L}_x$ has an invariant probability measure $\mu_x(dy) = q(x, y)dy$. If for a constant K, $|q(x_1, y) - q(x_2, y)| \leq K\rho(x_1, x_2)$ for all $x_1 \in M$, $x_2 \in M$, $y \in G$, then $|\mu^{x_1} - \mu^{x_2}|_{TV} \leq K\rho(x_1, x_2)$.

Let μ_x be an invariant probability measure for $\mathscr{L}_x$. We study the regularity of the densities of the invariant probability measures with respect to the parameter, especially the continuity of the invariant probability measures in the total variation norm. This can be more easily obtained if $\mathscr{L}_x$ are Fredholm operators on the same Hilbert space and if there is a uniform estimate on the resolvent. For a family of uniformly strict elliptic operators, these are possible.

Remark 3.4 For the existence of an invariant probability measure, we may use Krylov-Bogoliubov theorem which is valid for a Feller semi-group: Let $P_t(x, \cdot)$ be the transition probabilities. If for some probability measure μ_0 and for a sequence of numbers T_n with $T_n \to \infty$, $\{Q_n(\cdot) = \frac{1}{T_n}\int_0^{T_n}\int_M P_t(x, \cdot)d\mu_0(x)dt, n \geq 1\}$ is tight, then any limit point is an invariant probability measure. The existence of an invariant probability measure is trivial for a Feller Markov process on a compact

space. Otherwise, a Laypunov function is another useful tool. See [28, 42, 44, 45] for relevant existence and uniqueness theorems.

Remark 3.5 Our operators $\mathscr{L}_x$ are Fredholm from their domains to L_2. On a compact manifold $\mathscr{L}_x$ is a bounded operator from $W^{2,2}$ to L_2 but this is only an extension of $\mathscr{L}_x$, where $W^{2,2}$ denotes the standard Sobolev space of functions, twice weakly differentiable with derivatives in L_2. We have $W^{2,2} \subset \mathrm{Dom}(\mathscr{L}_x) \subset W^{\delta(x),2}$. Due to the directions of degeneracies the domain of $\mathscr{L}_x$, given by its graph norm, can be larger than $W^{2,2}$. Since the points of the degeneracies of $\mathscr{L}_x$ move, in general, with x, their domain also change with x. Suppose that $\mathscr{L}_x$ has zero Fredholm index, then $\mathscr{L}_x$ is an isometry from $[\ker(\mathscr{L}_x)]^\perp$ to its image and $\mathscr{L}_x^*$ is invertible on $N^\perp$, the annihilator of the kernel of $\mathscr{L}_x$. Set

$$A(x) = \left|(\mathscr{L}_x^*)^{-1}_{N_x^\perp}\right|_{op}.$$

In the following proposition we consider the continuity of μ^x.

Proposition 3.6 *Let G be compact. Suppose that $Y_i \in BC^\infty$ and the conclusions of Proposition 3.3. Suppose also that each $\mathscr{L}_x$ has a unique invariant probability measure $\mu^x(dy)$.*

(i) *Let $q(x, y)$ denote the kernel of $\mu^x(dy)$. Then q and its derivatives in y are locally bounded in x.*
If the rank r is bounded from above, γ is bounded from below, then q and its derivatives in y are bounded, i.e. $\sup_x |\nabla^{(k)}\rho(x, \cdot)|_\infty$ is finite for any $k \in \mathscr{N}$.
(ii) *The kernel q is smooth in both variables.*
(iii) *Let D be a compact subset of N. There exists a number c such that for any $x_1, x_2 \in D$, $|\mu^{x_1} - \mu^{x_2}|_{TV} \le c\rho(x_1, x_2)$.*
(iv) *Suppose furthermore that r is bounded from above, γ is bounded from below, and A is bounded, then μ^x is globally Lipschitz continuous in x and $q \in BC^\infty(N \times G)$.*

Proof Each function q solves the equation $\mathscr{L}_x^* q = 0$ where $\mathscr{L}_x^*$ is the L^2 adjoint of $\mathscr{L}_x$. Since $\mathscr{L}_x^*$ is hypo-elliptic, then for each x, $q(x, \cdot)$ is C^∞. In other words, $q(x, \cdot)$ is a function from M to $C^\infty(G, \mathbb{R})$. We observe that $q(x, \cdot)$ are probability densities, so bounded in L^1. If we take s to be a number smaller than $-n/2$, n being the dimension of the manifold, then $|q(x, \cdot)|_s \le C|q(x, \cdot)|_{L^1(G)}$. We apply the sub-elliptic estimates in part (1) of Proposition 3.3 to q:

$$\|u\|_{s+\delta(x)} \le c_0(x)(\|\mathscr{L}_x^* u\|_s + \|u\|_s),$$

where $\delta(x)$ and $c(x)$ are constants, and obtain that $|q(x, \cdot)|_{s+\delta(x)} \le C(x)$. Iterating this we see that for all s,

$$|q(x, \cdot)|_s \le C(\delta(x), r(x), \gamma(x), Y).$$

The function $C(x)$ depends on the L^∞ norms of the vector fields Y_i and their covariant derivatives, and also on $\gamma(x)$. Also, δ can be taken to be $\frac{1}{r(x)+1}$ and $r(x)$ is locally bounded. By the Sobolev embedding theorems, q and the norms of its derivatives in y are locally bounded in x. (If furthermore r and γ are bounded, Y_i and their derivatives in x are bounded, then both δ and C can be taken as a constant, in which case q and their derivatives in y are bounded.)

Since q is in L^1, its distributional derivative in the x-variable exists and will be denoted by $\partial_x q$. For each x, $\mathscr{L}_x^* q = 0$, and so the distributional derivative in x of $\mathscr{L}_x^* q$ vanishes and

$$\partial_x(\mathscr{L}_x^*)q(x, y) + \mathscr{L}_x^* \partial_x q(x, y) = 0.$$

Set

$$g(x, y) = -\left(\partial_x(\mathscr{L}_x^*)\right) q(x, y).$$

Then g is smooth in y, whose Sobolev norms in y are locally bounded in x. Since the distributional derivative of q in x satisfies $\int_G \mathscr{L}_x^*(\partial_x q)(x, y)\, dy = 0$ for every x, $\int_G g(x, y)\, dy$ vanishes also. Since the index of $\mathscr{L}_x$ is zero, the invariant measure is unique, the dimension of the kernel of $\mathscr{L}_x$ is 1. The kernel consists of only constants and so $g(x, \cdot)$ is an annihilator of the kernel of $\mathscr{L}$. By Fredholm's alternative, this time applied to $\mathscr{L}_x^*$, we see that for each x we can solve the Poisson equation

$$\mathscr{L}_x^* G(x, y) = g(x, y).$$

Furthermore, by the sub-elliptic estimates, $|G(x, \cdot)|_{L_2(G)} \leq A(x)|g(x, \cdot)|_{L_2(G)}$ for some number $A(x)$. Since A is locally bounded, then $G(x, y)$ has distributional derivative in x. But $\partial_x q(x, y)$ also solves $\mathscr{L}_x^* \partial_x q(x, y) = g(x, y)$, by the uniqueness of solutions we see that $\partial_x q(x, y) = G(x, y)$. Thus the distributional derivative of q in x is a locally integrable function. Iterating this procedure and use sub-elliptic estimates to pass to the supremum norm we see that $q(x, y)$ is C^∞ in x with its derivatives in x locally bounded, in particular for a locally bounded function c_1,

$$\sup_{y \in G} |\partial_x q(x, y)| \leq A(x)c_1(x).$$

Finally, let f be a measurable function with $|f| \leq 1$. Then

$$\left| \int_G f(y)q(x_1, y)dy - \int_G f(y)q(x_2, y)dy \right| \leq \sup_{x \in D} A(x) \sup_{x \in D} c_1(x)\ \rho(x_1, x_2),$$

where D is a relatively compact open set containing a geodesic passing through x_1 and x_2. We use the fact that the total variation norm between two probability measures μ and ν is $\frac{1}{2} \sup_{|g| \leq 1} \left| \int g\, d\mu - \int g\, d\nu \right|$ where the supremum is taken

over the family of measurable functions with values in $[-1, 1]$ to conclude that $|\mu^{x_1} - \mu^{x_2}|_{TV} \leq \sup_{x\in D} A(x)\, \rho(x_1, x_2)$ and conclude the proof. □

Example 3.7 *An example of a fast diffusion satisfying all the conditions of the proposition is the following on* S^1 *and take* $x \in \mathbb{R}$*:*

$$dy_t = \sin(y_t + x)dB_t + \cos(y_t + x)dt.$$

Then $\mathscr{L}_x = \cos(x + y)\frac{\partial}{\partial y} + \frac{1}{2}\sin^2(x + y)\frac{\partial^2}{\partial y^2}$ *satisfies Hörmander's condition, has a unique invariant probability measure and* $r(x) = 1$*. Furthermore the resolvent of* $\mathscr{L}_x$ *is bounded in* x*.*

Definition 3.8 The operator $\mathscr{L}_x$ is said to satisfy the parabolic Hörmander's condition if $\{Y_1(x, \cdot), \ldots, Y_{m_2}(x, \cdot)\}$ together with the brackets and iterated the brackets of $\{Y_0(x, \cdot), Y_1(x, \cdot), \ldots, Y_{m_2}(x, \cdot)\}$ spans the tangent space of N at every point. Let $P^x(t, y_0, y)$ denote the semigroup generated by $\mathscr{L}$.

Remark 3 Suppose that each $\mathscr{L}_x$ is symmetric, satisfies the parabolic Hörmander's condition and the following uniform Doeblin's condition: there exists a constant $c \in (0, 1]$, $t_0 > 0$, and a probability measure ν such that

$$P_t^x(y_0, U) \geq c\nu(U),$$

for all $x \in N$, $y_0 \in G$ and for every Borel set U of G, Suppose that $Y_j \in BC^\infty$. Then $A(x)$ is bounded. In fact for any f with $\int f(y)\mu^x(dy) = 0$, the function $P_t^x f(y_0) = \int_G f(y)P^x(t, y_0, dy)$ converges to 0 as $t \to \infty$ with a uniform exponential rate. Since $\mathscr{L}_x$ satisfies the parabolic Hörmander's condition, $\mathscr{L} - \frac{\partial}{\partial x}$ satisfies Hörmander's condition on $M \times \mathbb{R}$. Then by the sub-elliptic estimates for $\mathscr{L} - \frac{\partial}{\partial t}$, $P_t^x f$ converges also in L_2. Let R^x denote the resolvent of $\mathscr{L}_x$. Since

$$\langle R^x f, f\rangle_{L_2} = \int_G \int_0^\infty P_t^x f(y) f(y)\, dt\, dy,$$

then R^x is uniformly bounded. Since $\mathscr{L}_x$ is symmetric, this gives a bound on $A(x)$. We refer to the book [6] for studies on Poincaré inequalities for Markov semigroups.

Corollary 3.9 *Let* G *be compact. Then* q *is smooth in both variables and in* $BC^\infty(N \times G)$*. In particular* $\mu^x = q(x, y)dy$ *is globally Lipschitz continuous.*

Just note that the semigroups P_t^x converges to equilibrium with uniform rate. The spectral gap of $\mathscr{L}_x$ is bounded from below by a positive number.

The following is a version of the law of large numbers.

Theorem 3.10 *Let* G *be compact. Suppose that* $\sum_{j=1}^{m_2} |Y_j|_\infty$ *is finite, and the conclusions of Proposition 3.3. Suppose that each* $\mathscr{L}_x$ *has a unique invariant probability measure* μ_x*.*

Let $s > 1 + \frac{\dim(G)}{2}$. Then there exists a constant $C(x)$ such that for every $x \in N$ and for every smooth real valued function $f : N \times G \to \mathbb{R}$ with compact support in the first variable (or independent of the first variable),

$$\sqrt{\mathbf{E}\left(\frac{1}{T}\int_t^{t+T} f(x, z_r^x)\, dr - \int_G f(x, y)\mu_x(dy)\right)^2} \leq C(x)\|f(x,\cdot)\|_s \frac{1}{\sqrt{T}} \tag{9}$$

where z_r^x is an $\mathscr{L}_x$ diffusion and $C(x)$ is locally bounded.

Proof In the proof we take $t = 0$ for simplicity. We only need to work with a fixed $x \in N$. We may assume that $\int_G f(x, y)\mu_x(dy) = 0$. Since $\mathscr{L}_x$ is hypo-elliptic and since μ_x is the unique invariant probability measure then, for any smooth function f with $\int_G f(x, y)\, \mu_x(dy) = 0$, $\mathscr{L}_x g(x, \cdot) = f(x, \cdot)$ has a smooth solution. If f is compactly supported in the first variable, so is g. We may then apply Itô's formula to the smooth function $g(x, \cdot)$, allowing us to estimate $\frac{1}{T}\int_0^T f(y_r^x)dr$ whose $L^2(\Omega)$ norm is controlled by the norm of g in C^1 and the norms $|Y_j(x, \cdot)|_\infty$. The $\mathscr{L}_x$ diffusion satisfies the equation:

$$\frac{1}{T}\int_0^T f(x, z_r^x)dr = \frac{1}{T}\left(g(x, z_T^x) - g(x, y_0)\right) - \frac{1}{T}\left(\sum_{k=1}^{m_2}\int_0^T dg(x,\cdot)(Y_k(x, z_r^x))dW_r^k\right).$$

Since $|Y_j(x, \cdot)|_\infty$ is bounded, it is sufficient to estimate the stochastic integral term by Burkholder-Davis-Gundy inequality:

$$\mathbf{E}\left(\sum_{k=1}^{m_2}\int_0^T dg(x,\cdot)(Y_k(x, z_r^x))dW_r^k\right)^2 \leq \sum_{k=1}^{m_1}|Y_k|_\infty^2 \int_0^T \mathbf{E}|dg(x, z_r^x)|^2\, ds.$$

It remains to control the supremum norm of $dg(x, \cdot)$. By the Sobolev embedding theorem this is controlled by the L_2 Sobolev norms $\|f(x, \cdot)\|_s$ where $s > 1+\frac{\dim(G)}{2}$. Let D be a compact set containing the supports of the functions $f(\cdot, y)$. We can choose a number $\delta > 0$, chosen according to $\sup_{x\in D} r(x)$, such that the sub-elliptic estimates holds for every $x \in D$. There exist constants c_1, c_2, c_3 such that for every $x \in D$,

$$|dg(x,\cdot)|_\infty \leq c_1\, \|g(x,\cdot)\|_{s+\delta} \leq c_2(x)\, (\|f(x,\cdot)\|_s + |g(x,\cdot)|_{L_2}) \leq c_3(x)\, \|f(x,\cdot)\|_s.$$

The constant c_2 may depend on s. The constant $c_3(x)$ is locally bounded. We have used the following fact. The spectrum of $\mathscr{L}_x$ is discrete, the dimension of the kernel space of $\mathscr{L}_x$ is 1 and hence the only solutions to $\mathscr{L}_x h = 0$ are constants. We know that the spectral distance is continuous, which is not the right reason for $c_3(x)$ to be locally bounded. To see that we may assume that f is not a constant and observe

that $|\mathscr{L}_x^{-1}|_{op} \leq k(x)$ where $k(x)$ is a finite number. This number is locally bounded following the fact that the semi-group $P_t^x f$ converges to zero exponentially and the kernels for the probability distributions of $\mathscr{L}_x$ are smooth in the parameter x. □

For the study of the limiting process in stochastic averaging we would need to know the regularity of the average of a Lipschitz continuous function with respect to one of its variables. The following illustrates what we might need.

Proposition 3.11 *Let $\{\mu^x, x \in M\}$ be a family of probability measures on G. Let $f : N \times G \to \mathbb{R}$ be a measurable function.*

(1) *Let f be a bounded function, Lipschitz continuous in the first variable, i.e. $|f(x_1, y) - f(x_2, y)| \leq K_1(y)\rho(x_1, x_2)$ with $\sup_{x\in M} |K_1|_{L_1(\mu_x)} < \infty$. Then*

$$\left| \int_G f(x_1, y)\, \mu^{x_1}(dy) - \int_G f(x_2, y)\, \mu^{x_2}(dy) \right| \leq K_2\, \rho(x_1, x_2) + |f|_\infty |\mu^{x_1} - \mu^{x_2}|_{TV}.$$

(2) *Suppose furthermore that μ_x depends continuously on x in the total variation metric. Let f be bounded continuous such that*

$$|f(x_1, z) - f(x_2, z)| \leq K_3\, \rho(x_1, x_2), \quad \forall z \in G, x_1, x_2 \in M,$$

for a positive number K_2. Then $\int_0^T \int_G f(x_s, z)\mu^{x_s}(dz)ds$ exists, and if D is the support of f then

$$\left| \sum_{i=0}^{N-1} \Delta t_i \int_G f(x_{t_i}, z)\ \mu^{x_{t_i}}(dz) - \int_0^T \int_G f(x_s, z)\mu^{x_s}(dz)ds \right|$$

$$\leq T K_3 \sup_{0\leq i<N-1} \sup_{s\in[t_i,t_{i+1})} [\rho(x_s, x_{t_i})] + |f|_\infty \cdot \sup_{0\leq i<N-1} \sup_{s\in[t_i,t_{i+1})} \left(\left|\mu^{x_s} - \mu^{x_{t_i}}\right|_{TV} \chi_{x_s\in D} \right).$$

(3) *Suppose that μ_x depends continuously on x in the Wasserstein 1-distance. Then for any bi-Lipschitz continuous f, $\int_0^T \int_G f(x_s, z)\mu^{x_s}(dz)ds$ exists and the estimate in part (1) holds with the total variation distance replaced by W_1, the Wasserstein 1-distance.*

Proof Just observe that:

$$\left| \int_G f(x_1, y)\, \mu^{x_1}(dy) - \int_G f(x_2, y)\, \mu^{x_2}(dy) \right|$$

$$\leq \int K_1(y)\mu^{x_1}(dy)\, \rho(x_1, x_2) + |f|_\infty |\mu^{x_1} - \mu^{x_2}|_{TV},$$

obtaining the required inequality in part (1) . For any non-negative numbers s, t,

$$\left|\int_G f(x_t, z)\ \mu^{x_t}(dz) - \int_G f(x_s, z)\mu^{x_s}(dz)\right|$$
$$\leq K_3\ \rho(x_t, x_s) + \left|\int_G f(x_s, z)\mu^{x_t}(dz) - \int_G f(x_s, z)\mu^{x_s}(dz)\right|.$$

This holds pathwise. Since each function $f(x_s(\omega, \cdot))$ is bounded by $|f|_\infty$,

$$\left|\int_G f(x_t, z)\ \mu^{x_t}(dz) - \int_G f(x_s, z)\mu^{x_s}(dz)\right| \leq K_3\ \rho(x_t, x_s) + |f|_\infty\ \left|\mu^{x_s} - \mu^{x_t}\right|_{TV} \chi_{x_s \in D}.$$

Since x_s is sample continuous, $x \mapsto \mu^x$ is continuous and f is a bounded and continuous, $\int_G f(x_s, z)\mu^{x_s}(dz)$ is continuous in s and so integrable in s. Consequently,

$$\left|\sum_{i=0}^{N-1} \Delta t_i \int_G f(x_{t_i}, z)\ \mu^{x_{t_i}}(dz) - \int_0^T \int_G f(x_s, z)\mu^{x_s}(dz)ds\right|$$
$$\leq \sum_{i=0}^{N-1} \Delta t_i K_3[\rho(x_s, x_{t_i})] + \sum_{i=0}^{N-1} \Delta t_i |f|_\infty[\chi_{x_s \in D}|\mu^{x_s} - \mu^{x_{t_i}}|_{TV}]$$
$$\leq T K_3 \sup_{s \in [t_i, t_{i+1})} \mathbf{E}[\rho(x_s, x_{t_i})] + |f|_\infty \sup_{s \in [t_i, t_{i+1})} \left(\chi_{x_s \in D}\left|\mu^{x_s} - \mu^{x_{t_i}}\right|_{TV}\right).$$

Finally we use the fact that f is Lipschitz in the second variable and the following dual formulation for the Wasserstein 1-distance $W_1(\mu, \nu)$ of two probability measures μ and ν,

$$W_1(\mu, \nu) = \sup_{|g|_{\text{Lip}}=1} \left|\int g\, d\mu - \int g\, d\nu\right|,$$

where $|g|_{\text{Lip}}$ denotes the Lipschitz constant of g. We obtain

$$\left|\int_G f(x_t, z)\ \mu^{x_t}(dz) - \int_G f(x_s, z)\mu^{x_s}(dz)\right| \leq K_3\ \rho(x_t, x_s) + K_4\ W_1\left(\mu^{x_s}, \mu^{x_t}\right).$$

The required assertion and estimate now follows by the argument in part (2). □

Put Propositions 3.6 and 3.10 together we obtain Theorem 1.

Finally we would like to refer to [7] for the convergence in total variation in the Law of large numbers for independent random variable, see also [56]. See the books [9, 29] for stochastic flows in sub-Riemmian geometry. It would be interesting to

study problems in this section under the 'uniformly finitely generated' conditions, see e.g. [22, 59]. See also [2, 19].

4 Basic Estimates for SDEs on Manifolds

To obtain an averaging theorem associated to a family of stochastic processes $\{x_t^\epsilon, \epsilon > 0\}$ on a manifold N, we first prove that the family of stochastic processes is pre-compact and we then proceed to identify the limiting processes. To this end we first obtain uniform estimates on the family of slow variables, on the space of continuous functions on the manifold, and also obtain estimates on the limiting Markov processes. In this section we obtain essential estimates for a general SDE and these estimates will be in terms of bounds on the driving vector fields.

Throughout this section we assume that M is a connected smooth and complete Riemannian manifold, $B_t = (B_t^1, \ldots, B_t^m)$ is an $\mathbb{R}^m$-valued Brownian motion. Let X_0 be a vector field and $X : M \times \mathbb{R}^m \to TM$ be a map linear in the second variable. For $x \in M$, let $\phi_t(x)$ denote the solution to the SDE

$$dx_t = \sum_{k=1}^{m} X_k(x_t) \circ dB_t^k + X_0(x_t)\, dt, \tag{10}$$

with initial value x. We also set $x_t = \phi_t(x_0)$.

The type of estimates we need are variation of the following $\mathbf{E}[\rho(x_s, x_t)]^2 \leq C|t-s|$ where the constant C depends on the SDE only on specific bounds for the driving vector fields. Since no ellipticity is assumed, it is essential to deal with the problem that $\rho(x, y)$ is only C^1, when x and y are on the cut locus of each other, and we cannot apply Itô's formula to ρ directly. If we are only interested in obtaining tightness results, this problem can be overcome by choosing an auxiliary distance function. Otherwise, e.g. for the convergence of the stochastic processes, we work with the Riemannian distance function $\rho : M \times M \to \mathbb{R}$. Let $M \times M$ be given the product Riemannian metric. Let $|f|_\infty$ denote the L_∞ norm of a function f.

Lemma 4.1

(1) *Suppose that M is a complete Riemannian manifold with bounded sectional curvature. Then for each $\delta > 0$ there exists a smooth distance like function $f_\delta : M \times M \to \mathbb{R}$ and a constant K_1 independent of δ such that*

$$\left|f_\delta - \rho\right|_\infty \leq \delta, \quad |\nabla f_\delta| \leq K_1, \quad |\nabla^2 f_\delta| \leq K_1.$$

If furthermore the curvature has a bounded covariant derivative, then we may also assume that $|\nabla^3 f_\delta| \leq K_1$.

(2) *If M is compact Riemannian manifold, there exists a smooth function $f : M \times M \to \mathbb{R}$ such that f agrees with ρ on a tubular neighbourhood of the diagonal set of the product manifold $M \times M$.*

Proof (1) For the distance function $\rho(\cdot, O)$, where O is a fixed point in M, this is standard, see [21, 78, 83]. To obtain the stated theorem it is sufficient to repeat the proof there for the distance function on the product manifold. The basic idea is as following. By a theorem of Greene and Wu [39], every Lipschitz continuous function with gradient less or equal to K can be approximated by C^∞ functions whose gradients are bounded by K. We apply this to the distance function ρ and obtain for each δ a smooth function $f_\delta : M \times M \to \mathbb{R}$ such that

$$|\rho - f_\delta|_\infty \leq \delta, \qquad |\nabla f_\delta|_\infty \leq 2.$$

We then convolve f_δ with the heat flow to obtain $f_\delta(x, y, t)$, apply Li-Yau heat kernel estimate for manifolds whose Ricci curvature is bounded from below and using harmonic coordinates on a small geodesic ball of radius a/K where K is the upper bound of the sectional curvature and a is a universal constant. For part (ii), M is compact. We take a smooth cut off function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that $h(t) = 1$ for $t < a$ and vanishes for $t > 2a$ where $2a$ is the injectivity radius of M and such that $|\nabla h|$ is bounded. The function $f := h \circ \rho$ is as required. □

Set $\tilde{X}_0 = \frac{1}{2}\sum_{i=1}^m \nabla X_i(X_i) + X_0$. We denote by ρ the Riemannian distance on M. Let T be a positive number and let $O \in M$. Let K', K, a_i and b_i denote constants.

Lemma 4.2 *Suppose that $\tilde{X}_0$ and X_i are C^1, where $i = 1, \ldots, m$. Suppose **one** of the following two conditions hold.*

(i) *The sectional curvature of M is bounded by K', and for every $x \in M$,*

$$|X_i(x)|^2 \leq K + K\rho(x, O), \quad |\tilde{X}_0(x)| \leq K + K\rho(x, O).$$

(ii) *Suppose that $\rho^2 : M \times M \to \mathbb{R}$ is smooth and*

$$\frac{1}{2}\sum_{i=1}^m \nabla d\rho^{2p}(X_i, X_i) + d\rho^{2p}(\tilde{X}_0) \leq K + K\rho^{2p}.$$

Then, the following statements hold.

(a) *There exists a constant c which depends only on K', T, p, and $\dim(M)$ such that for every pair of numbers s, t with $0 \leq s \leq t \leq T$,*

$$\mathbf{E}\rho^{2p}(x_t, O) \leq c(Kt + 1 + \rho^{2p}(x_0, O))e^{cKt},$$

$$\mathbf{E}\left\{\rho^{2p}(x_s, x_t)\Big|\mathscr{F}_s\right\} \leq c|t - s|(1 + K)e^{cK|t-s|}.$$

(b) *Suppose that in addition* $|X_i|$ *is bounded for every* $i = 1, \dots, m$. *Then, for every* $p \geq 1$, *there exists a constant* C, *which depends only on* p, K', m, *and* $\dim(M)$ *and a constant* $c(T)$, *such that for every* $s < t \leq T$,

$$\mathbf{E}\left(\sup_{s\leq u\leq t} \rho^{2p}(x_s, x_u)\right) \leq c + KC(T)\, e^{C(T)K}.$$

Also, $\mathbf{E}\left(\sup_{s\leq u\leq t} \rho^{2p}(O, x_u)\right) \leq c(\rho^{2p}(O, x_0) + Kc(T))\, e^{Kc(T)}$.

Proof Let $\delta \in (0, 1]$ and let $f_\delta : M \times M \to \mathbb{R}$ be a smooth function satisfying the estimates

$$\left|f_\delta - \rho\right|_\infty \leq \delta, \quad |\nabla f_\delta| \leq K_1, \quad |\nabla^2 f_\delta| \leq K_1$$

where K_1 is a constant depending on K' and $\dim(M)$. If ρ^2 is smooth we take $f_\delta = \rho$.

(a) Either hypothesis (i) or (ii) implies that the SDE (10) is conservative. For any $x \in M$ fixed we apply Itô's formula to the second variable of the function $f_\delta^2(x, y)$ on the time interval $[s, t]$:

$$f_\delta^{2p}(x, x_t) = f_\delta^{2p}(x, x_s) + \int_s^t \mathscr{L} f_\delta^{2p}(x, x_r)\, dr + \int_s^t 2f_\delta^{2p-1}(x, x_r)(df_\delta)(X_i(x_r))\, dB_r^i, \tag{11}$$

where d and $\mathscr{L}$ are applied to the second variable. Let τ_n denote the first time after s that $f_\delta(x, x_t) \geq n$ and we take the expectation of the earlier identity to obtain

$$\mathbf{E}[f_\delta^{2p}(x, x_{t\wedge\tau_n})] = \mathbf{E}[f_\delta^{2p}(x, x_s)] + \int_s^t \mathbf{E}\left[\chi_{r<\tau_n} \mathscr{L} f_\delta^{2p}(x, x_r)\right] dr.$$

Under hypothesis (ii), we use ρ in place of f_δ and conclude by Gronwall's inequality that $\mathbf{E}\rho^{2p}(x_t, O) \leq (\rho^{2p}(x, O) + Kt)e^{Kt}$. The second estimate follows from Markov property and taking $O = x_s$.

Let us now assume hypothesis (i) and let $C_1, C_2, \dots$ denote a constant depending on p. In the formula below, ∇ denotes differentiation w.r.t. the second variable,

$$\begin{aligned}\mathscr{L}[f_\delta^{2p}](x, y) =& p(2p-1)\sum_{i=1}^m f_\delta^{2p-2}(x, y)|\nabla f_\delta(X_i(y))|^2 \\ &+ p\sum_{i=1}^m f_\delta^{2p-1}(x, y)|\nabla^2 f_\delta(X_i(y), X_i(y))| + 2f_\delta^{2p-1}(x, y) df_\delta(\tilde{X}_0(y)).\end{aligned}$$

We first take $x = O$ and $s = 0$, to see that $\mathscr{L} f_\delta^{2p}(O, y) \le C_1 K f_\delta^{2p}(O, y) + C_1 K$. We may then apply Grownall's inequality followed by Fatou's lemma to obtain:

$$\mathbf{E} f_\delta^{2p}(x_t, O) \le (f_\delta^{2p}(x_0, O) + C_1 K t) e^{C_1 K t}.$$

Take $\delta = 1$, we conclude the first estimate from the following inequality:

$$\mathbf{E}[\rho^{2p}(x_t, O)] \le C_2 + C_2 \mathbf{E} f_1^{2p}(x_t, O) \le C_2 + C_3(\rho^{2p}(x_0, O) + 1 + Kt) e^{C_1 K t}.$$

Let $s < t$. Using the flow property, we see that

$$\mathbf{E}\{\rho^{2p}(x_s, x_t) | \mathscr{F}_s\} \le C_4 \delta^{2p} + C_4 \mathbf{E}\{f_\delta^{2p}(x_s, x_t) | \mathscr{F}_s\} \le C_4 \delta^{2p} + C_5 K (t-s) e^{C_5 K (t-s)}.$$

For any $s, t > 0$ we may choose δ_0 such that $\delta_0^{2p} < |t - s|$ and conclude that

$$\mathbf{E}\{\rho^{2p}(x_s, x_t) | \mathscr{F}_s\} \le C_6 (1 + K) |t - s| e^{C_6 K |t-s|}.$$

For part (b) we take $\delta = 1$ and take $p = 2$ in (11). Then

$$\begin{aligned} \mathbf{E} \sup_{u \le t} f_1^{2p}(O, x_u) =& C_1 f_\delta^{2p}(O, x_0) + C_1 \left(\int_0^t (K + K f_\delta^2(O, x_r))\, dr \right)^p \\ &+ C_1 \sum_i \mathbf{E} \left(\int_0^t 2 f_1^{2p-1}(O, x_r)(df_1)(X_i(x_r)) dr \right)^p \end{aligned}$$

Since $|X_i|$ is bounded for $i = 1, \ldots, m$, $|2 f_1(x, y)(df_1)(X_i(y))| \le 2 |f_1(x, y)| \cdot |X_i(y)|$. We conclude that

$$\mathbf{E} \sup_{0 \le u \le t} f_1^{2p}(O, x_u) \le C_2 (f_\delta^{2p}(O, x_0) + K C(T)) e^{K C(T)}.$$

This leads to the required estimates for $\mathbf{E}\left[\sup_{0 \le u \le t} \rho^{2p}(O, x_u)\right]$. Similarly, for some constants c_1 and c, depending on m and the bound of the sectional curvature, for some constants c and $C(T)$,

$$\mathbf{E}\left[\sup_{s \le u \le t} \rho^{2p}(x_s, x_u) \right] \le c_1 + c_1 K \mathbf{E}\left[\sup_{s \le u \le t} (f_1)^{2p}(x_s, x_u) \right] \le c + c K C(T) e^{K c(T)}.$$

We have completed the proof for part (b). □

These estimates will be applied in the next section to both of our slow and fast variables. For the slow variables, we have the uniform bounds on the driving vector fields and hence we obtain a uniform moment estimate (in ϵ) of the distance traveled

by the solutions. For the fast variables, the vector fields are bounded by $\frac{1}{\epsilon}$ and we expect that the evolution of the y-variable in an interval of size Δt_i to be controlled by the following quantity $\frac{\Delta t_i}{\epsilon} e^{\frac{\Delta t_i}{\epsilon}}$.

5 Proof of Theorem 2

We proceed to prove the main averaging theorem, this is Theorem 2 in Sect. 1.1.

In this section N and G are smooth complete Riemannian manifolds and $N \times G$ is the product manifold with the product Riemannian metric. We use ρ to denote the Riemannian distance on N, or on G, or on $N \times G$. This will be clear in the context and without ambiguity. For each $y \in G$ let $X_i(\cdot, y)$ be smooth vector fields on N and for each $x \in N$ let $Y_i(x, \cdot)$ be smooth vector fields on G, as given in the introduction. Let $x_0 \in N$ and $y_0 \in G$. We denote by $(x_t^\epsilon, y_t^\epsilon)$ the solution to the equations:

$$\begin{cases} dx_t^\epsilon = \sum_{k=1}^{m_1} X_k(x_t^\epsilon, y_t^\epsilon) \circ dB_t^k + X_0(x_t^\epsilon, y_t^\epsilon)\, dt, \quad x_0^\epsilon = x_0; \\ dy_t^\epsilon = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} Y_k(x_t^\epsilon, y_t^\epsilon) \circ dW_t^k + \frac{1}{\epsilon} Y_0(x_t^\epsilon, y_t^\epsilon)\, dt, \quad y_0^\epsilon = y_0. \end{cases} \tag{12}$$

Let us first study the slow variables $\{x_t^\epsilon, \epsilon \in (0, 1]\}$. We use O to denote a reference point in N.

Lemma 5.1 *Under Assumption 1, the family of stochastic processes* $\{x_t^\epsilon, \epsilon \in (0, 1]\}$ *is tight on any interval* $[0, T]$ *where* T *is a positive number. Furthermore there exists a number* C *such that for any* $p > 0$,

$$\sup_{\epsilon \in (0,1]} \sup_{s,t \in [0,T]} \mathbf{E} \rho^2(x_s^\epsilon, x_t^\epsilon) \le C|t - s|, \qquad \sup_{\epsilon \in (0,1]} \sup_{s,t \in [0,T]} \mathbf{E} \rho^{2p}(x_s^\epsilon, x_t^\epsilon) < \infty.$$

Any limiting process of x_t^ϵ, *which we denote by* $\bar{x}_t$, *has infinite life time and satisfies the same estimates:* $\mathbf{E}\rho^2(\bar{x}_s, \bar{x}_t) \le C(t - s)$ *and* $\sup_{s,t \in [0,T]} \mathbf{E}\rho^{2p}(\bar{x}_s, \bar{x}_t)$ *is finite.*

Proof Assumption 1 states that: the sectional curvature of N is bounded, $|X_i(x, y)|^2 \le K + K\rho(x, O)$ and $|\tilde{X}_0(x, y)| \le K + K\rho(x, O)$. Or $\rho^2 : N \times N \to \mathbb{R}$ is smooth, and

$$\frac{1}{2} \sum_{i=1}^{m} \nabla d\rho^2(X_i(\cdot, y), X_i(\cdot, y)) + d\rho^2(\tilde{X}_0(\cdot, y)) \le K + K\rho^2(\cdot, O).$$

In either case, the bounds are independent of the y-variable. We apply Lemma 4.2 to each x_t^ϵ to obtain estimates that are uniform in ϵ: there exists a constant C such that for all $0 \le s \le t \le T$ and for every $\epsilon > 0$, $\mathbf{E}\rho^2(x_s^\epsilon, x_t^\epsilon) \le C|t-s|$. Then use a chaining argument we obtain the following estimate for some positive constant α: $\mathbf{E}\left[\sup_{|s-t|\neq 0} \frac{\rho(x_t^\epsilon, x_s^\epsilon)}{|t-s|^\alpha}\right] < \infty$, this proves the tightness. Since $\mathbf{E}[\rho(x_t^\epsilon, O)^2]$ is uniformly bounded, we see x_t has infinite lifetime and $\mathbf{E}[\rho(x_t, O)^2]$ is finite. From the uniform estimates $\mathbf{E}\rho^2(x_s^\epsilon, x_t^\epsilon) \le C(t-s)$ and $\mathbf{E}\rho^4(x_s^\epsilon, x_t^\epsilon)^2 \le C(t-s)^2$, we easily obtain $\mathbf{E}\rho^2(\bar{x}_s, \bar{x}_t) \le C(t-s)$ and the other required estimates for $\bar{x}_s$. □

Let us fix $x \in N$. For $t \ge s$, let $\phi_{s,t}^x(y)$ denote the solution to the equation

$$dz_t = \sum_{k=1}^{m_2} Y_k(x, z_t) \circ dW_t^k + Y_0(x, z_t)\, dt, \quad z_s = y. \tag{13}$$

Write $z_t^x = \phi_{0,t}^x(z_0)$, its Markov generator is $\mathscr{L}_0^x = \frac{1}{2}\sum_{k=1}^{m_1}(Y_i(x,\cdot))^2 + Y_0(x,\cdot)$. Let $\phi_{s,t}^{\epsilon,x}$ denote the solution flow to the SDE:

$$dy_t = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} Y_k(x, y_t) \circ dW_t^k + \frac{1}{\epsilon} Y_0(x, y_t)\, dt, \quad y_s = y_0 \tag{14}$$

Observe that the time changed solution flow $\phi_{\frac{s}{\epsilon}, \frac{t}{\epsilon}}^x(\cdot)$ agrees with $\phi_{s,t}^{\epsilon,x}(\cdot)$. On each sub-interval $[t_i, t_{i+1})$ we set

$$z_t^{x_{t_i}^\epsilon} = \phi_{\frac{t_i}{\epsilon}, t}^{x_{t_i}^\epsilon}(y_{t_i}^\epsilon), \quad y_t^{x_{t_i}^\epsilon} = \phi_{t_i/\epsilon,\, t/\epsilon}^{x_{t_i}^\epsilon}(y_{t_i}^\epsilon). \tag{15}$$

In the following locally uniform law of large numbers (LLN), any rate of convergence $\lambda(t)$ is allowed.

Assumption 3 (Locally Uniform LLN) *Suppose that there exists a family of probability measures μ_x on G which is continuous in the total variation norm. Suppose that for any smooth function $g : G \to \mathbb{R}$ and for any initial point $z_0 \in G$ and $t_0 \ge 0$,*

$$\left| \frac{1}{t}\mathbf{E}\int_{t_0}^{t+t_0} g\left(\phi_{t_0,s}^x(z_0)\right) ds - \int_G g(z)\mu_x(dz) \right|_{L_2(\Omega)} \le \alpha(x)\, \|g\|_s\, \lambda(t).$$

Here $\lambda(t)$ is a constant such that $\lim_{t\to\infty}\lambda(t) = 0$, s is a non-negative number, and $\alpha(x)$ is a real number locally bounded in x.

Remark 4 In Proposition 3.10 we proved that if each $\mathscr{L}_x$ satisfies Hörmander's condition and if μ_x is the invariant probability measure for $\mathscr{L}_x$ (assume uniqueness), the locally uniform LLN holds with $\lambda(t) = \frac{1}{\sqrt{t}}$.

Suppose that $f : N \times G \to \mathbb{R}$ is bounded measurable, we define $\bar{f}(x) = \int_G f(x,z)\,\mu^x(dz)$.

Lemma 5.2 *Suppose the locally uniform LLN assumption. Let $f : N \times G \to \mathbb{R}$ be a smooth function with compact support (it is allowed to be independent of the first variable). Let $t_0 = 0 < t_1 < \cdots < t_N = T$ be a partition of equal size Δt_i. Then, for some number c,*

$$\mathbf{E}\sum_{i=0}^{N-1}\left|\int_{t_i}^{t_{i+1}} f\left(x^\epsilon_{t_i}, y_r^{x^\epsilon_{t_i}}\right) ds - \Delta t_i\, f\left(x^\epsilon_{t_i}\right)\right| \le c\,T\,\lambda(\frac{\Delta t_i}{\epsilon})\sup_{x\in D}\left\|f(x,\cdot)-\bar{f}(x)\right\|_s.$$

Proof Set $\bar{\alpha} = \sup_{x\in D}\alpha(x)$ ant $C = \sup_{x\in D}\left\|f(x,\cdot)-\bar{f}(x)\right\|_s$, both are finite numbers by the assumptions on f and on $\alpha(x)$. Firstly we observe that

$$\left|\mathbf{E}\left\{\frac{\epsilon}{\Delta t_i}\int_{\frac{t_i}{\epsilon}}^{\frac{t_{i+1}}{\epsilon}} f(x^\epsilon_{t_i}, y_r^{x^\epsilon_{t_i}})\,dr - \bar{f}(x^\epsilon_{t_i})\Big|\mathscr{F}_{t_i}\right\}\right| \le \alpha(x^\epsilon_{t_i})\,\lambda(\frac{\Delta t_i}{\epsilon})\,\chi_{x^\epsilon_{t_i}\in D}\left\|f(x^\epsilon_{t_i},\cdot)-\bar{f}(x^\epsilon_{t_i})\right\|_s.$$

Summing up over i and making a time change we obtain that

$$\begin{aligned}\left|\mathbf{E}\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} f\left(x^\epsilon_{t_i}, y_r^{x^\epsilon_{t_i}}\right) dr - \Delta t_i\,\bar{f}(x^\epsilon_{t_i})\right| &= \sum_{i=0}^{N-1}\mathbf{E}\left|\epsilon\int_{t_i/\epsilon}^{t_{i+1}/\epsilon} f\left(x^\epsilon_{t_i}, z_r^{x^\epsilon_{t_i}}\right) dr - \Delta t_i\,\bar{f}(x^\epsilon_{t_i})\right| \\ &\le \bar{\alpha}C\lambda(\frac{\Delta t_i}{\epsilon})\sum_{i=0}^{N-1}\Delta t_i,\end{aligned}$$

and thus conclude the proof. □

For the application of the LLN, we must ensure the size of the sub-interval to be sufficiently large and we should consider $\Delta t_i/\epsilon$ to be of order ∞ as $\epsilon \to 0$. Then we must ensure that $z_{\frac{t}{\epsilon}}^{x^\epsilon_{t_i}} = y_r^{x^\epsilon_{t_i}}$ is an approximation for the fast variable y_t^ϵ on the sub-interval $[t_i, t_{i+1}]$. A crude counting shows that the distance of the two, beginning with the same initial value, is bounded above by $\frac{\Delta t_i}{\epsilon}$. To obtain better estimates, we must choose the size of the interval carefully and use the slower evolutions of the slow variables on the sub-intervals and the Lipschitz continuity of the driving vector fields Y_i. We describe the intuitive idea for $\mathbb{R}^n \times \mathbb{R}^d$, assuming all vector fields are in BC^∞. We use the Lipschitz continuity of the vector fields $\frac{1}{\epsilon}Y_i$. On $[0,r]$, we have a pre-factor of $\frac{1}{\epsilon}$ from the stochastic integrals and $\frac{r}{\epsilon}$ from the deterministic interval (by Holder's inequality). Then there exists a constant C such that

$$\mathbf{E}\left|y_r^\epsilon - y_r^{x^\epsilon_{t_i}}\right|^2 \le C(\frac{1}{\epsilon}+\frac{\Delta t_i}{\epsilon^2})\int_{t_i}^r \mathbf{E}\left|x_s^\epsilon - x_{t_i}^\epsilon\right|^2 ds + C(\frac{1}{\epsilon}+\frac{\Delta t_i}{\epsilon^2})\int_{t_i}^r \mathbf{E}\left|y_s^\epsilon - y_s^{x^\epsilon_{t_i}}\right|^2 ds.$$

By Lemma 4.2, $\mathbf{E}\left|x_s^\epsilon - x_{t_i}^\epsilon\right|^2 \le \tilde{C}\Delta t_i$ on $[t_i, t_{i+1}]$ where $\tilde{C}$ is a constant and so

$$\mathbf{E}\left|y_r^\epsilon - y_r^{x_{t_i}^\epsilon}\right|^2 \le C\tilde{C}\Delta t_i\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right)e^{C\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right)}.$$

If we take Δt_i to be of the order $\epsilon|\ln\epsilon|^a$ for a suitable $a > 0$, then the above qunatity converges to zero uniformly in r as $\epsilon \to 0$. See. e.g. [32, 34, 46, 89].

In the next lemma we give the statement and the details of the computation under our standard assumptions. In particular we assume that the sectional curvature of G is bounded. Let C, c, c' denote constants.

Lemma 5.3 *Let* $0 = t_0 < t_1 < \cdots < t_N = T$ *and* $\epsilon \in (0, 1]$. *Let*

$$\alpha_i^\epsilon(C) := C\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right)e^{C\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right)} \sup_{s\in[t_i,t_{i+1}]} \mathbf{E}\rho^2\left(x_s^\epsilon, x_{t_i}^\epsilon\right).$$

1. *Suppose Assumption 2. Then there exist constants c and C such that:*

$$\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \le \alpha_i^\epsilon(C) + c\sqrt{K}\left(\alpha_i^\epsilon(C)\right)^{\frac{1}{2}}\frac{\Delta t_i}{\epsilon}e^{c\frac{\Delta t_i}{\epsilon}}$$

   where K is the bound on the sectional curvature of G.
2. *Suppose furthermore that there exists a constant c' such that*

$$\sup_{i=0,1,\ldots N-1}\ \sup_{s,t\in[t_i,t_{i+1}]}\ \sup_{\epsilon\in(0,1]} \mathbf{E}\rho^2(x_s^\epsilon, x_t^\epsilon) \le c'\,|t-s|.$$

   Then there exists a constant $C > 0$ such that for every $\epsilon \in (0, 1]$,

$$\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \le C\sqrt{\Delta t_i}\left(\frac{(\Delta t_i)^2}{\epsilon^2} + \frac{(\Delta t_i)^3}{\epsilon^3}\right)^{\frac{1}{2}} e^{C\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right)}, \quad \forall r \in [t_i, t_{i+1}], \forall i.$$

   In particular, if Δt_i is of the order $\epsilon|\ln\epsilon|^a$ where $a > 0$, then $\sup_i \sup_{r\in[t_i,t_{i+1}]} \mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right)$ is of order ϵ^δ where $\delta \in (0, \frac{1}{2})$.

Proof Since the sectional curvature of G is bounded above by K, its conjugate radius is bounded from below by $\frac{\pi}{\sqrt{K}}$. Let us consider a distance function on N that agrees with the Riemannian distance, which we denote by ρ, on the tubular neighbourhood of the diagonal of $N \times N$ with radius $\frac{\pi}{2\sqrt{K}}$. More precisely let $\tau := \tau^\epsilon$ be the first exit time when the distance between y_r^ϵ and $y_r^{x_{t_i}^\epsilon}$ is greater than or

equal to $A = \frac{\pi}{2\sqrt{K}}$. We use the identity

$$\mathbf{E}\rho^2\left(y^\epsilon_{r\wedge\tau}, y^{x^\epsilon_{t_i}}_{r\wedge\tau}\right) = \mathbf{E}\left[\rho^2\left(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r\right)\chi_{r<\tau}\right] + A^2\, P(\tau \le r),$$

to obtain that

$$P(\tau \le r) \le \frac{1}{A^2}\mathbf{E}\left[\rho^2\left(y^\epsilon_{r\wedge\tau}, y^{x^\epsilon_{t_i}}_{r\wedge\tau}\right)\right].$$

Thus,

$$\mathbf{E}\rho^2\left(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r\right) \le \mathbf{E}\left[\rho^2\left(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r\right)\chi_{r<\tau}\right] + \mathbf{E}\left[\rho^4\left(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r\right)\chi_{r\ge\tau}\right]^{\frac{1}{2}}\sqrt{P(\tau \le r)}.$$

By the earlier argument, it is sufficient to estimate $\mathbf{E}\rho^2\left(y^\epsilon_{r\wedge\tau}, y^{x^\epsilon_{t_i}}_{r\wedge\tau}\right)$, and we will show that $\mathbf{E}\rho^2\left(y^\epsilon_{r\wedge\tau}, y^{x^\epsilon_{t_i}}_{r\wedge\tau}\right)$ converges to zero sufficiently fast as $\epsilon \to 0$ to compensate with the possible divergence from the factor $\left(\mathbf{E}\rho^4\left(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r\right)\right)^{\frac{1}{2}}$.

On $\{r < \tau\}$, x, y are not on each other's cut locus, we may apply Itô's formula to the pair of stochastic processes $(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r)$ and obtain

$$\left[\rho(y^\epsilon_r, y^{x^\epsilon_{t_i}}_r)\right]^2 = \int_{t_i}^r d\rho^2\left(\frac{1}{\sqrt{\epsilon}}\sum_{k=1}^{m_2} Y_k(x^\epsilon_s, y^\epsilon_s)\circ dW^k_s + \frac{1}{\epsilon}Y_0(x^\epsilon_s, y^\epsilon_s)\,ds\right)$$
$$+ \int_{t_i}^r d\rho^2\left(\frac{1}{\sqrt{\epsilon}}\sum_{k=1}^{m_2} Y_k(x^\epsilon_{t_i}, y^{x^\epsilon_{t_i}}_s)\circ dW^k_s + \frac{1}{\epsilon}Y_0(x^\epsilon_{t_i}, y^{x^\epsilon_{t_i}}_s)\,ds\right).$$

Here the notation d in the first $d\rho^2$ refers to differentiation w.r.t. the first variable, as a gradient we use $\nabla^{(1)}(\rho^2)$, and the d in the second $d\rho^2$ is with respect to the second variable whose gradient is denoted by $\nabla^{(2)}(\rho^2)$. However $\nabla^{(1)}(\rho^2)(x, y) = -/\!/\nabla^{(2)}(\rho^2)(x, y)$, where $/\!/$ denotes the parallel translation of the relevant gradient vector along the geodesic from y to x. In the following let us denote by $d\rho^2$ the differential of ρ^2 w.r.t to the first variable. Using the assumption that each Y_k, $k = 1, \ldots, m_2$, has bounded first order derivative, and the fact that $\nabla\rho$ and $\nabla^2\rho$ are bounded, the latter follows from the assumption that the sectional curvature is

bounded, we see:

$$\left| d\rho^2(Y_k)\left(x_s^\epsilon, y_s^\epsilon\right) - d\rho^2(/\!/ Y_k)\left(x_{t_i}^\epsilon, y_s^{x_{t_i}^\epsilon}\right)\right| \leq 2\rho(y_s^\epsilon, y_s^{x_{t_i}^\epsilon})\left(\rho(x_s^\epsilon, x_{t_i}^\epsilon) + \rho(y_s^\epsilon, y_s^{x_{t_i}^\epsilon})\right).$$

It is useful to observe that Y_i is a vector field on G depending on $x \in N$, so the (product) distance function on $N \times G$ is needed for the estimate. On the other hand we only need to control the Hessian of the Riemannian distance on G and the assumption on the boundedness of the sectional curvature of G suffices.

A similar estimate applies to the first order differential involving $\tilde{Y}_0$, the sum of the Stratnovich correction for the stochastic integrals and Y_0. Again we use the assumption that each Y_k where k ranges from 1 to m_2 is bounded, and $\tilde{Y}_0$ has bounded first order covariant derivative. To summing up, for a constant C independent of ϵ and i, we have

$$\mathbf{E}\rho^2\left(y_{r\wedge\tau}^\epsilon, y_{r\wedge\tau}^{x_{t_i}^\epsilon}\right) \leq C(\frac{1}{\epsilon} + \frac{\Delta t_i}{\epsilon^2})\left(\mathbf{E}\int_{t_i}^{r\wedge\tau} \rho^2\left(x_s^\epsilon, x_{t_i}^\epsilon\right) ds + \mathbf{E}\int_{t_i}^{r\wedge\tau} \rho^2\left(y_s^\epsilon, y_s^\epsilon\right) ds\right).$$

Use Gronwall's inequality we obtain that,

$$\mathbf{E}\rho^2\left(y_{r\wedge\tau}^\epsilon, y_{r\wedge\tau}^{x_{t_i}^\epsilon}\right) \leq C\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right) \sup_{s\in[t_i,t_{i+1}]} \mathbf{E}\rho^2\left(x_s^\epsilon, x_{t_i}^\epsilon\right) e^{C(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2})}. \tag{16}$$

We can now plug in the uniform estimates that $\mathbf{E}\rho^2\left(x_s^\epsilon, x_{t_i}^\epsilon\right) \leq C|t_i - s|$ we see that

$$\mathbf{E}\rho^2\left(y_{r\wedge\tau}^\epsilon, y_{r\wedge\tau}^{x_{t_i}^\epsilon}\right) \leq C\Delta t_i\left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right) e^{C(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2})}.$$

Observe that the constant here is independent of ϵ, i and independent of $r \in [t_i, t_{i+1}]$. A similar estimates hold for $\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right)\chi_{\tau>r}$.

On $\{r > \tau\}$ we use a more crude estimate, which we obtain without using estimates on the slow variables at time s and time t_i. It is sufficient to estimate $\mathbf{E}\rho^4\left(y_r^\epsilon, y_{t_i}^\epsilon\right)$ and $\mathbf{E}\rho^4\left(y_r^{x_{t_i}^\epsilon}, y_{t_i}^\epsilon\right)$. Observing that on $[t_i, t_{i+1}]$, the processes begin with the same initial point and the driving vector fields of the SDEs to which they are solutions are $\frac{1}{\epsilon}Y_i(x_r^\epsilon, \cdot)$ and $\frac{1}{\epsilon}Y_i(x_{t_i}^\epsilon, \cdot)$ respectively. We have assumed that $\sum_{k=1}^m |Y_k|$ and $\tilde{Y}_0$ are bounded. We then apply Lemma 4.2 to these SDEs. In Lemma 4.2 we take $K = \frac{c}{\epsilon}$ where c is a constant. Then we have

$$\mathbf{E}\rho^4\left(y_r^{x_{t_i}^\epsilon}, y_{t_i}^\epsilon\right) + \mathbf{E}\rho^4\left(y_r^\epsilon, y_{t_i}^\epsilon\right) \leq c\left(\Delta t_i + \frac{\Delta t_i}{\epsilon}\right) e^{c\frac{\Delta t_i}{\epsilon}}. \tag{17}$$

Again, the constant is independent of ϵ, i and independent of $r \in [t_i, t_{i+1}]$. We put the two estimates together to see that

$$\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \leq C\Delta t_i \left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right) e^{C(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2})}$$
$$+ \frac{2\sqrt{K}}{\pi}\left(C\Delta t_i \left(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2}\right) e^{C(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2})}\right)^{\frac{1}{2}} \sqrt{c}\left(\Delta t_i + \frac{\Delta t_i}{\epsilon}\right)^{\frac{1}{2}} e^{\frac{1}{2}e^{\frac{\Delta t_i}{\epsilon}}}.$$

For ϵ small the first term is small. The second factor in the second term on the right hand side is large. We conclude that for another constant $\tilde{C}$,

$$\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \leq \tilde{C}\sqrt{\Delta t_i}(1+\epsilon)^{\frac{1}{2}} \left(\frac{(\Delta t_i)^2}{\epsilon^2} + \frac{(\Delta t_i)^3}{\epsilon^3}\right)^{\frac{1}{2}} e^{\tilde{C}(\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2})}.$$

Let us suppose that $\Delta t_i \sim \epsilon |\ln \epsilon|^a$. Then the exponent $\frac{\Delta t_i}{\epsilon} + \frac{(\Delta t_i)^2}{\epsilon^2} \sim |\ln \epsilon|^{2a}$. So for a constant C',

$$\mathbf{E}\rho^2\left(y_r^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \leq C'\sqrt{\epsilon}\,|\ln\epsilon|^{2a}\, e^{\tilde{C}|\ln\epsilon|^{2a}}.$$

The right hand side is of order ϵ^δ for $\delta < \frac{1}{2}$. We conclude the proof. □

The next lemma is on the convergence of Riemannian sums in the stochastic averaging procedure and the continuity of stochastic averages of a function with respect to a family of measures μ_x.

Lemma 5.4 *Suppose that for a sequence of numbers $\epsilon_n \downarrow 0$, $x_\cdot^{\epsilon_n}$ converges almost surely in $C([0,T]; N)$ to a stochastic process $x_\cdot$. Suppose that there exists a constant $p \geq 1$ s.t. for $|s-t|$ sufficiently small,*

$$\mathbf{E}\left[\sup_{0\leq r\leq T} \rho^{2p}(x_r^\epsilon, O)\right] < \infty, \quad \mathbf{E}\rho(x_s^\epsilon, x_t^\epsilon)^2 \leq C|t-s|, \quad \forall \epsilon \in (0,1].$$

Let μ_x be a family of probability measures on G, continuous in x in the total variation norm. Let $f : N \times G \to \mathbb{R}$ be a BC^1 function. Let $0 = t_0 < t_1 < \cdots < t_N = T$ and let $C_1 = |f|_\infty K_2 + |\nabla f|_\infty$. Then, the following statements hold:

(i)

$$\sup_{t\in[0,T]} \mathbf{E}\left|\int_G f(x_t^{\epsilon_n}, z)\, \mu^{x_t^{\epsilon n}}(dz) - \int_G f(\bar{x}_t, z)\, \mu^{\bar{x}_t}(dz)\right| \to 0.$$

In particular, the following converges in L^1,

$$\left| \int_0^t \int_G f(x_s^{\epsilon_n}, z)\ \mu^{x_s^{\epsilon_n}}(dz)\, ds - \int_0^t \int_G f(\bar{x}_s, z)\ \mu^{\bar{x}_s}(dz)\, ds \right| \to 0.$$

(ii) *The following convergence is uniform in* ϵ*:*

$$\mathbf{E} \left| \sum_{i=0}^{N-1} \Delta t_i \int_G f(x_{t_i}^{\epsilon}, z)\ \mu^{x_{t_i}^{\epsilon}}(dz) - \int_0^T \int_G f(x_s^{\epsilon}) \mu^{x_s^{\epsilon}}(dz) ds \right| \to 0.$$

Consequently, the Riemannian sum $\sum_{i=0}^{N-1} \Delta t_i \int_G f(\bar{x}_{t_i}, z)\ \mu^{\bar{x}_{t_i}}(dz)$ *converges in* L^1 *to* $\int_0^T \int_G f(\bar{x}_s, z) \mu^{\bar{x}_s}(dz) ds$.

Proof Suppose that x^{ϵ_n} converges to $\bar{x}$.. We simplify the notation by assuming that $x^{\epsilon} \to x$ almost surely. We may assume that N is not compact, the compact case is easier. Let D_n be a family of relatively compact open set such that $D_n \subset B_{a_n} \subset B_{a_{n+2}} \subset D_{n+1}$ where B_{a_n} is the geodesic ball centred at O of radius a_n where $a_n \to \infty$. This exists by a theorem of Greene and Wu. For any $t \in [0, T]$ and for any $\epsilon \in (0, 1]$,

$$\begin{aligned}
&\left| \int_G f(x_t^{\epsilon}, z)\ \mu^{x_t^{\epsilon}}(dz) - \int_G f(\bar{x}_t, z)\ \mu^{\bar{x}_t}(dz) \right| \\
&\leq \int_G \left| f(x_t^{\epsilon}, z) - f(\bar{x}_t, z) \right|\ \mu^{\bar{x}_t}(dz) + \left| \int_G f(\bar{x}_t, z) \mu^{\bar{x}_t}(dz) - \int_G f(\bar{x}_t, z) \mu^{x_t^{\epsilon}}(dz) \right| \\
&\leq |\nabla f|_{\infty}\ \rho(x_t^{\epsilon}, \bar{x}_t) + |f|_{\infty} |\mu^{\bar{x}_t} - \mu^{x_t^{\epsilon}}|_{TV}.
\end{aligned}$$

We have control over $\rho(x_t^{\epsilon}, \bar{x}_t)$, it is bounded by $\rho(x_t^{\epsilon}, O)$ and $\rho(\bar{x}_t, O)$. By the assumption, they are bounded in L^p, uniformly in $\epsilon \in (0, 1]$ and in $t \in [0, T]$. Similarly we also have uniform control over $P(\bar{x}_t \notin D_n)$ and $P(x_t^{\epsilon} \notin D_n)$, they are bounded by $c\frac{1}{n}$ where c is a constant. We observe that

$$|\mu^{\bar{x}_t} - \mu^{x_t^{\epsilon}}|_{TV} \leq |\mu^{\bar{x}_t} - \mu^{x_t^{\epsilon}}|_{TV}\, \chi_{\bar{x}_t \in D_n} \chi_{\bar{x}_t^{\epsilon} \in D_n} + 2(\chi_{\bar{x}_t \notin D_n} + \chi_{x_t^{\epsilon} \notin D_n})$$

and there exists c_n such that $|\mu^{\bar{x}_t} - \mu^{x_t^{\epsilon}}|_{TV}\, \chi_{\bar{x}_t \in D_n} \chi_{\bar{x}_t^{\epsilon} \in D_n} \leq c_n \rho(x_t^{\epsilon_n}, \bar{x}_t)$. We take n large, so that $P(\bar{x}_t \notin D_n)$ and $P(x_t^{\epsilon} \notin D_n)$ are as small as we want. Then for n fixed we see that the $c_n \rho(x_t^{\epsilon_n}, \bar{x}_t)$ converges, as $\epsilon \to 0$, in L^1. Thus, $\sup_{0 \leq t \leq T} \mathbf{E} |\mu^{\bar{x}_t} - \mu^{x_t^{\epsilon}}| \to 0$ and

$$\sup_{0 \leq t \leq T} \mathbf{E} \left| \int_G f(x_t^{\epsilon_n}, z)\ \mu^{x_t^{\epsilon_n}}(dz) - \int_G f(\bar{x}_t, z)\ \mu^{\bar{x}_t}(dz) \right|$$

converges to zero. This proves part (i). Since $\mathbf{E}\rho(x_s^\epsilon, x_{t_i}^\epsilon)^2 \le C|t_{i+1} - t_i|$,

$$\mathbf{E}\left|\sum_{i=0}^{N-1} \Delta t_i \int_G f(x_{t_i}^\epsilon, z)\ \mu^{x_{t_i}^\epsilon}(dz) - \int_0^T \int_G f(x_s^\epsilon)\mu^{x_s^\epsilon}(dz)ds\right|$$

$$\le T\,|\nabla f|_\infty \sup_{s\in[t_i,t_{i+1})} \mathbf{E}[\rho(x_s^\epsilon, x_{t_i}^\epsilon)] + |f|_\infty T \sup_{s\in[t_i,t_{i+1})} \mathbf{E}\left[|\mu^{x_s^\epsilon} - \mu^{x_{t_i}^\epsilon}|_{TV}\right] \to 0.$$

The convergence can be proved, again by breaking the total variation norm into two parts, in one part the processes are in D_n, and in the other part they are not. Since x_t^ϵ converges to x_t as a stochastic process on $[0, T]$, we also have that $\mathbf{E}\rho(x_s, x_{t_i}) \le C|t_{i+1} - t_i|$. We apply the same argument to $\bar{x}_t$ to obtain that

$$\mathbf{E}\left|\sum_{i=0}^{N-1} \Delta t_i \int_G f(\bar{x}_{t_i}, z)\ \mu^{\bar{x}_{t_i}}(dz) - \int_0^T \int_G f(\bar{x}_s)\mu^{\bar{x}_s}(dz)ds\right| \to 0.$$

This concludes the proof. □

Suppose we assume furthermore that there exists a constant K such that

$$|\mu^{x_1} - \mu^{x_2}|_{TV} \le K(1 + \rho(x_1, O) + \rho(x_2, O))\rho(x_1, x_2).$$

Then explicit estimates can be made for the convergence in Lemma 5.4, e.g.

$$\left|\int_G f(x_t^\epsilon, z)\ \mu^{x_t^\epsilon}(dz) - \int_G f(\bar{x}_t, z)\ \mu^{\bar{x}_t}(dz)\right|$$

$$\le |\nabla f|_\infty\ \rho(x_t^\epsilon, \bar{x}_t) + |f|_\infty K(1 + \rho^p(\bar{x}_t, O) + \rho^p(x_t^\epsilon, O))\rho(x_t^\epsilon, \bar{x}_t).$$

To this we may apply Hölder's inequality and obtain:

$$\mathbf{E}\left|\int_0^T \int_G f(x_t^{\epsilon_n}, z)\ \mu^{x_t^{\epsilon_n}}(dz)\,dt - \int_0^T \int_G f(\bar{x}_t, z)\ \mu^{\bar{x}_t}(dz)\,dt\right|$$

$$\le |\nabla f|_\infty\ \mathbf{E}\int_0^T \rho(x_t^\epsilon, \bar{x}_t)\ dt + |f|_\infty K\mathbf{E}\left|\int_0^T (1 + \rho^p(\bar{x}_t, O) + \rho^p(x_t^\epsilon, O))\rho(x_t^\epsilon, \bar{x}_t)\,dt\right|$$

$$\le |\nabla f|_\infty\ T\mathbf{E}\sup_{s\le t}\rho(x_t^\epsilon, \bar{x}_t) + |f|_\infty KT\sqrt{\mathbf{E}\sup_{t\le T}(1 + \rho^p(\bar{x}_t, O) + \rho^p(x_t^\epsilon, O))^2}\sqrt{\mathbf{E}\sup_{t\le T}\rho^2(x_t^\epsilon, \bar{x}_t)}.$$

In the proposition below we are interested in the time average concerning a product function $f_1 f_2$, where $f_1 : N \to \mathbb{R}$ is C^∞ with has compact support and $f_2 : G \to \mathbb{R}$ is smooth.

Proposition 5.5 *Suppose the following conditions.*

(1) μ_x *is a family of probability measures on* G *for which the locally uniform LLN assumption (Assumption 3) holds.*
(2) *Assumption 2.*
(3) *There exist constants* $p \geq 1$ *and* c *such that for* $s, t \in [r_1, r_2]$ *where* $r_2 - r_1$ *is sufficiently small,*

$$\sup_{\epsilon\in(0,1]} \sup_{s,t\in[r_1,r_2]} \mathbf{E}\rho^2(x_s^\epsilon, x_t^\epsilon) \leq c|t-s|, \qquad \sup_{0\leq s\leq T} \sup_{\epsilon\in(0,1]} \mathbf{E}\rho^{2p}(x_s^\epsilon, O) < \infty.$$

(4) ϵ_n *is a sequence of numbers converging to* 0 *with* $\sup_{t\leq T} \rho(x_t^{\epsilon_n}, \bar{x}_t)$ *converges to zero almost surely.*
(5) *Let* $f : N \times G \to \mathbb{R}$ *be a smooth and globally Lipschitz continuous function. Suppose that either* f *is independent of the first variable or for each* $y \in G$*, the support of* $f(\cdot, y)$ *is contained in a compact set* D*.*

Then the following random variables converge to zero in L^1*:*

$$\int_0^T f(x_s^\epsilon, y_s^\epsilon)ds - \int_0^T \int_G f(\bar{x}_s, z)\, \mu^{\bar{x}_s}(dz)\, ds.$$

Proof Let $0 = t_0 < t_1 < \cdots < t_N = T$ and let $\Delta t_i = t_{i+1} - t_i$. Then, recalling the notation given in (15),

$$\begin{aligned}
&\int_0^T f(x_s^\epsilon, y_s^\epsilon)ds = \sum_{n=0}^{N-1} \int_{t_i}^{t_{i+1}} f(x_s^\epsilon, y_s^\epsilon)ds \\
&= \sum_{n=0}^{N-1} \int_{t_i}^{t_{i+1}} \left[f(x_s^\epsilon, y_s^\epsilon) - f(x_{t_i}^\epsilon, y_s^\epsilon) \right] ds + \sum_{n=0}^{N-1} \int_{t_i}^{t_{i+1}} \left[f(x_{t_i}^\epsilon, y_s^\epsilon) - f\left(x_{t_i}^\epsilon, y_r^{x_{t_i}^\epsilon}\right) \right] ds \\
&\quad + \sum_{n=0}^{N-1} \left[\int_{t_i}^{t_{i+1}} f\left(x_{t_i}^\epsilon, y_r^{x_{t_i}^\epsilon}\right) ds - \Delta t_i \int_G f(x_{t_i}^\epsilon, z)\mu^{x_{t_i}^\epsilon}(dz) \right] \\
&\quad + \left[\sum_{n=0}^{N-1} \Delta t_i \int_G f(x_{t_i}^\epsilon, z)\mu^{x_{t_i}^\epsilon}(dz) - \int_0^T \int_G f(x_s^\epsilon, z)\mu^{x_s^\epsilon}(dz)\, ds \right] \\
&\quad + \left[\int_0^T \int_G f(x_s^\epsilon, z)\mu^{x_s^\epsilon}(dz)\, ds - \int_0^T \int_G f(\bar{x}_s, z)\mu^{\bar{x}_s}(dz)\, ds \right] + \int_0^T \int_G f(\bar{x}_s, z)\mu^{\bar{x}_s}(dz)\, ds.
\end{aligned}$$

Using the fact that f is Lipschitz continuous in the first variable and the assumptions on the moments of $\rho(x_t^\epsilon, x_s^\epsilon)$ we see that for a constant K,

$$\sum_{i=0}^{N-1} \mathbf{E} \int_{t_i}^{t_{i+1}} \left| f(x_r^\epsilon, y_r^\epsilon) - f(x_{t_i}^\epsilon, y_r^\epsilon) \right| dr \le K \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbf{E}\rho(x_r^\epsilon, x_{t_i}^\epsilon)\, dr$$
$$\le K\ T\mathbf{E}\rho(x_r^\epsilon, x_{t_i}^\epsilon) \le T K c \max_i \sqrt{\Delta t_i}.$$

By choosing $\Delta t_i = o(\epsilon)$ we see that the first term on the right hand side converges to zero. The converges of the second term follows directly from Lemma 5.3 by choosing $\Delta t_i \sim \epsilon |\ln \epsilon|^a$ where $a > 0$ and Assumption 2. By Lemma 5.2 and Assumption 3, the third term converges if we choose $\frac{\epsilon}{\Delta t_i} = o(\epsilon)$. The convergence of the fourth and fifth terms follow respectively from part (i) and part (ii) of Lemma 5.4. □

We are now ready to prove the main averaging theorem, Theorem 2 in Sect. 1.1. This proof has the advantage for being concrete, from this an estimate for thew rate of convergence is also expected.

Theorem 5.6 *Suppose the following statements hold.*

(a) *Assumptions 1 and 2.*
(b) *There exists a family of probability measures μ_x on G for which the locally uniform LLN assumption (Assumption 3) holds.*

Then, as $\epsilon \to 0$, the family of stochastic processes $\{x_t^\epsilon, \epsilon > 0\}$ converges weakly on any compact time intervals to a Markov process with generator $\bar{\mathscr{L}}$.

Proof By Prohorov's theorem, a set of probability measures is tight if and only if its relatively weakly compact, i.e. every sequence has a sub-sequence that converges weakly. It is therefore sufficient to prove that every limit process of the stochastic processes x_t^ϵ is a Markov process with the same Markov generator. Every sequence of weakly convergent stochastic processes on an interval $[0, T]$ can be realised on a probability space as a sequence of stochastic processes that converge almost surely on $[0, T]$ with respect to the supremum norm in time. It is sufficient to prove that if a subsequence $\{x_t^{\epsilon_n}\}$ converges almost surely on $[0, T]$, the limit is a Markov process with generator $\bar{\mathscr{L}}$. For this we apply Stroock-Varadhan's martingale method [74, 82]. To ease notation we may assume that the whole family x_t^ϵ converges almost surely. Let f be a real valued smooth function on N with compact support. Let $\bar{x}_t$ be the limit Markov process. We must prove that $f(\bar{x}_t) - f(x_0) - \int_0^t \bar{\mathscr{L}} f(\bar{x}_r) dr$ is a martingale. In other words we prove that for any bounded measurable random variable $G_s \in \mathscr{F}_s$ and for any $s < t$, $\mathbf{E}\left(G_s(f(\bar{x}_t) - f(\bar{x}_s) - \int_s^t \bar{\mathscr{L}} f(x_r) dr)\right) = 0$. On the other hand, for each $\epsilon > 0$,

$$f(x_t^\epsilon) - f(x_s^\epsilon) - \int_s^t \left(\frac{1}{2} \sum_{i=1}^{m_1} (X_i(\cdot, y_r^\epsilon))^2 f + X_0(\cdot, y_r^\epsilon) f \right)(x_r^\epsilon)\, dr$$

is a martingale. Let us introduce the notation:

$$F(x_r^\epsilon, y_r^\epsilon) = \left(\frac{1}{2} \sum_{i=1}^{m_1} (X_i(\cdot, y_r^\epsilon))^2 f + X_0(\cdot, y_r^\epsilon) f \right)(x_r^\epsilon).$$

Since x_t^ϵ converges to $\bar{x}_s$ it is sufficient to prove that as $\epsilon \to 0$,

$$\mathbf{E}\left[G_s \left(\int_s^t F(x_r^\epsilon, y_r^\epsilon) dr - \int_s^t \bar{\mathscr{L}} f(x_r) dr \right) \right] \to 0.$$

Even simpler we only need to prove that $\int_s^t F(x_r^\epsilon, y_r^\epsilon) dr$ converges to $\int_s^t \bar{\mathscr{L}} f(x_r) dr$ in L^1. Under Assumption 1, we may apply Lemma 5.1 from which we see that conditions (3) and (4) of Proposition 5.5 hold. Since f has compact support, F has compact support in the first variable. We may apply Proposition 5.5 to the function F to complete the proof. □

We remark that the locally uniform law of large numbers hold if G is compact, if $\mathscr{L}_x$ satisfies *strong Hörmander's condition*, or if $\mathscr{L}_x$ satisfies Hörmander's condition with the additional assumption that $\mathscr{L}_x$ has a unique invariant probability measure.

We obtain the following Corollary.

Corollary 1 *Let G be compact. Suppose Assumptions 1 and 2. Suppose that $\mathscr{L}_x$ satisfies Hörmander's condition and that it has a unique invariant probability measure. Then $\{x_t^\epsilon, \epsilon > 0\}$ converges weakly, on any compact time intervals, to a Markov process with generator $\bar{\mathscr{L}}$.*

From the proof of Theorem 5.6, the Markov generator $\bar{\mathscr{L}}$ given below.

$$\bar{\mathscr{L}} f(x) = \int_G \left(\frac{1}{2} \sum_{i=1}^{m_1} X_i^2(\cdot, y) f + X_0(\cdot, y) f \right)(x) \, \mu_x(dy). \tag{18}$$

Appendix

It is possible to write the operator $\bar{\mathscr{L}}$ given by (18) as a sum of squares of vector fields. For this we need an auxiliary family of vector fields $\{E_1, \ldots, E_{n_1}\}$ with the property that at each point x they span the tangent space $T_x N$. Let us write each vector field $X_i(\cdot, y)$ in this basis and denote by $X_i^k(\cdot, y)$ its coordinate functions, so

$X_i(x,y) = \sum_{k=1}^{n_1} X_i^k(x,y)E_k(x)$. Set

$$a_{k,l}(x,y) = \sum_{i=1}^{m_1} X_i^k(x,y)X_i^l(x,y),$$

$$b_0^k(x,y) = \frac{1}{2}\sum_{l=1}^{n_1}\sum_{i=1}^{m_1} X_i^l(x,y)\left(\nabla X_k^l(\cdot,y))(E_l(x))\right) + X_0^k(x,y),$$

where ∇ denotes differentiation with respect to the first variable.We observe that

$$\frac{1}{2}\sum_{i=1}^{m_1}(X_i(\cdot,y)^2 f)(x) + (X_0(\cdot,y)f)(x) = \frac{1}{2}\sum_{k,l=1}^{n_1} a_{k,l}(x,y)(E_k E_l f)(x) + \sum_{k=1}^{n_1} b_0^k(x,y)(E_k f)(x).$$

If μ_x is a family of probability measures on G, we set

$$\bar{\mathscr{L}} = \frac{1}{2}\sum_{k,l=1}^{n_1}\left(\int_G a_{k,l}(x,y)\mu_x(dy)\right)E_k E_l + \sum_{k=1}^{n_1}\left(\int_G b_0^k(x,y)dy\right)E_k. \tag{19}$$

The auxiliary vector fields can be easily constructed. For example, we may use the gradient vector fields coming from an isometric embedding $i : N \to \mathbb{R}^{n_1}$. Then they have the following properties. For $e \in \mathbb{R}^{n_1}$, we define $E(x)(e) = \sum_{i=1}^{n_1} E_i(x)e_i$ where $\{e_i\}$ is an orthonormal basis of $\mathbb{R}^{n_1}$. Then $\mathbb{R}^{n_1}$ has a splitting of the form $\ker[X(x)]^\perp \oplus X(x)$ and $X(e)$ has vanishing derivative for $e \in \ker[X(x)]^\perp$. We may also use a 'moving frames' instead of the gradient vector fields. This is particularly useful if N is an Euclidean space, or a compact space, or a Lie group. For such spaces and their moving frames, the assumption that $X_1, \ldots, X_k$ and their two order derivatives, X_0 and ∇X_0 are bounded can be expressed by the boundedness of the functions $a_{k,l}$ and b_0^k and their derivatives.

6 Re-visit the Examples

6.1 A Dynamical Description for Hypo-elliptic Diffusions

Let us consider two further generalisations to the dynamical theory for Brownian motions described in Sect. 2.1. Both cases allow degeneracy in the fast variables. One of which has the same type of reduced random ODE and is closer to Theorem 2A. We state this one first and will take M compact for simplicity.

Proposition 6.1 *Let M be compact. Suppose that in (5), we replace the orthonormal basis $\{A_1, \ldots, A_N\}$ and A_0 by a vectors $\{A_1, \ldots A_{m_2}\} \subset \mathfrak{so}(n)$ with the property that these vectors together with their commutators generates $\mathfrak{so}(n)$. (Take*

$A_0 = 0$ *for simplicity. Then, as* $\epsilon \to 0$*, the rescaled position stochastic processes,* $x^\epsilon_{\frac{t}{\epsilon}}$*, converges to a scaled Brownian motion. Their horizontal lifts from* u_0 *converge also.*

Proposition 6.2 *The scale is determined by the eigenvalues of the symmetry matrix* $\sum_{i=1}^{m_2}(A_i)^2$.

Proof The reduced equation is as before:

$$\begin{cases} \dfrac{d}{dt}\tilde{x}^\epsilon_t = H_{\tilde{x}^\epsilon_t}(g_{\frac{t}{\epsilon}}e_0), \quad \tilde{x}^\epsilon_0 = u_0, \\ dg_t = \displaystyle\sum_{k=1}^{m_2} g_t A_k \circ dw^k_t + g_t A_0\, dt, \quad g_0 = Id. \end{cases}$$

We observe that the operator $\sum_{i=1}^{m_2}(A_i^*)^2 + A_0^*$ satisfies Hörmander's condition and has a unique invariant probability measure. It is symmetric w.r.t the bi-invariant Haar measure dg, and the only invariant measure is dg. Then we apply Theorem 6.4 from [66] to conclude.

Suppose instead we consider the following SDE, in which the horizontal part involves a stochastic integral

$$\begin{cases} du^\epsilon_t = H_{u^\epsilon_t}(e_0)\, dt + \displaystyle\sum_{j=1}^{m_1} H(u^\epsilon_t)(e_j) \circ dB^j_t + \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} A^*_k(u^\epsilon_t) \circ dW^k_t + A^*_0(u^\epsilon_t)\, dt, \\ u^\epsilon_0 = u_0. \end{cases} \tag{20}$$

where $e_j \in \mathbb{R}^n$.

Proposition 6.3 *Suppose that* M *has bounded sectional curvature. Suppose that* $\{A_0, A_1, \ldots, A_{m_2}\}$ *and their iterated brackets (commutators) generate the vector space* $\mathfrak{so}(n)$*. Suppose that* $\{e_1, \ldots, e_{m_1}\}$ *is an orthonormal set. Then as* $\epsilon \to 0$*, the position component of* u^ϵ_t*,* x^ϵ_t*, converges to a rescaled Brownian motion, scaled by* $\frac{m_1}{n}$ *where* $n = \dim(M)$*. Their horizontal lifts converge also to a horizontal Brownian motion with the same scale.*

Proof Set $x^\epsilon_t = \pi(u^\epsilon_t)$, where π takes an frame to its base point. Then x^ϵ_t is the position process. Then

$$dx^\epsilon_t = \sum_{i=1}^{m_1} u^\epsilon_t(e_i) \circ dB^i_t + u^\epsilon_t e_0\, dt.$$

Let $\tilde{x}^\epsilon_t$ denote the stochastic horizontal lifts of x^ϵ_t. Then from the nature of the horizontal vector fields and the horizontal lifts, this procedure introduces a twist

to the Euclidean vectors e_i. If g_t solves:

$$dg_t = \sum_{k=1}^{m_2} g_t A_k \circ dw_t^k + g_t A_0\, dt$$

with initial value the identity, then x_t^ϵ satisfies the equation

$$d\tilde{x}_t^\epsilon = \mathfrak{h}_{\tilde{x}_t^\epsilon} dx_t^\epsilon = H(\tilde{x}_t^\epsilon)(g_{\frac{t}{\epsilon}} e_0)dt + \sum_{i=1}^{m_1} H(\tilde{x}_t^\epsilon)(g_{\frac{t}{\epsilon}} e_i) \circ dB_t^i.$$

Since g_t does not depend on the slow variable, the conditions of the Theorem is satisfied provided M has bounded sectional curvature.

The limiting process, in this case, will not be a fixed point. It is a Markov process on the orthonormal frame bundle with generator

$$\bar{\mathscr{L}} f(u) = \sum_{i=1}^{m_1} \int_{SO(n)} \nabla df(H(u)(ge_i), H(u)(ge_i))dg,$$

where ∇ is a flat connection, ∇H_i vanishes. Let dg denote the normalised bi-invariant Haar measure. Using this connection and an orthonormal basis $\{e_i\}$ of $\mathbb{R}^n$, extending our orthonormal set $\{e_1, \ldots, e_{m_1}\}$ we see that

$$\bar{\mathscr{L}} f(u) = \sum_{k,l=1}^{n} \nabla df(H(E_k), H(E_l)) \sum_{i=1}^{m_1} \int_{SO(n)} \langle e_k, ge_i\rangle \langle ge_i, e_l\rangle dg.$$

It is easy to see that

$$\sum_{i=1}^{m_1} \int_{SO(n)} \langle e_k, ge_i\rangle \langle ge_i, e_l\rangle dg = \sum_{i=1}^{m_1} \delta_{k,l} \int_{SO(n)} \langle e_k, ge_i\rangle^2 dg = \frac{m_1}{n} \delta_{k,l}.$$

This means that

$$\bar{\mathscr{L}} f(u) = \frac{m_1}{n} \sum_{k,l=1}^{n} \nabla df(H(e_k), H(e_k)).$$

Thus the $\bar{\mathscr{L}}$ diffusion has Markov generator $\frac{1}{2}\frac{m_1}{n}\Delta^H$ where Δ^H is the horizontal diffusion and which means that $\pi(u_t^\epsilon)$ converges to a scaled Brownian motion as we have guessed. □

Problem 2 The vertical vector fields in (20) are left invariant. Instead of left invariant vertical vector fields we may take more general vector fields and consider the following SDEs. Let $f : OM \to \mathbb{R}$ be smooth functions, $e_j \in \mathbb{R}^n$ are unit vectors. Let us consider the equation

$$\begin{cases} du_t^\epsilon = H_{u_t^\epsilon}(e_0)\,dt + \sum_{j=1}^{m_1} H(u_t^\epsilon)(e_j) \circ dB_t^j + \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} f_k(u_t^\epsilon) A_k^*(u_t^\epsilon) \circ dW_t^k + A_0^*(u_t^\epsilon)\,dt, \\ u_0^\epsilon = u_0. \end{cases} \tag{21}$$

Then the horizontal lift of its position processes will, in general, depend on the slow variables. It would be interesting to determine explicit conditions on f_k for which the averaging procedure is valid and if so what is the effective limit?

6.2 Inhomogeneous Scaling of Riemannian Metrics

Returning to Sect. 2.3 we pose the following problem.

Problem 3 With Theorem 5.6, we can now study a fully coupled system:

$$dg_t^\epsilon = \frac{1}{\sqrt{\epsilon}} \sum_{k=1}^{m_2} (a_k A_k)(g_t^\epsilon) \circ dB_t^k + \frac{1}{\epsilon}(a_0 A_0)(g_0^\epsilon)dt + (b_0 Y_0)(g_t^\epsilon)dt + \sum_{k=1}^{m_1} (b_k Y_k)(g_t^\epsilon) \circ dW_t^k,$$

where a_k, b_k are smooth functions. It would be interesting to study the convergence of the slow variables, vanishing of the averaged processes, and the nature of the limits in terms of a_k and b_k.

6.3 An Averaging Principle on Principal Bundles

We return to the example in Sect. 2.4. In the following proposition, ∇ denotes the flat connection on the principal bundle P.

Proposition 6.4 *Let G be a compact Lie group and dg its Haar measure. Assume that M has bounded sectional curvature. Suppose that $\mathscr{L}_u$ satisfies Hörmander's condition and has a unique invariant probability measure. Suppose that θ_k^j are*

bounded with bounded derivatives. Define

$$a_{i,j}(u) = \int_G \sum_{l=1}^{m_1} \langle X_l(u, g), H_i(u)\rangle \langle X_l(ug), H_j(u)\rangle \, dg,$$

$$b(u) = \int_G \left(\frac{1}{2} \sum_{l=1}^{m_1} \nabla_{X_l} X_l(ug) + X_0(ug) \right) dg.$$

Then $\tilde{x}^\epsilon_{\frac{t}{\epsilon}}$ *converges weakly to a Markov process on* P *with the Markov generator*

$$\bar{\mathscr{L}} f(u) = df(b(u)) + \frac{1}{2} \sum_{i,j=1}^{n} a_{i,j}(u) \nabla df(H_i(u), H_j(u)).$$

Proof The convergence is a trivial consequence of Theorem 5.6. To identify the limit let $f : P \to \mathbb{R}$ be any smooth function with compact support. Then

$$f(\tilde{x}^\epsilon_t) = f(g_0) + \sum_{l=1}^{m_1} \int_0^t df\left(X_l((\tilde{x}^\epsilon_s g^\epsilon_s)\right) dB^l_s + \sum_{l=1}^{m_1} \int_0^t \nabla df\left(X_l(\tilde{x}^\epsilon_s g^\epsilon_s), X_l(\tilde{x}^\epsilon_s g^\epsilon_s)\right) ds$$
$$+ \sum_{l=1}^{m_1} \int_0^t df\left(\nabla_{X_l} X_l(\tilde{x}^\epsilon_s g^\epsilon_s) + X_0(\tilde{x}^\epsilon_s g^\epsilon_s)\right) ds.$$

Finally we take coordinates of X_l w.r.t the parallel vector fields H_i, cf. the Appendix of Sect. 5, to complete the proof.

Conclusions and Other Open Questions In conclusion, the examples we studied treat some of the simplest and yet universal models, they can be studied using the method we have just developed. Even for these simple models many questions remain to be answered, including the questions stated in Sects. 1.1, 2.1 and 2.3. For example we do not know the geometric nature of the limiting object. Concerning Theorem 5.6, we expect the conditions of the theorem improved for more specific examples of manifolds, and expect an upper bound for the rate of convergence if the resolvents of the operators $\mathscr{L}_x$ is bounded in x and if the rank of the operators and their quadratic forms are bounded, and also expect an averaging principle for slow-fast SDEs driven by Lévy processes, cf. [49].

References

1. Angst, J., Bailleul, I, Tardif, C.: Kinetic Brownian motion on Riemannian manifolds. Electron. J. Probab. **20**(110), 40 (2015)
2. Albeverio, S., Daletskii, A., Kalyuzhnyi, A.: Random Witten Laplacians: traces of semigroups, L^2-Betti numbers and index. J. Eur. Math. Soc. (JEMS) **10**(3), 571–599 (2008)
3. Arnol'd, V.I.: Mathematical Methods of Classical Mechanics. Springer, New York (1989)
4. Arnaudon, M.: Semi-martingales dans les espaces homogènes. Ann. Inst. H. Poincaré Probab. Statist. **29**(2), 269–288 (1993)
5. Atiyah, M.F.: Algebraic topology and operators in Hilbert space. In: Atiyah, M.F., Taam, C.T., et al. (eds.) Lectures in Modern Analysis and Applications. I, pp. 101–121. Springer, Berlin (1969)
6. Bakry, D., Gentil, I., Ledoux, M.: Analysis and Geometry of Markov Diffusion Operators. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 348. Springer, Cham (2014)
7. Bally, V., Caramellino, L.: Asymptotic development for the CLT in total variation distance. Bernoulli **22**(4), 2442–2485 (2016)
8. Barret, F., von Renesse, M.: Averaging principle for diffusion processes via Dirichlet forms. Potential Anal. **41**(4), 1033–1063 (2014)
9. Baudoin, F.: An Introduction to the Geometry of Stochastic Flows. Imperial College Press, London (2004)
10. Baudoin, F., Hairer, M., Teichmann, J.: Ornstein-Uhlenbeck processes on Lie groups. J. Funct. Anal. **255**(4), 877–890 (2008)
11. Bérard-Bergery, L., Bourguignon, J.-P.: Laplacians and Riemannian submersions with totally geodesic fibres. Illinois J. Math. **26**(2), 181–200 (1982)
12. Berglund, N., Gentz, B.: Noise-Induced phenomena in Slow-Fast Dynamical Systems. Probability and Its Applications (New York). Springer, London (2006). A sample-paths approach
13. Berline, N., Getzler, E., Vergne, M.: Heat Kernels and Dirac Operators. Springer, New York (1992)
14. Birrell, J., Hottovy, S., Volpe, G., Wehr, J.: Small mass limit of a Langevin equation on a manifold. Ann. Henri Poincaré **18**(2), 707–755 (2017)
15. Bismut, J.-M.: The hypoelliptic Laplacian on a compact Lie group. J. Funct. Anal. **255**(9), 2190–2232 (2008)
16. Bismut, J.-M., Lebeau, G.: Laplacien hypoelliptique et torsion analytique. C. R. Math. Acad. Sci. Paris **341**(2), 113–118 (2005)
17. Borodin, A.N.: A limit theorem for the solutions of differential equations with a random right-hand side. Teor. Verojatnost. i Primenen. **22**(3), 498–512 (1977)
18. Borodin, A.N., Freidlin, M.I.: Fast oscillating random perturbations of dynamical systems with conservation laws. Ann. Inst. H. Poincaré Probab. Statist. **31**(3), 485–525 (1995)
19. Cass, T., Friz, P.: Densities for rough differential equations under Hörmander's condition. Ann. Math. (2) **171**(3), 2115–2141 (2010)
20. Catellier, R., Gubinelli, M.: Averaging along irregular curves and regularisation of ODEs. Stochastic Process. Appl. **126**(8), 2323–2366 (2016)
21. Chow, B., Chu, S.-C., Glickenstein, D., Guenther, C., Isenberg, J., Ivey, T., Knopf, D., Lu, P., Luo, F., Ni, L.: The Ricci flow: techniques and applications. Part III. In: Geometric-Analytic Aspects. Mathematical Surveys and Monographs, vol. 163. American Mathematical Society, Providence (2010)
22. Crisan, D., Ottobre, M.: Pointwise gradient bounds for degenerate semigroups (of UFG type). Proc. A **472**(2195), 20160442, 23 (2016)
23. David Elworthy, K., Le Jan, Y., Li, X.-M.: The Geometry of Filtering. Frontiers in Mathematics. Birkhäuser, Basel (2010)
24. Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for linear kinetic equations conserving mass. Trans. Am. Math. Soc. **367**(6), 3807–3828 (2015)

25. Dolgopyat, D., Kaloshin, V., Koralov, L.: Sample path properties of the stochastic flows. Ann. Probab. **32**(1A), 1–27 (2004)
26. Dowell, R.M.: Differentiable approximations to Brownian motion on manifolds. PhD thesis, University of Warwick (1980)
27. Duong, M.H., Lamacz, A., Peletier, M.A., Sharma, U.: Variational approach to coarse-graining of generalized gradient flows. Calc. Var. **56**, 100 (2017). Published first online
28. Eckmann, J.-P., Hairer, M.: Uniqueness of the invariant measure for a stochastic PDE driven by degenerate noise. Commun. Math. Phys. **219**(3), 523–565 (2001)
29. Elworthy, K.D., Le Jan, Y., Li, X.-M.: On the Geometry of Diffusion Operators and Stochastic Flows. Lecture Notes in Mathematics, vol. 1720. Springer, New York (1999)
30. Elworthy, K.D.: Stochastic Differential Equations on Manifolds. London Mathematical Society Lecture Note Series, vol. 70. Cambridge University Press, Cambridge (1982)
31. Émery, M.: Stochastic Calculus in Manifolds. Universitext. Springer, Berlin (1989) With an appendix by P.-A. Meyer
32. Freidlin, M.I.: The averaging principle and theorems on large deviations. Uspekhi Mat. Nauk **33**(5(203)), 107–160, 238 (1978)
33. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 260. Springer, New York (1984). Translated from the Russian by Joseph Szücs
34. Freidlin, M.I., Wentzell, A.D.: Random perturbations of dynamical systems. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 260, 3rd edn. Springer, Heidelberg (2012). Translated from the 1979 Russian original by Joseph Szücs
35. Friz, P., Gassiat, P., Lyons, T.: Physical Brownian motion in a magnetic field as a rough path. Trans. Am. Math. Soc. **367**(11), 7939–7955 (2015)
36. Fu, H., Duan, J.: An averaging principle for two-scale stochastic partial differential equations. Stoch. Dyn. **11**(2–3), 353–367 (2011)
37. Fu, H., Liu, J.: Strong convergence in stochastic averaging principle for two time-scales stochastic partial differential equations. J. Math. Anal. Appl. **384**(1), 70–86 (2011)
38. Gonzales-Gargate, I.I., Ruffino, P.R.: An averaging principle for diffusions in foliated spaces. Ann. Probab. **44**(1), 567–588 (2016)
39. Greene, R.E., Wu, H.: C^∞ approximations of convex, subharmonic, and plurisubharmonic functions. Ann. Sci. École Norm. Sup. (4) **12**(1), 47–84 (1979)
40. Gu, Y., Mourrat, J.-C.: Pointwise two-scale expansion for parabolic equations with random coefficients. Probab. Theory Relat. Fields **166**(1–2), 585–618 (2016)
41. Hairer, M., Pavliotis, G.A.: Periodic homogenization for hypoelliptic diffusions. J. Statist. Phys. **117**(1–2), 261–279 (2004)
42. Hairer, M., Mattingly, J.C.: Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. Ann. Math. (2) **164**(3), 993–1032 (2006)
43. Hairer, M., Pardoux, E.: Homogenization of periodic linear degenerate PDEs. J. Funct. Anal. **255**(9), 2462–2487 (2008)
44. Hairer, M., Pillai, N.S.: Ergodicity of hypoelliptic SDEs driven by fractional Brownian motion. Ann. Inst. Henri Poincaré Probab. Stat. **47**(2), 601–628 (2011)
45. Hairer, M., Mattingly, J.C., Scheutzow, M.: Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. Probab. Theory Related Fields **149**(1–2), 223–259 (2011)
46. Has'minskii, R.Z.: On the principle of averaging the Itô's stochastic differential equations. Kybernetika (Prague) **4**, 260–279 (1968)
47. Helland, I.S.: Central limit theorems for martingales with discrete or continuous time. Scand. J. Statist. **9**(2), 79–94 (1982)
48. Hörmander, L.: Hypoelliptic second order differential equations. Acta Math. **119**, 147–171 (1967)
49. Högele, M., Ruffino, P.: Averaging along foliated Lévy diffusions. Nonlinear Anal. **112**, 1–14 (2015)

50. Ikeda, N., Ogura, Y.: A degenerating sequence of Riemannian metrics on a manifold and their Brownian motions. In: Diffusion Processes and Related Problems in Analysis, Vol. I. Progress in probability, vol. 22, pp. 293–312. Birkhäuser, Boston (1990)
51. Kelly, D., Melbourne, I.: Deterministic homogenization for fast-slow systems with chaotic noise. J. Funct. Anal. **272**(10), 4063–4102 (2017)
52. Kifer, Y.: Random Perturbations of Dynamical Systems. Progress in Probability and Statistics, vol. 16. Birkhäuser, Boston (1988)
53. Kipnis, C., Varadhan, S.R.S.: Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Commun. Math. Phys. **104**(1), 1–19 (1986)
54. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry. Vol I. Interscience Publishers, a division of John Wiley & Sons, New York/London (1963)
55. Komorowski, T., Landim, C., Olla, S.: Fluctuations in Markov processes. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 345. Springer, Heidelberg (2012). Time symmetry and martingale approximation
56. Korepanov, A., Kosloff, Z., Melbourne, I.: Martingale-coboundary decomposition for families of dynamical systems. Annales I'Institut Henri Poincaré, Analyse non-linéaire, pp. 859–885 (2018)
57. Kramers, H.A.: Brownian motion in a field of force and the diffusion model of chemical reactions. Physica **7**, 284–304 (1940)
58. Kuehn, C.: Multiple Time Scale Dynamics. Volume 191 of Applied Mathematical Sciences. Springer, Cham (2015)
59. Kusuoka, S., Stroock, D.: Applications of the Malliavin calculus. II. J. Fac. Sci. Univ. Tokyo Sect. IA Math. **32**(1), 1–76 (1985)
60. Kurtz, T.G.: A general theorem on the convergence of operator semigroups. Trans. Am. Math. Soc. **148**, 23–32 (1970)
61. Langevin, P.: Sur la thèorie du mouvement brownien. C. R. Acad. Sci. (Paris), 146 (1908)
62. Li, X.-M.: An averaging principle for a completely integrable stochastic Hamiltonian system. Nonlinearity **21**(4), 803–822 (2008)
63. Li, X.-M.: Effective diffusions with intertwined structures. arxiv:1204.3250 (2012)
64. Li, X.-M.: Random perturbation to the geodesic equation. Ann. Probab. **44**(1), 544–566 (2015)
65. Li, X.-M.: Homogenisation on homogeneous spaces. J. Math. Soc. Jpn. **70**(2), 519–572 (2018)
66. Li, X.-M.: Limits of random differential equations on manifolds. Probab. Theory Relat. Fields **166**(3–4), 659–712 (2016). https://doi.org/10.1007/s00440-015-0669-x
67. Liverani, C., Olla, S.: Toward the Fourier law for a weakly interacting an harmonic crystal. J. Am. Math. Soc. **25**(2), 555–583 (2012)
68. Mazzeo, R.R., Melrose, R.B.: The adiabatic limit, Hodge cohomology and Leray's spectral sequence for a fibration. J. Differ. Geom. **31**(1), 185–213 (1990)
69. Mischler, S., Mouhot, C.: Exponential stability of slowly decaying solutions to the kinetic-Fokker-Planck equation. Arch. Ration. Mech. Anal. **221**(2), 677–723 (2016)
70. Myers, S.B., Steenrod, N.E.: The group of isometries of a Riemannian manifold. Ann. Math. (2) **40**(2), 400–416 (1939)
71. Nelson, E.: Dynamical Theories of Brownian Motion. Princeton University Press, Princeton (1967)
72. Ogura, Y., Taniguchi, S.: A probabilistic scheme for collapse of metrics. J. Math. Kyoto Univ. **36**(1), 73–92 (1996)
73. Papanicolaou, G.C., Kohler, W.: Asymptotic theory of mixing stochastic ordinary differential equations. Commun. Pure Appl. Math. **27**, 641–668 (1974)
74. Papanicolaou, G.C., Stroock, D., Varadhan, S.R.S.: Martingale approach to some limit theorems. In: Papers from the Duke Turbulence Conference (Duke University,1976), ii+120pp. Duke University, Durham (1977)
75. Papanicolaou, G.C., Varadhan, S.R.S.: A limit theorem with strong mixing in Banach space and two applications to stochastic differential equations. Commun. Pure Appl. Math. **26**, 497–524 (1973)

76. Pavliotis, G.A., Stuart, A.M.: Multiscale Methods: Averaging and Homogenization. Texts in Applied Mathematics, vol. 53. Springer, New York (2008)
77. Ruffino, P.R.: Application of an averaging principle on foliated diffusions: topology of the leaves. Electron. Commun. Probab. **20**(28), 5 (2015)
78. Schoen, R., Yau, S.-T.: Lectures on differential geometry. In: Conference Proceedings and Lecture Notes in Geometry and Topology, I. International Press, Cambridge (1994). Lecture notes prepared by Wei Yue Ding, Kung Ching Chang [Gong Qing Zhang], Jia Qing Zhong and Yi Chao Xu, Translated from the Chinese by Ding and S. Y. Cheng, Preface translated from the Chinese by Kaising Tso
79. Skorokhod, A.V., Hoppensteadt, F.C., Salehi, H.: Random Perturbation Methods with Applications in Science and Engineering. Applied Mathematical Sciences, vol. 150. Springer, New York (2002)
80. Stratonovich, R.L.: Selected problems in the theory of fluctuations in radio engineering. Sov. Radio, Moscow (1961). In Russian
81. Stratonovich, R.L.: Topics in the Theory of Random Noise. Vol. I: General Theory of Random Processes. Nonlinear transformations of signals and noise. Revised English edition. Translated from the Russian by Richard A. Silverman. Gordon and Breach Science Publishers, New York/London (1963)
82. Stroock, D., Varadhan, S.R.S.: Theory of diffusion processes. In: Stochastic Differential Equations. C.I.M.E. Summer Schools, vol. 77, pp. 149–191. Springer, Heidelberg (2010)
83. Tam, L.-F.: Exhaustion functions on complete manifolds. In: Lee, Y.-I., Lin, C.-S., Tsui, M.-P. (eds.) Recent Advances in Geometric Analysis. Advanced Lectures in Mathematics (ALM), vol. 11, pp. 211–215. Internat Press, Somerville (2010)
84. Tanno, S.: The first eigenvalue of the Laplacian on spheres. Tôhoku Math. J. (2) **31**(2), 179–185 (1979)
85. Uhlenbeck, G.E., Ornstein, L.S.: Brownian motion in a field of force and the diffusion model of chemical reactions. Phys. Rev. **36**, 823–841 (1930)
86. Urakawa, H.: The first eigenvalue of the Laplacian for a positively curved homogeneous Riemannian manifold. Compositio Math. **59**(1), 57–71 (1986)
87. van Erp, E.: The Atiyah-Singer index formula for subelliptic operators on contact manifolds. Part I. Ann. Math. (2) **171**(3), 1647–1681 (2010)
88. van Erp, E.: The index of hypoelliptic operators on foliated manifolds. J. Noncommut. Geom. **5**(1), 107–124 (2011)
89. Veretennikov, A.Yu.: On an averaging principle for systems of stochastic differential equations. Mat. Sb. **181**(2), 256–268 (1990)
90. Villani, C.: Hypocoercive diffusion operators. In: International Congress of Mathematicians, vol. III, pp. 473–498. European Mathematical Society, Zürich (2006)
91. Weinan, E.: Principles of Multiscale Modeling. Cambridge University Press, Cambridge (2011)
92. Yosida, K.: Functional Analysis. Die Grundlehren der Mathematischen Wissenschaften, Band 123. Academic Press/Springer, New York/Berlin (1965)

Free Probability, Random Matrices, and Representations of Non-commutative Rational Functions

Tobias Mai and Roland Speicher

Abstract A fundamental problem in free probability theory is to understand distributions of "non-commutative functions" in freely independent variables. Due to the asymptotic freeness phenomenon, which occurs for many types of independent random matrices, such distributions can describe the asymptotic eigenvalue distribution of corresponding random matrix models when their dimension tends to infinity. For non-commutative polynomials and rational functions, an algorithmic solution to this problem is presented. It relies on suitable representations for these functions.

1 Introduction

We want to understand distributions of functions in non-commuting variables. This phrase needs some explanations.

Firstly, let us specify what our "non-commuting variables" will usually be. We are mostly interested in either (random) matrices of size $N \times N$ or in operators on Hilbert spaces; one of our main points later will be that such operators correspond usually to the limit $N \to \infty$ of our random matrices.

Then, which "functions" of those variables do we want to consider? Since our variables do in general not commute, taking functions in such non-commuting variables is not a straightforward thing. In fact, we see that this question actually splits into two: we first need to clarify what our non-commutative functions should be as objects in their own right and secondly, we must explain how these non-commutative functions can be evaluated in the given collection of non-commuting variables. Everything which goes beyond polynomials is a non-trivial issue.

T. Mai · R. Speicher (✉)
Fachrichtung Mathematik, Universität des Saarlandes, Saarbrücken, Germany
e-mail: mai@math.uni-sb.de; speicher@math.uni-sb.de

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_19

Here, we will mostly address non-commutative polynomials and non-commutative rational functions, but our hope is that in the long run we will also have access to non-commutative analytic functions. The basis for a non-commutative analogue of complex function theory, intended to provide some sort of multivariate functional calculus in analogy to the well-known analytic functional calculus for a single operator, was laid in the 1970s in work of Joseph L. Taylor [33, 34]; but only recently this was revived and is under heavy development (with motivations coming from different directions, in particular free probability theory, but also control theory). We refer the reader who is interested in this subject to [25]. In this article we will not go beyond non-commutative rational functions.

Finally, we should be precise what we mean by "distribution" of our functions in our variables. There are essentially two versions of this. In the most general setting, we have to talk about algebraic/combinatorial distributions, which is just given by the collection of moments of our considered random variables. In more restricted analytic settings this might be identified with an analytic distribution, which is just a probability measure. To make this more precise we first have to set our frame.

Definition 1 A *non-commutative probability space* $(\mathscr{A}, \varphi)$ consists of a complex algebra $\mathscr{A}$ with unit $1_{\mathscr{A}}$ and a linear functional $\varphi : \mathscr{A} \to \mathbb{C}$ satisfying $\varphi(1_{\mathscr{A}}) = 1$. Elements $x \in \mathscr{A}$ are called *non-commutative random variables* and φ is usually addressed as *expectation*.

Example 1 Let us give some examples for this.

1. The classical setting is captured in this algebraic form via $(L^{\infty}(\Omega, P), E)$, where (Ω, Σ, P) is a classical probability space and E the usual expectation that is given by $E[X] = \int_{\Omega} X(\omega)\, dP(\omega)$.
2. A typical genuine non-commutative example is $(M_N(\mathbb{C}), \mathrm{tr}_N)$, where tr_N is the normalized trace on $M_N(\mathbb{C})$; i.e., $\mathrm{tr}((a_{ij})_{i,j=1}^{N}) = \frac{1}{N}\sum_{k=1}^{N} a_{kk}$.
3. The combination of the two examples leads to one of our most important examples, namely $(L^{\infty}(\Omega) \otimes M_N(\mathbb{C}), E \otimes \mathrm{tr}_N)$, whose elements are random matrices having entries that are bounded random variables. In order to include also random variables that are not necessarily bounded but have moments of all orders, we must replace $L^{\infty}(\Omega)$ by $L^{\infty-}(\Omega) := \bigcap_{p \geq 1} L^p(\Omega)$.

Definition 2 We call $(\mathscr{A}, \varphi)$ a *C^*-probability space* if $\mathscr{A}$ is a unital C^*-algebra and φ is a state (i.e. $\varphi(x^*x) \geq 0$ for all $x \in \mathscr{A}$). The former means that $\mathscr{A}$ consists of bounded operators on some Hilbert space $\mathscr{H}$ and a state on $\mathscr{A}$ can, via the GNS construction, be realized in the form $\varphi(x) = \langle \Omega, x\Omega \rangle$ for some unit vector $\Omega \in \mathscr{H}$.

Now we can be more precise on what we mean with "distributions".

Definition 3 Let $(\mathscr{A}, \varphi)$ be a non-commutative probability space. Let $(x_i)_{i \in I}$ be a family of non-commutative random variables. We call the unital linear mapping

$$\mu_{(x_i)_{i \in I}} : \mathbb{C}\langle \chi_i \mid i \in I \rangle \to \mathbb{C}, \quad \chi_{i_1} \cdots \chi_{i_k} \mapsto \varphi(x_{i_1} \cdots x_{i_k})$$

that is defined on the free associative algebra $\mathbb{C}\langle \chi_i \mid i \in I \rangle$ generated by the non-commuting variables $(\chi_i)_{i\in I}$ the *(joint) distribution* of $(x_i)_{i\in I}$.

In the general algebraic frame, we can only talk about distributions in such a combinatorial fashion, whereas in the analytic setting of a C^*-probability space $(\mathcal{A}, \varphi)$, we can, by the Riesz representation theorem for positive linear functionals on continuous functions (see, for example, Prop. 3.13 in [31]), identify the distribution of a single selfadjoint operator $x \in \mathcal{A}$ with the unique *Borel probability measure* μ_x on the real line $\mathbb{R}$ that satisfies

$$\varphi(x^k) = \int_{\mathbb{R}} t^k \, d\mu_x(t) \qquad \text{for all } k \in \mathbb{N}_0. \tag{1}$$

Using the same symbol both for "combinatorial" and "analytic" distributions is of course an abuse of notation. This can be excused, since both of them contain the same information, and usually, it is clear which of the two we are using; if we want to be precise, we refer to the latter as the *analytic distribution* of x.

Similarly, in the classical multivariate case (for several commuting selfadjoint variables $x_1, \ldots, x_n$ in a C^*-setting), we can identify the combinatorial distribution $\mu_{x_1,\ldots,x_n}$ with a probability measure on $\mathbb{R}^n$. In the general case, where our variables $x_1, \ldots, x_n$ do not commute, this is not possible any more. It is tempting to think of the distribution $\mu_{x_1,\ldots,x_n}$ in such a situation as a "non-commutative probability measure", but actually we have no idea what this should mean. As a kind of analytic substitute, we will try to analyze the distribution of $(x_1, \ldots, x_n)$ by investigating the analytic distributions of all $p(x_1, \ldots, x_n)$ for a large class of selfadjoint functions of $x_1, \ldots, x_n$. Clearly, the more functions we can deal with, the better we understand $\mu_{x_1,\ldots,x_n}$. Looking on polynomials and rational functions is a first step in this direction.

2 Random Matrices

Random matrices are $N \times N$ matrices, whose entries are chosen randomly (according to a prescribed distribution). Usually, one looks on sequences of such matrices for growing N. One of the basic observations in the subject is that for $N \to \infty$ something interesting happens. Before becoming more concrete on this, let us give a bit of history of the subject.

Fig. 1 The Oberwolfach workshop "Random Matrices" in 2000 was one of the first general appearances of the subject in mathematics. (Source: Archives of the Mathematisches Forschungsinstitut Oberwolfach.)

2.1 Some History

1928	Wishart introduced random matrices in statistics, for finite N;
1955	Wigner introduced random matrices in physics, for a statistical description of nuclei of heavy atoms, and investigated the $N \to \infty$ asymptotics of these "Wigner matrices";
1967	Marchenko and Pastur described the $N \to \infty$ asymptotics of "Wishart matrices";
1972	Montgomery and Dyson discovered relation between zeros of the Riemann zeta function and eigenvalues of random matrices;
since 2000	random matrix theory developed into a central subject in mathematics, with many different connections (Fig. 1).

2.2 Wigner's Semi-circle Law

As said before, random matrices are sequences of $N \times N$ matrices whose entries are chosen randomly (according to a prescribed distribution). A fundamental observation in the subject is that many random matrices show for $N \to \infty$ almost surely a *deterministic* (and interesting) behaviour. Let us give an example for this via one of the most important random matrix ensembles, the Wigner matrices, which were introduced by Eugene Wigner in 1955 [38].

Definition 4 A *Wigner random matrix* $X_N = \frac{1}{\sqrt{N}}(x_{ij})_{i,j=1}^{N}$ is a real random matrix, which is symmetric ($X_N^* = X_N$, i.e. $x_{ij} = x_{ji}$ for all $i, j = 1, \ldots, N$)

and apart from this symmetry condition all its entries $\{x_{ij} \mid 1 \leq i \leq j \leq N\}$ are chosen independent and identically distributed.

Surprisingly, the common distribution of the entries does not matter for many results. The nicest distribution is the Gaussian (which leads to what is called "Gaussian orthogonal ensemble" GOE). We will here instead produce our random matrices by independent coin tosses for the entries; i.e., our common distribution for the entries is the symmetric Bernoulli distribution $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. Here is one realization (via independent coin tosses by the authors) for such a 10×10 Wigner matrix:

$$\frac{1}{\sqrt{10}} \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 \end{pmatrix}$$

The main quantity one is usually interested in for (random) matrices are the eigenvalues. For a matrix $A \in M_N(\mathbb{C})$, the information about its eigenvalues $\lambda_1, \ldots, \lambda_N$ (counted with multiplicity) is encoded in the *empirical eigenvalue distribution*

$$\mu_A = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}.$$

Note that this probability measure is nothing but the analytical distribution of A with respect to the normalized trace tr_N in case A is selfadjoint.

The left picture in Fig. 2 shows the histogram of the 10 eigenvalues for the above matrix. Of course, since the matrix is random, the eigenvalue distribution is also random, so depends on the chosen realization. The right picture in Fig. 2 is the histogram of the 10 eigenvalues of another such matrix created by coin tosses.

Clearly, the two pictures do not have much similarity. But now let's do the same for two different realizations of a 100×100 matrix, see Fig. 3, and for two different realizations of a 3000×3000 matrix, see Fig. 4. (Instead of tossing coins we preferred in those cases to use matlab for producing the matrices.)

Those histograms should make clear what we mean with the statement that for $N \to \infty$ the eigenvalue distribution of a Wigner matrix converges almost surely to a deterministic limit μ (which is called "semicircle distribution"); more precisely, we have that μ_{X_N} converges in the weak topology for probability measures to μ

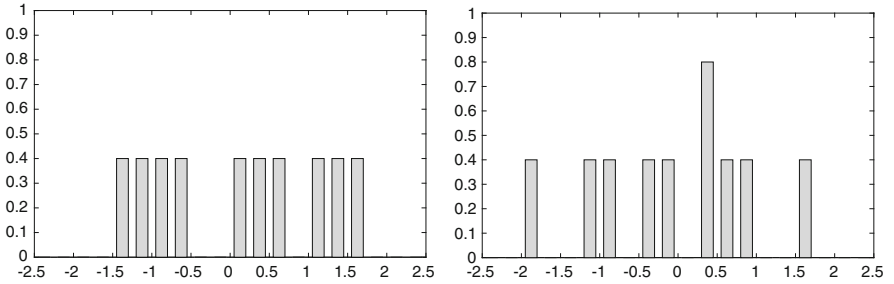


Fig. 2 The histogram of the 10 eigenvalues of a 10×10-Wigner matrix; for two different realizations of the matrix

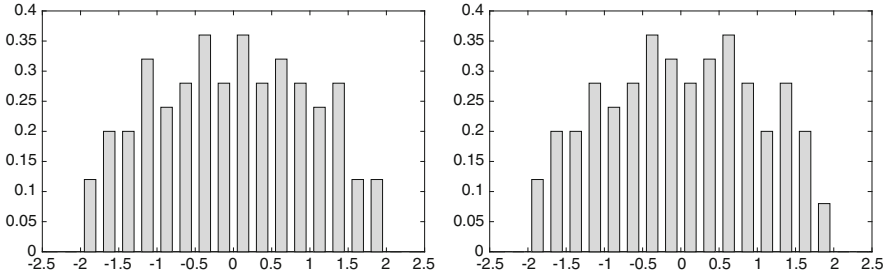


Fig. 3 The histogram of the 100 eigenvalues of a 100×100-Wigner matrix; for two different realizations of the matrix

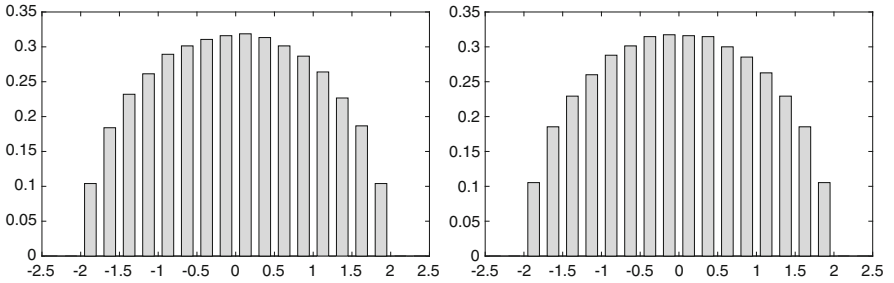


Fig. 4 The histogram of the 3000 eigenvalues of a 3000×3000-Wigner matrix; for two different realizations of the matrix

(and this happens for almost all realizations of X_N). This almost sure convergence is a concrete instance of concentration phenomena in high dimensions and is usually not too hard to prove. What is more interesting is the determination and description of this deterministic limit μ. Let us address the question how we can describe the limit.

2.3 Convergence in Distribution to the Large N Limit

In the above treated one-matrix case X_N, the usual classical way of describing the almost sure limit of μ_{X_N} is by a probability measure μ. Here is an alternative to this, which we will favor in the following: instead of just describing μ, we try to find some nice operator x on a Hilbert space $\mathscr{H}$ with state φ such that the distribution of x with respect to φ coincides with μ; i.e. that $\mu = \mu_x$; then we can say that X_N converges to x in distribution. Note that this is like in the classical central limit theorem where often one prefers to talk about the convergence of normalized sums to a normal variable instead of just saying that the distribution of the normalized sums converge to a normal distribution.

Of course, this is just language. However, in the multi-variate non-commutative case this shift in perspective is more fundamental. So let us consider two independent copies X_N, Y_N of our Wigner matrices. As those do not commute, there is no nice analytic object describing their joint distribution (which is given by all mixed moments with respect to tr_N) and hence the determination of the almost sure limit of μ_{X_N,Y_N} would consist in trying to find some (combinatorial) description of the limits of the moments. Again, we propose an alternative: try to find some nice operators x, y on a Hilbert space with some state φ, such that almost surely

$$\lim_{N\to\infty} \mathrm{tr}_N(q(X_N, Y_N)) = \varphi(q(x, y))$$

for all monomials, and hence for all polynomials, q. Then we can say again that the pair (X_N, Y_N) converges in distribution to the pair (x, y).

The question is of course how big are our chances to have such nice limiting operators. Note that the important point here is "nice"; by the GNS-construction we can always find some abstract operators somewhere out there with the correct limiting moments. What we really want are operators, which can be handled and are useful.

The surprising fact in this context is the fundamental observation of Voiculescu [35] from 1991 that indeed limits of random matrices can often be described by "nice" and "interesting" operators on Hilbert spaces. (Actually, those operators describe usually interesting von Neumann algebras, which was the initial starting point of Voiculescu.)

2.4 Semi-circle Law and One-Sided Shift

Figure 5 shows again the histogram of our large Wigner matrices compared to the semicircle density.

We claim that the real part of one of the most important Hilbert space operators – if suitably rescaled – has actually this semicircle distribution. More precisely, in this case our limit operator x can be written in the form $x = l + l^*$, where l is the

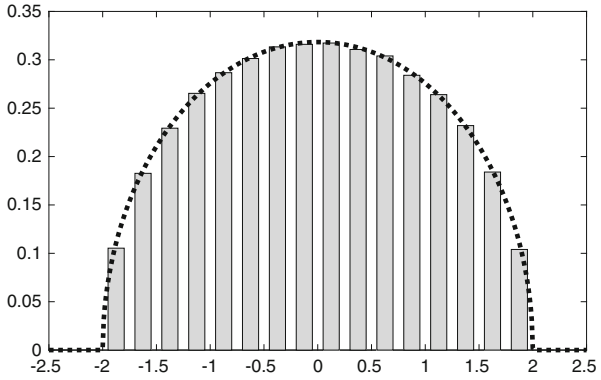


Fig. 5 Wigner's Theorem [38] says that the empirical eigenvalue distribution of Wigner matrices converges to the semicircle distribution; the Theorem of Füredi and Komlós [15] says that we also have almost sure convergence of the operator norms; i.e., there are no outlier eigenvalues outside the limit spectrum

one-sided shift on the Hilbert space $\bigoplus_{n\geq 0} \mathbb{C}e_n$ with orthonormal basis $(e_n)_{n\in\mathbb{N}_0}$; the action of the shift is given by the action on the basis: $le_n = e_{n+1}$ for all $n \in \mathbb{N}_0$; this implies that the action of the adjoint operator l^* is given by: $l^*e_{n+1} = e_n$ for all $n \in \mathbb{N}_0$ and $l^*e_0 = 0$. A canonical state on the algebra generated by those operators is the vector state $\varphi(a) = \langle e_0, ae_0\rangle$, corresponding to the distinguished basis element e_0. It turns out (and is actually a nice exercise) that the moments of x with respect to φ are given by the moments of the semicircle distribution; namely both are equal to the famous *Catalan numbers*. More concretely, odd moments are zero in both cases and for even moments we have

$$\varphi(x^{2n}) = \frac{1}{n+1}\binom{2n}{n} = \frac{1}{2\pi}\int_{-2}^{+2} t^{2n}\sqrt{4-t^2}\,dt.$$

In our language, we can now express the theorem of Wigner from 1955 [38] by saying that $X_N \to x$. Wigner did not equate the limiting moments of the Wigner matrices to the moments of our operator x, but just calculated them as the Catalan numbers.

We want to point out that the eigenvalue distribution μ_{X_N} gives only the averaged behaviour over all eigenvalues and its limiting behaviour does not allow to infer what happens to the largest eigenvalues of our Wigner matrices. Wigner's semicircle law would still allow that there is one exceptional large eigenvalue which has nothing to do with the limiting spectrum $[-2, +2]$. The mass $1/N$ of such an eigenvalue would disappear in the limit. However, there have been strengthenings of Wigner's result, which also tell us that such outliers are almost surely non-existent. More precisely, Füredi and Komlós showed in 1981 [15] that almost surely the largest eigenvalue of X_N converges to the edge of the spectrum, namely 2. Since

the operator norm of the limit operator is 2, $\|x\| = 2$, we can paraphrase the result of Füredi and Komlós in our language as $\|X_N\| \to \|x\|$ almost surely.

2.5 *Several Independent Wigner Matrices and Full Fock Space*

Let us now consider the multi-variate situation. Voiculescu showed in [35] that the limit of two independent Wigner matrices X_N, Y_N can be described by a canonical multi-dimensional version of the one-sided shift; namely, by two copies of the one-sided shift in different directions. More precisely, we consider now the full Fock space $\mathscr{F}(\mathscr{H})$ over an underlying Hilbert space $\mathscr{H}$, given by

$$\mathscr{F}(\mathscr{H}) := \bigoplus_{n=0}^{\infty} \mathscr{H}^{\otimes n},$$

where $\mathscr{H}^{\otimes 0}$ is a one-dimensional Hilbert space which we write in the form $\mathscr{H}^{\otimes 0} = \mathbb{C}\Omega$ for some distinguished unit vector of norm one; Ω is usually called the *vacuum vector*. On this full Fock space one has, for each $f \in \mathscr{H}$, a *creation operator* $l(f)$ given by

$$l(f)\Omega = f, \qquad l(f) f_1 \otimes \cdots \otimes f_n = f \otimes f_1 \otimes \cdots \otimes f_n.$$

The adjoint of $l(f)$ is the *annihilation operator* $l^*(f)$, i.e. $l(f)^* = l^*(f)$, which is given by

$$l^*(f)\Omega = 0, \qquad l^*(f) f_1 \otimes \cdots \otimes f_n = \langle f, f_1 \rangle f_2 \otimes \cdots \otimes f_n,$$

where, in particular, $l^*(f) f_1 = \langle f, f_1 \rangle \Omega$. Let g_1 and g_2 be two orthogonal unit vectors in $\mathscr{H}$; then we put $x := l(g_1) + l^*(g_1)$ and $y := l(g_2) + l^*(g_2)$. Again, we have a canonical state given by the vacuum vector Ω, $\varphi(a) = \langle \Omega, a\Omega \rangle$. It turns now out (as a special case of Voiculescu's result [35] on asymptotic freeness) that we have $(X_N, Y_N) \to (x, y)$. Note that both x and y have with respect to φ a semicircular distribution; the basis vectors e_n from the one-sided shift correspond in the present setting to $g_1^{\otimes n}$ (for x) or to $g_2^{\otimes n}$ (for y).

Let us point out that in the same way as the one sided-shift l is one of the most important operators in single operator theory, the creation operators $l(g_1)$ and $l(g_2)$ are actually important (and nice) operators in the theory of operator algebras; they are closely related to the *Cuntz algebra*, which is one of the most important C^*-algebras and their real parts generate as von Neumann algebra the *free group factor* $L(\mathbb{F}_2)$, which is a main object of interest in Voiculescu's free probability theory.

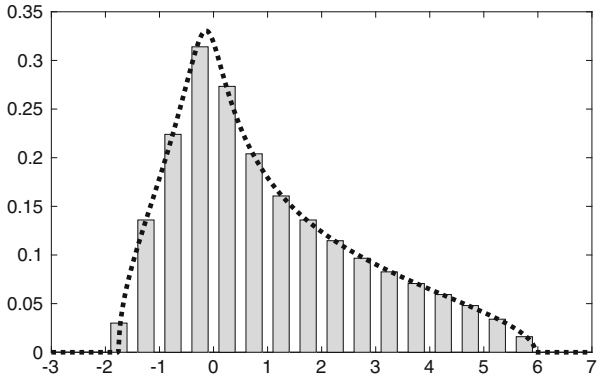


Fig. 6 Voiculescu's multivariate version of Wigner's Theorem says that the empirical eigenvalue distribution of $p(X_N, Y_N)$ for two independent Wigner matrices converges to the distribution of $p(x, y)$; the Theorem of Haagerup and Thorbjørnsen [18] says that we also have almost sure convergence of the operator norms; i.e., there are no outlier eigenvalues outside the limit spectrum. Here we have $p(x, y) = xy + yx + x^2$

As indicated at the end of Sect. 1, in order to get a better understanding of the limit distribution $\mu_{x,y}$ we will now try to deal with selfadjoint polynomials $p(x, y)$ in x and y. The convergence in distribution of (X_N, Y_N) to (x, y) implies that also $p(X_N, Y_N)$ converges to $p(x, y)$ for all such polynomials. For example, consider the polynomial $p(x, y) = xy + yx + x^2$. Then the theorem of Voiculescu tells us that $p(X_N, Y_N) \to p(x, y)$. This is again something which can be visualized by comparing the histogram of eigenvalues of $p(X_N, Y_N) = X_N Y_N + Y_N X_N + X_N^2$ with the analytic distribution of the selfadjoint operator $p(x, y) = xy + yx + x^2$; see Fig. 6. At the moment it should not be clear to the reader how to get the distribution of $p(x, y)$ explicitly; understanding how we can get the dotted curve in Fig. 6 will be one of the main points of the rest of this article.

Again, the behavior of the largest eigenvalue of $p(X_N, Y_N)$ is not captured by Voiculescu's result on the convergence in distribution of (X_N, Y_N) to (x, y). As in the classical case, there is some strengthening, which addresses this question. Namely, Haagerup and Thorbjørnsen have shown in [18] that we have almost sure convergence of the largest eigenvalue $\|p(X_N, Y_N)\|$ to the corresponding limit quantity $\|p(x, y)\|$. Whereas in the one-dimensional case we are only dealing with one limiting probability measure, for which the edge of the spectrum is clear, in the present, multivariate case we want now a statement covering a whole family of probability measures μ_p varying with the considered polynomial p; since we have no concrete description of those measures, there is also no explicit description of the edge of the support of those measures in useful classical terms – in the non-commutative setting, however, this can be easily described as the operator norm of $p(x, y)$.

2.6 Are Those Limit Operators x, y Really Useful?

Still one might have the feeling that talking about operators as the limit of the X_N, Y_N instead of limits of distributions of $p(X_N, Y_N)$ might be more an issue of language than real insights. So the question remains: What are those limit operators good for? Here are some supporting facts in favour of their relevance.

Theorem 1 ([35]) *For many random matrix models X_N, Y_N (like for independent Wigner matrices) the limit operators x, y are free in the sense of Voiculescu's free probability theory.*

The notion of "freeness" is defined as follows: let $(\mathscr{A}, \varphi)$ be any non-commutative probability space; a family $(\mathscr{A}_i)_{i\in I}$ of unital subalgebras of a $\mathscr{A}$ is called *free*, if $\varphi(a_1 \cdots a_k) = 0$ holds whenever we have $a_j \in \mathscr{A}_{i_j}$ and $\varphi(a_j) = 0$ for $j = 1, \ldots, k$ with some $k \in \mathbb{N}$ and indices $i_1, \ldots, i_k \in I$ satisfying $i_1 \neq i_2 \neq \cdots \neq i_k$; accordingly, a family $(x_i)_{i\in I}$ of non-commutative random variables in $\mathscr{A}$ is called *free*, if $(\mathscr{A}_i)_{i\in I}$ is free in the previous sense, where $\mathscr{A}_i$ denotes the subalgebra of $\mathscr{A}$ generated by $1_{\mathscr{A}}$ and x_i. For more details the reader should consult some of the references [21, 30, 31, 36] for the subject. Here, we want to emphasize that free probability theory has developed a couple of tools to work effectively with free random variables. In particular, for x and y free we have

- *free convolution*: the distribution of $x + y$ can effectively be calculated in terms of the distribution of x and the distribution of y;
- *matrix-valued free convolution*: the matrix-valued distribution of $\alpha_0 \otimes 1 + \alpha_1 \otimes x + \alpha_2 \otimes y$ (where the coefficients $\alpha_0, \alpha_1, \alpha_2$ are now not just complex numbers, but matrices of arbitrary size) can be calculated in terms of the distribution of x and the distribution of y.

Still, this does not sound like a convincing argument in favour of x, y. What we want is to be able to deal with arbitrary polynomials in x and y. The above tells us that we can deal with linear polynomials in x and y, which seems to be much less. However, the fact the we have included the matrix-valued version above has striking consequences if we combine this with some powerful techniques of purely algebraic nature, which we summarize here for the sake of simplicity under the name "linearization". The latter is of such a fundamental relevance that we will treat it in a section of its own.

3 Linearization and the Calculation of the Distribution of $p(x, y)$

3.1 Idea of Linearization

The linearization philosophy says that we can transform a *non-linear problem* into a *matrix-valued linear problem*. More precisely, if we want to understand a non-linear polynomial $p(x_1, \ldots, x_m)$ in non-commuting variables $x_1, \ldots, x_m$, then we can assign to it (in a non-unique way) a linear polynomial $\hat{p} := \alpha_0 \otimes 1 + \alpha_1 \otimes x_1 + \cdots + \alpha_m \otimes x_m$ (where we have to allow matrix-valued coefficients), such that $\hat{p}$ contains all "relevant information" about $p(x_1, \ldots, x_m)$; note that $\hat{p}$ is by definition an element of $M_n(\mathbb{C}) \otimes \mathbb{C}\langle x_1, \ldots, x_m \rangle$ for some n, but we usually identify that space with $M_n(\mathbb{C}\langle x_1, \ldots, x_m \rangle)$. Relevant information for us is the spectrum of the operators, hence we would like to decide whether $p(x_1, \ldots, x_m)$, and more generally $z - p(x_1, \ldots, x_m)$ for $z \in \mathbb{C}$, is invertible. To see how such questions on invertibility can be shifted from $p(x_1, \ldots, x_m)$ to some $\hat{p}$ let us consider some examples.

Example 2 Consider first the simple polynomial $p(x, y) = xy$. We try to decide for which $z \in \mathbb{C}$ the element $z - xy$ is invertible. For this we write

$$\begin{pmatrix} z - xy & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z & -x \\ -y & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix}. \tag{2}$$

Of course, $z - xy$ is invertible if and only if the matrix on the left-hand side is invertible. On the right-hand side we have a product of three matrices; however, the first and the third are always invertible, as one has for all x and all y

$$\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ -y & 1 \end{pmatrix}.$$

Hence $z - xy$ is invertible if and only if the middle matrix

$$\begin{pmatrix} z & -x \\ -y & 1 \end{pmatrix} = \begin{pmatrix} z & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & x \\ y & -1 \end{pmatrix} = \Lambda(z) - \hat{p}$$

is invertible, where we put

$$\Lambda(z) = \begin{pmatrix} z & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \hat{p} = \begin{pmatrix} 0 & x \\ y & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \otimes 1 + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \otimes x + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \otimes y.$$

But this $\hat{p}$ is now a matrix-valued linear polynomial in x and y. Furthermore, we infer from the identity (2) that the resolvent $(z - xy)^{-1}$ appears as the $(1, 1)$-entry of the 2×2-matrix $(\Lambda(z) - \hat{p})^{-1}$.

Example 3 Let us consider now the more interesting $p(x, y) = xy + yx + x^2$ and ask for which $z \in \mathbb{C}$ the element $z - p(x, y)$ becomes invertible. Again we have a factorization into linear terms on matrix level

$$\begin{pmatrix} z - p(x,y) & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & y + \frac{x}{2} & x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} z & -x & -y - \frac{x}{2} \\ -x & 0 & 1 \\ -y - \frac{x}{2} & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ y + \frac{x}{2} & 1 & 0 \\ x & 0 & 1 \end{pmatrix}. \tag{3}$$

As before the first and third term are triangular matrices which are always invertible; indeed,

$$\begin{pmatrix} 1 & 0 & 0 \\ y + \frac{x}{2} & 1 & 0 \\ x & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -y - \frac{x}{2} & 1 & 0 \\ -x & 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & y + \frac{x}{2} & x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -y - \frac{x}{2} & -x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence $p(x, y) = xy + yx + x^2$ is invertible if and only if the 3×3-matrix valued linear polynomial

$$\begin{pmatrix} z & -x & -y - \frac{x}{2} \\ -x & 0 & 1 \\ -y - \frac{x}{2} & 1 & 0 \end{pmatrix} = \begin{pmatrix} z & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & x & y + \frac{x}{2} \\ x & 0 & -1 \\ y + \frac{x}{2} & -1 & 0 \end{pmatrix} = \Lambda(z) - \hat{p}$$

is invertible, where we put

$$\Lambda(z) = \begin{pmatrix} z & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$\hat{p} = \begin{pmatrix} 0 & x & y + \frac{x}{2} \\ x & 0 & -1 \\ y + \frac{x}{2} & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \otimes 1 + \begin{pmatrix} 0 & 1 & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \otimes x + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \otimes y.$$

One should note that also the value of the inverse can be read off from inverting $\hat{p}$. Namely, we can easily infer from (3) that the resolvent $(z - p(x, y))^{-1}$ is the $(1, 1)$-entry of the 3×3-matrix $(\Lambda(z) - \hat{p})^{-1}$.

All the above can now actually be generalized to any polynomial $p(x, y)$. In view of the previous examples, this requires, of course, to have some general rule to produce matricial factorizations like in (2) and (3). For clarifying these relations, it is helpful to consider a block decomposition of the considered linearization $\hat{p}$ of the form

$$\hat{p} = \begin{pmatrix} 0 & u \\ v & Q \end{pmatrix}, \tag{4}$$

where the zero block in the upper left corner is of size 1×1 and all other blocks are of appropriate size. In each of the previous examples, we may observe

1. that the block Q is invertible without any conditions on x and y and
2. that its inverse Q^{-1} satisfies $p(x, y) = -uQ^{-1}v$.

Furthermore, we see that with these notations, the factorizations (2) and (3) take now the general form

$$\begin{pmatrix} z - p(x, y) & 0 \\ 0 & -Q \end{pmatrix} = \begin{pmatrix} 1 & -uQ^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z & -u \\ -v & -Q \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -Q^{-1}v & 1 \end{pmatrix}. \tag{5}$$

In this abstract frame, we can repeat the computations, which were carried out in the previous examples; this yields

$$\begin{pmatrix} (z - p(x, y))^{-1} & 0 \\ 0 & -Q^{-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ Q^{-1}v & 1 \end{pmatrix} \left(\begin{pmatrix} z & 0 \\ 0 & 0 \end{pmatrix} - \hat{p} \right)^{-1} \begin{pmatrix} 1 & uQ^{-1} \\ 0 & 1 \end{pmatrix} \tag{6}$$

and finally

$$(z - p(x, y))^{-1} = \left[(\Lambda(z) - \hat{p})^{-1} \right]_{1,1} \qquad \text{with} \qquad \Lambda(z) = \begin{pmatrix} z & 0 \\ 0 & 0 \end{pmatrix}. \tag{7}$$

In fact, the validity of the factorization (5) and thus the validity of the formulas in (6) and (7) only depend on the properties formulated in Item 1 and Item 2. This is known under the name *Schur-complement formula* and it allows us to generalize our arguments given above to any non-commutative polynomial $p(x_1, \ldots, x_m)$ in finitely many variables $x_1, \ldots, x_m$. For this, however, we need to be sure that $p(x_1, \ldots, x_m)$ enjoys a representation of the form

$$p(x_1, \ldots, x_m) = -uQ^{-1}v \tag{8}$$

with vectors u, v and an invertible matrix Q of compatible sizes, which are (affine) linear in the variables $x_1, \ldots, x_m$. According to (4), finding such a representation of $p(x_1, \ldots, x_m)$ is all we need in order to produce a linearization $\hat{p}$.

Theorem 2 *Each non-commutative polynomial $p(x_1, \ldots, x_m)$ admits a representation of the form (8). It can be constructed in the following way:*

1. *If $p(x_1, \ldots, x_m)$ is a* monomial *of the form*

$$p(x_1, \ldots, x_m) = \lambda x_{i_1} x_{i_2} \cdots x_{i_k}$$

with $\lambda \in \mathbb{C}$, $k \geq 1$, and $i_1, \ldots, i_k \in \{1, \ldots, m\}$, then

$$p(x_1, \ldots, x_m) = -\begin{pmatrix} 0 & 0 & \ldots & 0 & \lambda \end{pmatrix} \begin{pmatrix} & & & & x_{i_1} & -1 \\ & & & x_{i_2} & -1 & \\ & & \iddots & \iddots & & \\ & x_{i_k} & -1 & & & \\ -1 & & & & & \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

2. *If polynomials $p_1(x_1, \ldots, x_m), \ldots, p_k(x_1, \ldots, x_m)$ have representations*

$$p_j(x_1, \ldots, x_m) = -u_j Q_j^{-1} v_j \qquad \text{for } j = 1, \ldots, k,$$

then their sum

$$p(x_1, \ldots, x_m) := p_1(x_1, \ldots, x_m) + \cdots + p_k(x_1, \ldots, x_m)$$

is represented by

$$p(x_1, \ldots, x_m) = -\begin{pmatrix} u_1 & \ldots & u_k \end{pmatrix} \begin{pmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_k \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}.$$

3. *If p is selfadjoint and*

$$p(x_1, \ldots, x_m) = -u Q^{-1} v$$

any representation, then

$$p(x_1, \ldots, x_m) = -\begin{pmatrix} \frac{1}{2} u & v^* \end{pmatrix} \begin{pmatrix} 0 & Q^* \\ Q & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2} u^* \\ v \end{pmatrix}$$

yields another representation, which induces via (4) a selfadjoint linearization $\hat{p}$ of $p(x_1, \ldots, x_m)$.

The previous theorem constitutes the alternative approach of Anderson [2] to the "linearization trick" of [17, 18]. Whereas the original algorithm in [17, 18] was quite

complicated and did not preserve selfadjointness, Anderson's version streamlines their arguments and respects also selfadjointness. Let us summarize.

Theorem 3 ([2, 17, 18]) *Every polynomial $p(x_1, \ldots, x_m)$ has a (non-unique) linearization*

$$\hat{p} = \alpha_0 \otimes 1 + \alpha_1 \otimes x_1 + \cdots + \alpha_m \otimes x_m,$$

such that

$$(z - p(x_1, \ldots, x_m))^{-1} = \left[(\Lambda(z) - \hat{p})^{-1}\right]_{1,1}, \quad \text{where} \qquad \Lambda(z) = \begin{pmatrix} z & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}.$$

If p is selfadjoint, then $\hat{p}$ can also be chosen selfadjoint (meaning that the matrices $\alpha_0, \alpha_1, \ldots, \alpha_m$ appearing in $\hat{p}$ are all hermitian).

3.2 Calculation of the Distribution of $p(x, y)$

Let us now come back to our problem of calculating the distribution of a selfadjoint polynomial $p(x, y)$ in two free variables x and y. The distribution μ_p of $p = p(x, y)$ is a probability measure and the information about such probability measures is often encoded in certain functions: whereas in classical probability theory the function of our choice is usually the Fourier transform, in free probability and random matrix theory it is more adequate to use the so-called *Cauchy transform*; for any Borel probability measure μ on the real line $\mathbb{R}$, this is the analytic function

$$G_\mu : \mathbb{C}^+ \to \mathbb{C}^-, \quad z \mapsto \int_{\mathbb{R}} \frac{1}{z - t}\, d\mu(t),$$

which is defined on the upper complex half plane $\mathbb{C}^+ = \{z \in \mathbb{C} \mid \Im(z) > 0\}$ and whose values lie all in the lower complex half plane $\mathbb{C}^- = \{z \in \mathbb{C} \mid \Im(z) < 0\}$. Note that the Cauchy transform G_μ differs only by a minus sign from the so-called *Stieltjes transform* $S_\mu : \mathbb{C}^+ \to \mathbb{C}^+$, which is the more familiar object in random matrix theory. It is an important fact that the measure μ can be recovered from G_μ by the so-called *Stieltjes inversion formula*: for each $\epsilon > 0$, we have an absolutely continuous probability measure μ_ϵ given by

$$d\mu_\epsilon(t) = -\frac{1}{\pi} \Im(G_\mu(t + i\epsilon))\, dt,$$

and μ_ϵ converges weakly to μ as $\epsilon \searrow 0$, in the sense that

$$\int_{\mathbb{R}} f(t)\, d\mu(t) = \lim_{\epsilon \searrow 0} \int_{\mathbb{R}} f(t)\, d\mu_\epsilon(t)$$

for all bounded continuous functions $f : \mathbb{R} \to \mathbb{C}$. Thus, knowing the Cauchy transform of a probability measure is equivalent to the knowledge of the measure itself. Note that for the analytic distribution μ_x of some selfadjoint x in a C^*-probability space $(\mathscr{A}, \varphi)$ we have $G_{\mu_x}(z) = \varphi((z-x)^{-1})$; we often write G_x instead of G_{μ_x}.

The method of linearization formulated in Theorem 3 now allows us to connect the wanted Cauchy transform of $p(x, y)$ with its selfadjoint linearization $\hat{p}$ according to

$$G_{p(x,y)}(z) = \varphi\big((z - p(x, y))^{-1}\big) = \big[(1 \otimes \varphi)\big((\Lambda(z) - \hat{p})^{-1}\big)\big]_{1,1}, \tag{9}$$

where $1 \otimes \varphi$ acts entrywise as φ on each entry of the corresponding matrix. This puts the original scalar-valued problem concerning $p(x, y)$ into the setting of operator-valued free probability, where the expression $(1 \otimes \varphi)\big((\Lambda(z) - \hat{p})^{-1}\big)$ can be interpreted as (a boundary value of) the operator-valued Cauchy transform of $\hat{p}$: an *operator-valued non-commutative probability space* $(\mathscr{A}, E, \mathscr{B})$ consists of a complex unital algebra $\mathscr{A}$ with a distinguished subalgebra $1_{\mathscr{A}} \in \mathscr{B} \subseteq \mathscr{A}$ and a linear map $E : \mathscr{A} \to \mathscr{B}$, called *conditional expectation*, which satisfies $E[b] = b$ for all $b \in \mathscr{B}$ and $E[b_1 a b_2] = b_1 E[a] b_2$ for all $a \in \mathscr{A}$, $b_1, b_2 \in \mathscr{B}$; this generalizes Definition 1. If $\mathscr{A}$ and $\mathscr{B}$ are even C^*-algebras and if E is positive in the sense that $E[a^*a] \geq 0$ holds for each $a \in \mathscr{A}$, then $(\mathscr{A}, E, \mathscr{B})$ is called an *operator-valued C^*-probability space*, in analogy to Definition 2. In the latter case, if we take any selfadjoint $X \in \mathscr{A}$, then the *$\mathscr{B}$-valued Cauchy transform of X* is defined by

$$G_X : \mathbb{H}^+(\mathscr{B}) \to \mathbb{H}^-(\mathscr{B}), \quad b \mapsto E\big[(b - X)^{-1}\big],$$

where the upper respectively lower half plane in $\mathscr{B}$ are given by

$$\mathbb{H}^+(\mathscr{B}) = \{b \in \mathscr{B} \mid \Im(b) > 0\} \qquad \text{and} \qquad \mathbb{H}^-(\mathscr{B}) = \{b \in \mathscr{B} \mid \Im(b) < 0\}$$

with $\Im(b) = \frac{1}{2i}(b - b^*)$. Below, we will also use the so-called *h-transform of X*, which is given by

$$h_X : \mathbb{H}^+(\mathscr{B}) \to \overline{\mathbb{H}^+(\mathscr{B})}, \quad b \mapsto G_X(b)^{-1} - b.$$

Now, if N is the matrix size of the linearization $\hat{p}$, then the underlying C^*-probability space $(\mathscr{A}, \varphi)$ induces via $(M_N(\mathbb{C}) \otimes \mathscr{A}, 1 \otimes \varphi, M_N(\mathbb{C}))$ an operator-valued C^*-probability space, in which we may interpret (9) as

$$G_{p(x,y)}(z) = \lim_{\epsilon \searrow 0} \left[G_{\hat{p}}(\Lambda_\epsilon(z)) \right]_{1,1}, \qquad \text{where} \qquad \Lambda_\epsilon(z) = \begin{pmatrix} z & 0 & \dots & 0 \\ 0 & i\epsilon & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & i\epsilon \end{pmatrix}.$$

Note that the only reason for having introduced the limit $\epsilon \searrow 0$ is that we can move the point $\Lambda(z)$ to $\Lambda_\epsilon(z)$, which clearly belongs to the natural domain $\mathbb{H}^+(M_N(\mathbb{C}))$ of the $M_N(\mathbb{C})$-valued Cauchy transform $G_{\hat{p}}$. Hence, what we need in order to calculate the distribution of the non-linear scalar polynomial $p(x, y)$ is to calculate the operator-valued distribution (via its operator-valued Cauchy transform) of the operator-valued linear polynomial

$$\hat{p} = \alpha_0 \otimes 1 + \alpha_1 \otimes x + \alpha_2 \otimes y.$$

But this is exactly the realm of operator-valued free convolution (operator-valued freeness is defined in the same way as usual freeness, one just has to replace the state φ by the conditional expectation E), for which we have a well-developed analytic theory [4]. We only need to note that if x and y are free, then $X = \alpha_0 \otimes 1 + \alpha_1 \otimes x$ and $Y = \alpha_2 \otimes y$ are free in the operator-valued sense; furthermore their operator-valued Cauchy transforms are determined via the distribution of x and of y, respectively, by

$$G_X(b) = \int_{\mathbb{R}} (b - \alpha_0 - t\alpha_1)^{-1} \, d\mu_x(t) \qquad \text{and} \qquad G_Y(b) = \int_{\mathbb{R}} (b - t\alpha_2)^{-1} \, d\mu_y(t).$$

Theorem 4 ([4]) *Consider an operator-valued C^*-probability space $(\mathscr{A}, E, \mathscr{B})$ and selfadjoint variables $X, Y \in \mathscr{A}$, which are free in the operator-valued sense. Then the operator-valued Cauchy transform of $X + Y$ can be calculated from the operator-valued Cauchy transforms G_X and G_Y in the following way: there exists a unique pair of (Fréchet-)holomorphic maps $\omega_1, \omega_2 : \mathbb{H}^+(\mathscr{B}) \to \mathbb{H}^+(\mathscr{B})$, such that*

$$G_X(\omega_1(b)) = G_Y(\omega_2(b)) = G_{X+Y}(b), \quad b \in \mathbb{H}^+(\mathscr{B})$$

holds, where the values $\omega_1(b)$ and $\omega_2(b)$ of the subordination functions ω_1 and ω_2 *at any point $b \in \mathbb{H}^+(\mathscr{B})$ are the unique fixed points of the functions*

$$f_b : \mathbb{H}^+(\mathscr{B}) \to \mathbb{H}^+(\mathscr{B}), \quad w \mapsto h_Y(b + h_X(w)) + b,$$
$$g_b : \mathbb{H}^+(\mathscr{B}) \to \mathbb{H}^+(\mathscr{B}), \quad w \mapsto h_X(b + h_Y(w)) + b,$$

and they can be obtained, for any initial point $w \in \mathbb{H}^+(\mathscr{B})$*, by*

$$\omega_1(b) = \lim_{n\to\infty} f_b^{\circ n}(w) \qquad \text{and} \qquad \omega_2(b) = \lim_{n\to\infty} g_b^{\circ n}(w).$$

By applying this algorithm to $p(x, y) = xy + yx + x^2$ and its linearization $\hat{p} = X + Y$ with

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \otimes 1 + \begin{pmatrix} 0 & 1 & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \otimes x, \qquad Y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \otimes y$$

we produced the distribution of $p(x, y)$ as shown in Fig. 6. One should realize that the solution of the fixed point equations has to be done by numerical methods. Usually there is no hope of finding explicit solutions of those equations. Hence it is important to have a description of the solution which is amenable to easily implementable and controllable numerical methods. The fixed point equations from Theorem 4 provide such a controllable convergent scheme.

3.3 Historical Remark

After the successful implementation of the above program it was brought to our attention by J. William Helton and Victor Vinnikov that the linearization trick is not new at all, but a well-known idea in many other mathematical communities, known under various names like

- Higman's trick ("The units of group rings": Higman 1940 [22])
- recognizable power series (automata theory: Kleene 1956 [28]; Schützenberger 1961 [32]; Fliess 1974 [14]; Berstel and Reutenauer 1984 [7])
- linearization by enlargement (ring theory: Cohn 1985 [10, 11]; Cohn and Reutenauer 1994 [12, 13]; Malcolmson 1978 [29])
- descriptor realization (control theory: Kalman 1963 [26, 27]; Ball, Malakorn, and Groenewald 2005 [3]; Helton, McCullough, and Vinnikov 2006 [20]; Kaliuzhnyi-Verbovetskyi and Vinnikov 2009/2012 [23, 24]; Volcic 2015 [37])

However, in most of those contexts dealing with polynomials is (in contrast to our application in free probability) kind of trivial and the real domain for the linearization idea are non-commutative rational functions. Since our algorithm for calculating the distribution of a polynomial in free variables is actually an algorithm on the level of linearizations, this implies right away that all we have said before should work equally well for non-commutative rational functions in free variables. Let us address these issues in the next section.

4 Distributions of Non-commutative Rational Functions in Free Variables

Let us start with giving a bit of background on non-commutative rational functions before we address their distributions.

4.1 Non-commutative Rational Functions

Non-commutative rational functions were introduced by Amitsur [1] in 1966, whose methods were developed further by Bergman [6] in 1970, and they were studied extensively by Cohn [10, 11], Cohn and Reutenauer [12, 13], and Malcolmson [29]; see also [23, 24, 37].

Roughly speaking, non-commutative rational functions are given by rational expressions in non-commuting variables, like

$$r(x,y) := (4-x)^{-1} + (4-x)^{-1}y\big((4-x) - y(4-x)^{-1}y\big)^{-1}y(4-x)^{-1},$$

where two expressions are considered to be identical, when they can be transformed into each other by algebraic manipulations. The set of all non-commutative rational functions forms a skew field, the so-called *free field*. This – although it conveys the right idea – does not provide a rigorous definition of non-commutative rational functions to work with, since we presuppose here the existence of the free field as an algebraic frame, in which we can perform our algebraic manipulations. As a kind of substitute for this we can use matrix evaluations:

- Given a non-commutative rational expression r in m variables, we denote by $\mathrm{dom}(r)$ the subset of $\coprod_{n\in\mathbb{N}} M_n(\mathbb{C})^m$ consisting of all m-tuples $(X_1,\ldots,X_m)$, for which the evaluation $r(X_1,\ldots,X_m)$ is defined; if $\mathrm{dom}(r)\neq\emptyset$, we call the rational expression r *non-degenerate*.
- Two non-degenerate rational expressions r_1 and r_2 in m variables are considered to be equivalent if we have

$$r_1(X_1,\ldots,X_m) = r_2(X_1,\ldots,X_m) \qquad \text{for all } (X_1,\ldots,X_m)\in \mathrm{dom}(r_1)\cap\mathrm{dom}(r_2).$$

One can show that the free field is obtained as the set of all equivalence classes of non-degenerate rational expressions, with operations defined on representatives; we refer the reader to [24] for more details.

In the terminology of [10, 11], the free field is more precisely the universal skew field of fractions for the ring of non-commutative polynomials in the variables $x_1,\ldots,x_m$. That non-commutative rational functions form a skew field means that each $r(x_1,\ldots,x_m)\neq 0$ is invertible. However, deciding whether $r(x_1,\ldots,x_m)=0$

is not an easy task. For example, one has non-trivial rational identities, like

$$x_2^{-1} + x_2^{-1}(x_3^{-1}x_1^{-1} - x_2^{-1})^{-1}x_2^{-1} - (x_2 - x_3x_1)^{-1} = 0.$$

In the commutative situation, every rational function can be written as a fraction, i.e., the quotient of two polynomials. This is not true any more in the non-commutative case, and in general nested inversions are needed. So in the expression $r(x, y)$ from above we have a two-fold nested inversion. There are other ways of writing $r(x, y)$, but none of them can do without such a nested inversion. Whereas dealing with non-commutative rational functions just on the scalar level seems to be quite involved, going over to a matrix-level makes things again easier. In fact, it turns out that any non-commutative rational function can always be realized in the form of (8), namely in terms of matrices of polynomials, such that only one inverse is involved; in addition, we can achieve that the polynomials in the realization are linear. More precisely, over the free field, we can always find a representation of the form

$$r(x_1, \ldots, x_m) = -uQ(x_1, \ldots, x_m)^{-1}v, \tag{10}$$

where u, v are scalar row and column vectors, respectively, and $Q(x_1, \ldots, x_m)$ is an invertible matrix of corresponding size, whose entries are affine linear polynomials in the variables $x_1, \ldots, x_m$. For example, our $r(x, y)$ from above can be represented as

$$r(x, y) = -\begin{pmatrix} \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} -1 + \frac{1}{4}x & \frac{1}{4}y \\ \frac{1}{4}y & -1 + \frac{1}{4}x \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}. \tag{11}$$

Such representations appear for instance in [13, 29]; in [13], where they go under the name *pure linear representations*, they were used for an alternative construction of the free field. For non-commutative rational functions $r(x_1, \ldots, x_m)$, which are *regular at zero* (meaning that $0 \in \mathrm{dom}(r)$ holds – at least after suitable algebraic manipulations), such representations are called *non-commutative descriptor realizations*; see [3, 19, 20, 23, 24, 26, 27, 37].

4.2 Linearization for Non-commutative Rational Functions

As we have learned above, a realization like in (10) is according to (4) more or less the same as a linearization; the realization (11) of $r(x, y)$ yields directly a linearization $\hat{r}$ of the form

$$\hat{r} = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & -1 + \frac{1}{4}x & \frac{1}{4}y \\ 0 & \frac{1}{4}y & -1 + \frac{1}{4}x \end{pmatrix}.$$

Since this fits into the frame of our machinery for the calculation of distributions, one is tempted to believe that these methods extend also to non-commutative rational functions. This is indeed the case, but there is one hidden subtlety, which requires clarification: representations of the form (10) provide formulas for the non-commutative rational function $r(x_1, \ldots, x_m)$ and are thus valid only over the free field; it is not clear, if those formulas remain valid under evaluation of the involved rational expression r, i.e., when the variables $x_1, \ldots, x_m$ of the free field are replaced by elements from any non-commutative probability space $(\mathcal{A}, \varphi)$.

Such questions were addressed in [19]. The focus there was mainly on the case of non-commutative rational functions regular at zero, but most of the arguments pass directly to the frame of [13]. In any case, it turns out that rational identities are not necessarily preserved under evaluations on general algebras $\mathcal{A}$. However, it works well for the important class of *stably finite* algebras $\mathcal{A}$ (sometimes also addressed as *weakly finite*): if for $(X_1, \ldots, X_m) \in \mathcal{A}^m$ both $r(X_1, \ldots, X_m)$ is defined and $Q(X_1, \ldots, X_m)$ is invertible over $\mathcal{A}$, then $r(X_1, \ldots, X_m) = -uQ(X_1, \ldots, X_m)^{-1}v$ holds true. Notably, it was proven in [19] that, under certain conditions on the representation (10), the invertibility of $Q(X_1, \ldots, X_m)$ is automatically given, whenever $r(X_1, \ldots, X_m)$ is defined. It can be shown that if $(\mathcal{A}, \varphi)$ is a C^*-probability space, endowed with a faithful tracial state φ, then $\mathcal{A}$ must be stably finite.

When working in such a setting, our machinery applies. For the given r, the linearization $\hat{r}$ splits into a term X depending only on x and a term Y depending only on y: i.e., we have $\hat{r} = X + Y$ with

$$X = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \otimes 1 + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix} \otimes x, \qquad Y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{pmatrix} \otimes y.$$

If x and y are free, then X and Y are free in the operator-valued sense, and this is again an operator-valued free convolution problem, which can be solved as before by applying Theorem 4. The dotted curve in Fig. 7 shows the result of such a calculation for our r from above.

4.3 Rational Functions of Random Matrices and Their Limit

In Fig. 7 we compare again the distribution of $r(x, y)$ with the histogram of the eigenvalues of $r(X_N, Y_N)$ for independent Wigner matrices X_N, Y_N. One point one has to realize in the context of rational functions is that they cannot be evaluated on all operators. Clearly, we should only plug in operators for which all needed inverses make sense. So we have chosen an r which has two free semicirculars x and y in its domain. (For example, we have to invert $4 - x$; this is okay, because the spectrum of x is $[-2, 2]$.) If we approximate x, y by X_N, Y_N we hope that $r(X_N, Y_N)$ also

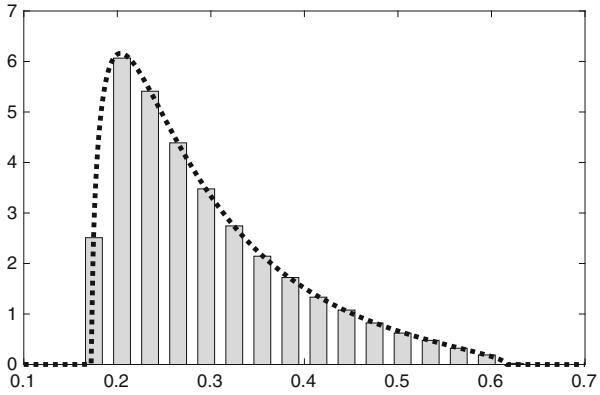


Fig. 7 As in Fig. 6, we consider the convergence of two independent Wigner matrices X_N, Y_N to two free semicircular operators x, y. We have then also for non-commutative rational functions r the almost sure convergence of $r(X_N, Y_N)$ to $r(x, y)$ in distribution as well as the convergence of the operator norms – again, there are no outlier eigenvalues outside the limiting spectrum. Here we have $r(x, y) = (4 - x)^{-1} + (4 - x)^{-1} y\big((4 - x) - y(4 - x)^{-1} y\big)^{-1} y(4 - x)^{-1}$

makes sense, at least for sufficiently large N. This is indeed the case, but relies on the fact that we also have good control on the largest eigenvalues. More precisely we have the following statement [39].

Proposition 1 ([39]) *Consider selfadjoint random matrices X_N, Y_N which converge to selfadjoint operators x, y in the following strong sense: for any selfadjoint polynomial p we have almost surely*

- $p(X_N, Y_N) \to p(x, y)$ *in distribution,*
- $\lim_{N\to\infty} \|p(X_N, Y_N)\| = \|p(x, y)\|$.

Then this strong convergence remains also true for rational functions: Let r be a selfadjoint non-commutative rational expression, such that $r(x, y)$ is defined. Then we have almost surely that

- $r(X_N, Y_N)$ *is defined eventually for large N,*
- $r(X_N, Y_N) \to r(x, y)$ *in distribution,*
- $\lim_{N\to\infty} \|r(X_N, Y_N)\| = \|r(x, y)\|$.

5 Non-selfadjoint Case: Brown Measure

The reader might wonder about our restriction to the case of selfadjoint polynomials (or rational functions). Why not consider arbitrary polynomials in random matrices or their limit operators, like

$$p(x_1, x_2, x_3, x_4) = x_1 x_2 + x_2 x_3 + x_3 x_4 + x_4 x_1?$$

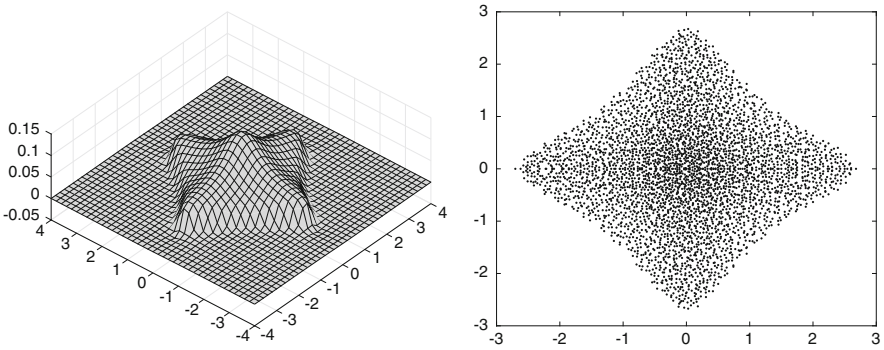


Fig. 8 The right plot shows the complex eigenvalues of the polynomial $p(X_1, X_2, X_3, X_4)$ in four independent Wigner matrices, each of size $N = 4000$. The left plot shows the Brown measure of the corresponding limit operator $p(s_1, s_2, s_3, s_4)$, calculated with our operator-valued free probability machinery. Here we have $p(x_1, x_2, x_3, x_4) = x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1$

Of course, we can (say for four independent Wigner matrices) just plug in our random matrices and calculate their eigenvalues. Those are now not real anymore, we will get instead a number of points in the complex plane, as in the right plot of Fig. 8.

The limit of four such independent Wigner matrices is given by four free semicircular elements s_1, s_2, s_3, s_4. The relevant information about $p := p(s_1, s_2, s_3, s_4)$ is given by its $*$-distribution, i.e., all moments in p and p^*. As p and p^* do not commute, this information cannot fully be captured by an analytic object, like a probability measure on $\mathbb{C}$. There is no straightforward substitute for the eigenvalue distribution for a non-normal operator. The full information about the non-normal operator p is given by its $*$-distribution, which is a highly non-trivial algebraic object. There is however a projection of this non-commutative algebraic object into the analytic classical world; namely, there exists a probability measure ν_p on $\mathbb{C}$, which captures some information about the $*$-distribution of p, and which is a canonical candidate for the limit of the eigenvalue distribution for the corresponding random matrix approximations. This ν_p is was introduced by Brown [9] in 1981 for operators in finite von Neumann algebras and is called the *Brown measure* of the operator p: let (M, τ) be a *tracial W^*-probability space*, i.e., a non-commutative probability space build out of a von Neumann algebra M and a faithful normal tracial state τ on M; for any given $x \in M$,

- the *Fuglede-Kadison determinant* $\Delta(x)$ is determined by

$$\log(\Delta(x)) = \int_{\mathbb{R}} \log(t)\, d\mu_{|x|}(t) \in \mathbb{R} \cup \{-\infty\},$$

  where $\mu_{|x|}$ denotes the analytic distribution of the operator $|x| = (x^*x)^{\frac{1}{2}}$ in the sense of (1).

- the *Brown measure* ν_x is the compactly supported Radon probability measure on $\mathbb{C}$, which is uniquely determined by the condition

$$\int_{\mathbb{C}} \psi(z)\, d\nu_x(z) = \frac{1}{2\pi} \int_{\mathbb{C}} \nabla^2 \psi(z) \log(\Delta(x-z))\, d\Re(z)\, d\Im(z)$$

  for all compactly supported C^∞-functions $\psi : \mathbb{C} \to \mathbb{C}$, where $\nabla^2\psi$ denotes the Laplacian of ψ, i.e., $\nabla^2\psi = \frac{\partial^2\psi}{\partial\Re(z)^2} + \frac{\partial^2\psi}{\partial\Im(z)^2}$.

It can be shown that the support of ν_x is always contained in the spectrum of the operator x. For matrices, the Brown measure coincides with the eigenvalue distribution. For selfadjoint operators, the Brown measure is just the spectral distribution with respect to the trace.

It turns out that we can refine the algorithm for selfadjoint polynomials or rational functions also to the non-selfadjoint case in order to calculate the Brown measure of arbitrary polynomials or rational functions in free variables. The result of this machinery for the polynomial $p(s_1, s_2, s_3, s_4)$ from above is shown in the left plot of Fig. 8. For more facts about the Brown measure as well as for the details on how free probability allows to calculate it, see [5, 8, 16, 19].

One problem in this context is that the construction of the Brown measure is not continuous with respect to convergence in $*$-distribution, i.e., knowing that our independent Wigner matrices X_1, X_2, X_3, X_4 converge in $*$-distribution to s_1, s_2, s_3, s_4 does not guarantee that the Brown measure of $p(X_1, X_2, X_3, X_4)$ converges to the Brown measure of $p(s_1, s_2, s_3, s_4)$. It is an open conjecture that this is indeed the case for all polynomials or even rational functions in independent Wigner matrices.

Acknowledgements This work was supported by the ERC Advanced Grant "Non-commutative Distributions in Free Probability" (grant no. 339760).

References

1. Amitsur, S.A.: Rational identities and applications to algebra and geometry. J. Algebra **3**, 304–359 (1966)
2. Anderson, G.W.: Convergence of the largest singular value of a polynomial in independent Wigner matrices. Ann. Probab. **41**(3B), 2103–2181 (2013)
3. Ball, J.A., Groenewald, G., Malakorn, T.: Structured noncommutative multidimensional linear systems. SIAM J. Control Optim. **44**(1), 1474–1528 (2005)
4. Belinschi, S.T., Mai, T., Speicher, R.: Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem. J. Reine Angew. Math. **732**, 21–53 (2017)
5. Belinschi, S.T., Sniady, P., Speicher, R.: Eigenvalues of non-hermitian random matrices and Brown measure of non-normal operators: hermitian reduction and linearization method. Linear Algebra Appl. **537**, 48–83 (2018)

6. Bergman, G.W.: Skew fields of noncommutative rational functions (preliminary version). Séminaire Schützenberger **1**, 1–18 (1969–1970)
7. Berstel, J., Reutenauer, C.: Rational Series and Their Languages. Springer, Berlin (1988)
8. Biane, P., Lehner, F.: Computation of some examples of Brown's spectral measure in free probability. Colloq. Math. **90**(2), 181–211 (2001)
9. Brown, L.G.: Lidskii's theorem in the type II case. In: Araki, H. (ed.) Geometric Methods in Operator Algebras, Proceedings of the US-Japan Seminar, Kyoto 1983. Pitman Research Notes in Mathematics Series, vol. 123, pp. 1–35. Longman Scientific and Technical/Wiley, New York/Harlow (1986)
10. Cohn, P.M.: Free Rings and Their Relations, 2nd edn. London Mathematical Society Monographs, vol. 19. Academic Press, London (Harcourt Brace Jovanovich, Publishers), XXII, p. 588 (1985)
11. Cohn, P.M.: Free Ideal Rings and Localization in General Rings. Cambridge University Press, Cambridge (2006)
12. Cohn, P.M., Reutenauer, C.: A normal form in free fields. Can. J. Math. **46**(3), 517–531 (1994)
13. Cohn, P.M., Reutenauer, C.: On the construction of the free field. Int. J. Algebra Comput. **9**(3–4), 307–323 (1999)
14. Fliess, M.: Sur divers produits de séries formelles. Bull. Soc. Math. Fr. **102**, 181–191 (1974)
15. Füredi, Z., Komloś, J.: The eigenvalues of random symmetric matrices. Combinatorica **1**, 233–241 (1981)
16. Haagerup, U., Larsen, F.: Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras. J. Funct. Anal. **176**(2), 331–367 (2000)
17. Haagerup, U., Schultz, H., Thorbjørnsen, S.: A random matrix approach to the lack of projections in $C^*_{\mathrm{red}}(\mathbb{F}_2)$. Adv. Math. **204**(1), 1–83 (2006)
18. Haagerup, U., Thorbjørnsen, S.: A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. Ann. Math. (2) **162**(2), 711–775 (2005)
19. Helton, J.W., Mai, T., Speicher, R.: Applications of realizations (aka Linearizations) to free probability. J. Funct. Anal. **274**(1), 1–79 (2018)
20. Helton, J.W., McCullough, S.A., Vinnikov, V.: Noncommutative convexity arises from linear matrix inequalities. J. Funct. Anal. **240**(1), 105–191 (2006)
21. Hiai, F., Petz, D.: The Semicircle Law, Free Random Variables and Entropy. American Mathematical Society (AMS), Providence (2000)
22. Higman, G.: The units of group-rings. Proc. Lond. Math. Soc. (2) **46**, 231–248 (1940)
23. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Singularities of rational functions and minimal factorizations: the noncommutative and the commutative setting. Linear Algebra Appl. **430**(4), 869–889 (2009)
24. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Noncommutative rational functions, their difference-differential calculus and realizations. Multidim. Syst. Signal Process. **23**(1–2), 49–77 (2012)
25. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Foundations of Free Noncommutative Function Theory. American Mathematical Society (AMS), Providence (2014)
26. Kalman, R.E.: Mathematical description of linear dynamical systems. J. Soc. Ind. Appl. Math. Ser. A Control **1**, 152–192 (1963)
27. Kalman, R.E.: Realization theory of linear dynamical systems. In: Control Theory and Topics in Functional Analysis, vol. II. Lecture Presented at the International Seminar Course, Trieste, vol. 1974, pp. 235–256 (1976)
28. Kleene, S.C.: Representation of Events in Nerve Nets and Finite Automata. Automata Studies, p. 341. Princeton University Press, Princeton (1956)
29. Malcolmson, P.: A prime matrix ideal yields a skew field. J. Lond. Math. Soc. II Ser. **18**, 221–233 (1978)
30. Mingo, J.A., Speicher, R.: Free Probability and Random Matrices. Fields Institute Monographs, vol. 35. Springer, New York (2017)

31. Nica, A., Speicher, R.: Lectures on the Combinatorics of Free Probability. Cambridge University Press, Cambridge (2006)
32. Schützenberger, M.P.: On the definition of a family of automata. Inf. Control **4**, 245–270 (1961)
33. Taylor, J.L.: A general framework for a multi-operator functional calculus. Adv. Math. **9**, 183–252 (1972)
34. Taylor, J.L.: Functions of several noncommuting variables. Bull. Am. Math. Soc. **79**, 1–34 (1973)
35. Voiculescu, D.: Limit laws for random matrices and free products. Invent. Math. **104**(1), 201–220 (1991)
36. Voiculescu, D., Dykema, K.J., Nica, A.: Free random variables. A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups. American Mathematical Society, Providence (1992)
37. Volcic, J.: Matrix coefficient realization theory of noncommutative rational functions. J. Algebra **499**, 397–437 (2018)
38. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. (2) **62**, 548–564 (1955)
39. Yin, S.: Non-commutative rational functions in strongly convergent random variables. Adv. Oper. Theory **3**(1), 190–204 (2018)

A Review on Comodule-Bialgebras

Dominique Manchon

Abstract We review some recent applications of the notion of comodule-bialgebra in several domains such as Combinatorics, Analysis and Quantum Field Theory.

1 Introduction

Let $\mathcal{B}$ be a unital counital bialgebra over some field k. Comodule-bialgebras on $\mathcal{B}$ are bialgebras in the monoidal category of left comodules over $\mathcal{B}$: if the coalgebra structure of $\mathcal{B}$ is enough to define the notion of left comodule, the algebra structure is needed in order to turn the class of left $\mathcal{B}$-comodules into a monoidal category, which is moreover symmetric under the usual flip of arguments when the bialgebra $\mathcal{B}$ is commutative. More precisely, let $\mathcal{H}$ be another unital counital bialgebra over k. We say [31, Definition 2.1.(e)] that $\mathcal{H}$ is a *comodule-bialgebra* on $\mathcal{B}$ if there exists a linear map

$$\Phi : \mathcal{H} \to \mathcal{B} \otimes \mathcal{H}$$

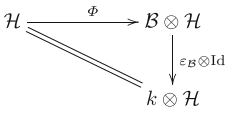
such that:

- Φ is a left coaction, i.e. the following diagrams commute:

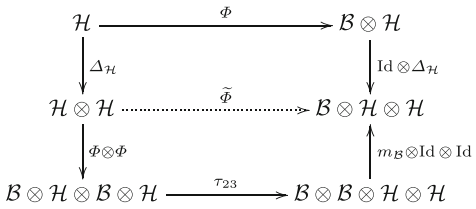
$$\begin{array}{ccc} \mathcal{H} & \xrightarrow{\Phi} & \mathcal{B} \otimes \mathcal{H} \\ \downarrow{\scriptstyle \Phi} & & \downarrow{\scriptstyle \Delta_{\mathcal{B}} \otimes \mathrm{Id}} \\ \mathcal{B} \otimes \mathcal{H} & \xrightarrow{\mathrm{Id} \otimes \Phi} & \mathcal{B} \otimes \mathcal{B} \otimes \mathcal{H} \end{array}$$

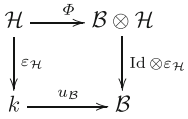
D. Manchon (✉)
C.N.R.S.-UMR 6620, Université Clermont-Auvergne, Aubière, France
e-mail: manchon@math.univ-bpclermont.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_20



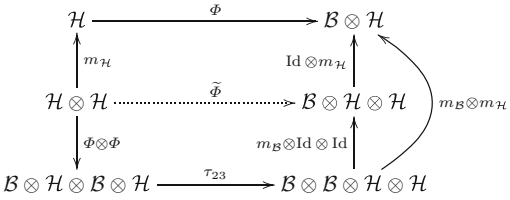
- The coproduct $\Delta_{\mathcal{H}}$ and the counit $\varepsilon_{\mathcal{H}}$ are morphisms of left $\mathcal{B}$-comodules. This amounts to the commutativity of the following two diagrams:

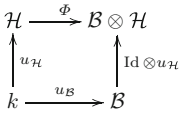




where τ_{23} stands for the flip of the two middle factors.

- Φ is a unital algebra morphism. This amounts to say that $m_{\mathcal{H}}$ and $u_{\mathcal{H}}$ are morphisms of left $\mathcal{B}$-comodules, where, as above, the comodule structure on k is given by the unit map $u_{\mathcal{B}}$, and the comodule structure on $\mathcal{H}\otimes\mathcal{H}$ is given by $\widetilde{\Phi}=(m_{\mathcal{B}}\otimes\mathrm{Id}\otimes\mathrm{Id})\circ\tau_{23}\circ(\Phi\otimes\Phi)$. In other words, the following two diagrams commute:





The comodule-bialgebra $\mathcal{H}$ is a *comodule-Hopf algebra* if moreover $\mathcal{H}$ is a Hopf algebra with antipode S such that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{H} & \xrightarrow{\Phi} & \mathcal{B} \otimes \mathcal{H} \\ \downarrow{\scriptstyle S} & & \downarrow{\scriptstyle I \otimes S} \\ \mathcal{H} & \xrightarrow{\Phi} & \mathcal{B} \otimes \mathcal{H} \end{array}$$

Weaker definitions are possible, for example by dropping the co-unitality of the bialgebra $\mathcal{H}$. Examples of non-co-unital comodule-bialgebras will be provided in Sect. 9.

Comodule-bialgebras and comodule-Hopf algebras are ubiquitous in various domains of Mathematics. They arise as soon as a pro-unipotent monoid scheme or group scheme acts on another pro-unipotent monoid or group scheme by morphisms: to be precise, the monoid scheme of characters of $\mathcal{B}$ acts by morphisms on the monoid scheme (resp. the group scheme) of characters of the bialgebra (resp. Hopf algebra) $\mathcal{H}$.

Two main examples of this situation are the following: comodule-bialgebras govern product and composition in mould calculus [14], and also composition and substitution of B-series in numerical analysis [10]. The latter is closely related to product and composition of arborified moulds[1] [17, 20]. Both group scheme and monoid scheme acting on it are given by characters of the quasi-shuffle algebra for mould calculs, by characters of the algebra of rooted forests for B-series. The group product is given by the deconcatenation (resp. Connes-Kreimer) coproduct, and the monoid product is mould composition (resp. B-series substitution), given by another compatible "internal" coproduct.

We will review the comodule-Hopf algebra structures at play in the two situations mentioned above, and give an account of some interesting generalizations: cycle free graphs [29], posets and finite topological spaces [18]. We will also give a brief account on how comodule-Hopf algebras occur in the theory of regularity structures [24], reflecting the action of the negative renormalization group on the positive renormalization group [3]. Finally, we give an example of an infinite family of comodule-bialgebras, indexed by pairs of integers (i, j) with $1 \leq i < j$, over the same bialgebra. These comodule-bialgebras are not co-unital except for $i = 1$.

2 The Adjoint Corepresentation

In this Section, following [31, Proposition 2.5] (see also [27, Exercise IX.8.5]), we will give the canonical example of $\mathcal{H}$-comodule-bialgebra structure on $\mathcal{H}$ for any *commutative* Hopf algebra $\mathcal{H}$. The antipode is necessary: it is the Hopf-algebraic

[1] By an unfortunate conflict of terminology, *composition* of B-series corresponds to *product* of arborified moulds, whereas *composition* of the latter corresponds to *substitution* of B-series.

counterpart of the action of any group on itself by conjugation:

$$\begin{aligned}\kappa : G \times G &\longrightarrow G\\ (\gamma, g) &\longmapsto \gamma g \gamma^{-1}.\end{aligned}$$

The coaction $\Phi : \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$ is given by:

$$\Phi = (m_{\mathcal{H}} \otimes \mathrm{Id}) \circ (\mathrm{Id} \otimes S \otimes \mathrm{Id}) \circ \tau_{23} \circ (\Delta_{\mathcal{H}} \otimes \mathrm{Id}) \circ \Delta_{\mathcal{H}}.$$

In Sweedler's notation,

$$\Phi(x) = \sum_{(x)} x_1 S(x_3) \otimes x_2.$$

If $\mathcal{H}$ is not commutative, the adjoint corepresentation Φ may fail to be an algebra morphism. In that case, $\mathcal{H}$ is only a *comodule-coalgebra* on itself.

3 J. Ecalle's Mould Calculus: Product, Composition, Arborification

We closely follow the original presentation of J. Ecalle in [14, Paragraph 4b]. Let Ω be an alphabet endowed with a commutative semigroup law written additively: for example we can choose $\Omega = \mathbb{N}^* = \{1, 2, 3, \ldots\}$. A word with letters in Ω will be denoted by:

$$\boldsymbol{\omega} = \omega_1 \cdots \omega_n.$$

For later use we denote by Ω^* the monoid of words with letters in Ω. The empty word is denoted by $\mathbf{1}$. The *length* of $\boldsymbol{\omega}$ is its number $|\boldsymbol{\omega}| = n$ of letters. The *weight* of the word $\boldsymbol{\omega}$ is the sum of its letters:

$$\|\boldsymbol{\omega}\| := \left[\sum_{i=1}^{n} \omega_i \right],$$

where the brackets indicate the internal sum in Ω, in contrast with formal linear combinations which will be widely used in the sequel. The weight takes values in the monoid $\overline{\Omega} := \Omega \sqcup \{0\}$. The unique word of length and weight zero is the empty word $\mathbf{1}$. The concatenation of two words $\boldsymbol{\omega} = \omega_1 \cdots \omega_p$ and $\boldsymbol{\omega}' = \omega_{p+1} \cdots \omega_{p+q}$ is given by:

$$\boldsymbol{\omega}.\boldsymbol{\omega}' = \omega_1 \cdots \omega_{p+q},$$

and we obviously have

$$|\boldsymbol{\omega}.\boldsymbol{\omega}'| = |\boldsymbol{\omega}| + |\boldsymbol{\omega}'| \text{ (addition in } \mathbb{N}), \qquad \|\boldsymbol{\omega}.\boldsymbol{\omega}'\| = \|\boldsymbol{\omega}\| + \|\boldsymbol{\omega}'\| \text{ (addition in } \overline{\Omega}). \tag{1}$$

3.1 Moulds: Product and Composition

A *mould* on the alphabet Ω is a function with values in some commutative unital algebra $\mathcal{A}$, depending on a variable number of variables in Ω. Considering the vector space $\mathcal{H}$ spanned by the words on Ω, a mould (strictly speaking, its linear extension) is a linear map M on $\mathcal{H}$ with values in $\mathcal{A}$. The evaluation of M at a word $\boldsymbol{\omega}$ will be denoted by $M^{\boldsymbol{\omega}}$. Mould multiplication and mould composition are respectively defined by:

$$(M \times N)^{\boldsymbol{\omega}} = \sum_{\boldsymbol{\omega}'.\boldsymbol{\omega}''=\boldsymbol{\omega}} M^{\boldsymbol{\omega}'} N^{\boldsymbol{\omega}''}, \tag{2}$$

$$(M \circ N)^{\emptyset} = M^{\emptyset}, \qquad (M \circ N)^{\boldsymbol{\omega}} = \sum_{s \geq 1} \sum_{\boldsymbol{\omega}=\boldsymbol{\omega}^1 \cdots \boldsymbol{\omega}^s} M^{\|\boldsymbol{\omega}^1\| \cdots \|\boldsymbol{\omega}^s\|} N^{\boldsymbol{\omega}^1} \cdots N^{\boldsymbol{\omega}^s}. \tag{3}$$

The composition (3) is defined as long as $N^{\emptyset} = 0$. Both operations are associative, and composition distributes on the right over multiplication, namely:

$$(M \times M') \circ N = (M \circ N) \times (M' \circ N) \tag{4}$$

for any triple of moulds (M, M', N). The unit for the product is the mould ε defined by $\varepsilon^{\emptyset} = 1$ and $\varepsilon^{\boldsymbol{\omega}} = 0$ for any nontrivial word $\boldsymbol{\omega}$. The unit for composition is the mould I defined by $I^{\omega} = 1$ for $\omega \in \Omega$ and $I^{\boldsymbol{\omega}} = 0$ if $\boldsymbol{\omega} = \emptyset$ or if $\boldsymbol{\omega}$ is a word of length ≥ 2.

Whereas the associativity of the mould product is easily checked (it is nothing but the convolution product dual to the deconcatenation coproduct), the two other properties involving mould composition are better seen when the latter is interpreted as a substitution of alphabets: any $\mathcal{A}$-valued mould M gives rise to a *word series* [34, 35] W^M given by:

$$W^M := \sum_{\boldsymbol{\omega} \in \Omega^*} M^{\boldsymbol{\omega}} \boldsymbol{\omega}, \tag{5}$$

which obviously determines the mould M. The space of noncommutative formal series with variables in Ω (word series) with coefficients in $\mathcal{A}$ is denoted by $\mathcal{A}\langle\langle \Omega \rangle\rangle$. The homogeneous components of W^M are given by:

$$\iota^M(\kappa) := \sum_{\boldsymbol{\omega} \in \Omega^*,\ ||\boldsymbol{\omega}||=\kappa} M^{\boldsymbol{\omega}} \boldsymbol{\omega}. \tag{6}$$

This gives rise to a linear map $\iota^M : \Omega \to \mathcal{A}\langle\langle\Omega\rangle\rangle$, which uniquely extends by $\mathcal{A}$-linearity, multiplicativity and completion, to an $\mathcal{A}$-algebra endomorphism $\jmath^M : \mathcal{A}\langle\langle\Omega\rangle\rangle \to \mathcal{A}\langle\langle\Omega\rangle\rangle$. Remark that the word series of the mould I is given by the formal sum of the letters:

$$W^I = \sum_{\omega\in\Omega} \omega. \tag{7}$$

From (5), (6) and (7) we immediately get for any mould M:

$$W^M = \jmath^M(W^I). \tag{8}$$

Proposition 1 *Let M, N be two $\mathcal{A}$-valued moulds on the alphabet Ω, where $\mathcal{A}$ is a commutative unital $\mathbf{k}$-algebra. Then:*

$$W^{M\times N} = W^M.W^N, \tag{9}$$

$$\jmath^M \circ \jmath^N = \jmath^{M\circ N}. \tag{10}$$

Proof Proving the first assertion is straightforward:

$$\begin{aligned} W^{M\times N} &= \sum_{\boldsymbol{\omega}\in\Omega^*} (M\times N)^{\boldsymbol{\omega}}\boldsymbol{\omega} \\ &= \sum_{\boldsymbol{\omega}\in\Omega^*} \sum_{\boldsymbol{\omega}'.\boldsymbol{\omega}''=\boldsymbol{\omega}} M^{\boldsymbol{\omega}'}N^{\boldsymbol{\omega}''}\boldsymbol{\omega} \\ &= \sum_{\boldsymbol{\omega}',\boldsymbol{\omega}''\in\Omega^*} M^{\boldsymbol{\omega}'}N^{\boldsymbol{\omega}''}\boldsymbol{\omega}'.\boldsymbol{\omega}'' \\ &= W^M.W^N. \end{aligned}$$

Now let κ be any letter in Ω, and compute:

$$\begin{aligned} \jmath^N \circ \jmath^M(\kappa) &= \jmath^N\Big(\sum_{\substack{\boldsymbol{\omega}\in\Omega^* \\ ||\boldsymbol{\omega}||=\kappa}} M^{\boldsymbol{\omega}}\boldsymbol{\omega}\Big) \\ &= \sum_{\substack{\boldsymbol{\omega}\in\Omega^* \\ ||\boldsymbol{\omega}||=\kappa}} M^{\boldsymbol{\omega}}\jmath^N(\boldsymbol{\omega}) \\ &= \sum_{r\geq 1} \sum_{\substack{\boldsymbol{\omega}\in\Omega^* \\ ||\boldsymbol{\omega}||=\kappa,\ |\boldsymbol{\omega}|=r}} M^{\boldsymbol{\omega}}\jmath^N(\omega_1)\cdots\jmath^N(\omega_r) \\ &= \sum_{r\geq 1} \sum_{\substack{\boldsymbol{\omega}\in\Omega^* \\ \left[||\boldsymbol{\omega}^1||+\cdots+||\boldsymbol{\omega}^r||\right]=\kappa}} M^{||\boldsymbol{\omega}^1||\cdots||\boldsymbol{\omega}^r||}N^{\boldsymbol{\omega}^1}\cdots N^{\boldsymbol{\omega}^r}\boldsymbol{\omega}^1\cdots\boldsymbol{\omega}^r \end{aligned}$$

$$= \sum_{r\geq 1} \sum_{\substack{\omega\in\Omega^* \\ ||\omega||=\kappa}} \Big(\sum_{\omega^1\cdots\omega^r=\omega} M^{||\omega^1||\cdots||\omega^r||} N^{\omega^1}\cdots N^{\omega^r} \Big)\omega$$

$$= \jmath^{M\circ N}(\kappa).$$

Both $\jmath^{M\circ N}$ and $\jmath^N \circ \jmath^M$ are algebra morphisms that coincide on letters from Ω, hence they are equal.

Remark 1 The associativity of the mould composition is an immediate consequence of Proposition 1, namely:

$$W^{M\circ(N\circ P)} = \jmath^P \circ \jmath^N \circ \jmath^M(W^I) = W^{(M\circ N)\circ P}. \tag{11}$$

The distributivity property (4) is also easily checked:

$$\begin{aligned} W^{(M\times M')\circ N} &= \jmath^N \circ \jmath^{M\times M'}(W^I) = \jmath^N(W^{M\times M'}) \\ &= \jmath^N(W^M).\jmath^N(W^{M'}) = \jmath^N \circ \jmath^M(W^I).\jmath^N \circ \jmath^{M'}(W^I) \\ &= W^{M\circ N}.W^{M'\circ N} \\ &= W^{(M\circ N)\times(M'\circ N)}. \end{aligned}$$

3.2 Arborification

The arborification of moulds appears in [15], as a tool to handle the analysis of local vector fields and diffeomorphisms. From a purely algebraic point of view, the arborification transform can be understood as follows [20]:

- The linear span $k < \Omega >$ of words on the alphabet Ω can be endowed with a connected filtered Hopf algebra structure in two ways. The coproduct is the deconcatenation:

$$\Delta\omega = \sum_{\omega'\omega''=\omega} \omega' \otimes \omega'',$$

and the product is either the shuffle ш or the quasi-shuffle product ⧢ : they are respectively defined by [26]

$$\mathbf{1} \text{ ш } \omega = \mathbf{1} \text{ ⧢ } \omega = \omega \text{ ш } \mathbf{1} = \omega \text{ ⧢ } \mathbf{1} = \omega$$

and by the recursive formulas:

$$a\omega' \text{ ш } b\omega'' = a(\omega' \text{ ш } b\omega'') + (a\omega' \text{ ш } \omega'')b,$$
$$a\omega' \text{ ⧢ } b\omega'' = a(\omega' \text{ ⧢ } b\omega'') + (a\omega' \text{ ⧢ } \omega'')b + [a+b](\omega' \text{ ⧢ } \omega''),$$

with $a, b \in \Omega$ and $\omega', \omega'' \in \Omega^*$. We denote by $\mathcal{H}_0^\Omega$ (resp. $\mathcal{H}^\Omega$) the shuffle Hopf algebra (resp. the quasi-shuffle Hopf algebra) thus obtained.

- Let $\mathcal{H}_<^\Omega$ be the linear span of rooted forests decorated[2] by Ω. It is the free commutative algebra over the linear span of Ω-decorated rooted trees. It is a graded commutative Hopf algebra [11, 12, 28], with the following coproduct:

$$\Delta(F) = \sum_{V_1 \sqcup V_2 = \mathcal{V}(F),\ V_1 < V_2} F_{|_{V_2}} \otimes F_{|_{V_1}}. \tag{12}$$

Here $\mathcal{V}(F)$ stands for the set of vertices of F, the restriction of the forest F to a subset of $\mathcal{V}(F)$ is obtained by keeping only the edges joining two vertices in the subset, and $V_1 < V_2$ means that for any $x \in V_1$ and $y \in V_2$, there is no path from one root to x through y. Such a couple (V_1, V_2) is called an *admissible cut* [32].

- For any $b \in \Omega$, the operator $B_+^b : \mathcal{H}_<^\Omega \to \mathcal{H}_<^\Omega$, defined by grafting all the trees of a forest on a common root decorated by b, verifies the following *cocycle equation*:

$$\Delta\big(B_+^b(F)\big) = (I \otimes B_+^b)\Delta(F) + B_+^b(F) \otimes \mathbf{1},$$

where $\mathbf{1}$ here stands for the empty forest. The operator $L^b : \mathcal{H}^\Omega \to \mathcal{H}^\Omega$ defined by $L^b(\omega) = \omega b$ verifies the same cocycle equation:

$$\Delta\big(L^b(\omega)\big) = (I \otimes L^b)\Delta(\omega) + L^b(\omega) \otimes \mathbf{1}.$$

This also stands for $\mathcal{H}_0^\Omega$, the coalgebra structure being the same. Now there are unique surjective Hopf algebra morphisms [21]

$$\mathfrak{a}_0 : \mathcal{H}_<^\Omega \longrightarrow\!\!\!\!\!\rightarrow \mathcal{H}_0^\omega, \qquad \mathfrak{a} : \mathcal{H}_<^\Omega \longrightarrow\!\!\!\!\!\rightarrow \mathcal{H}^\omega,$$

such that $\mathfrak{a}_0(\mathbf{1}) = \mathbf{1}$ and $\mathfrak{a}(\mathbf{1}) = \mathbf{1}$, and such that:

$$\mathfrak{a}_0 \circ B_+^b = L^b \circ \mathfrak{a}_0, \qquad \mathfrak{a} \circ B_+^b = L^b \circ \mathfrak{a},$$

called simple and contracting arborification respectively. For any mould $M_0 : \mathcal{H}_0^\Omega \to \mathcal{A}$, resp. $M : \mathcal{H}^\Omega \to \mathcal{A}$, the corresponding arborified mould will be $M_0^< := M_0 \circ \mathfrak{a}_0 : \mathcal{H}_<^\Omega \to \mathcal{A}$, resp. $M^< := M \circ \mathfrak{a} : \mathcal{H}_<^\Omega \to \mathcal{A}$.

A *symmetral* (resp. *symmetrel*) mould[3] is a character (i.e. a unital algebra morphism) from the shuffle (Hopf) algebra $\mathcal{H}_0^\Omega$ (resp. the quasi-shuffle algebra

[2] Only the vertices are decorated here: to be precise, an Ω-decorated forest is a pair (F, d) with F being a forest and $d : \mathcal{V}(F) \to \Omega$, where $\mathcal{V}(F)$ stands for the set of vertices of F.

[3] To be pronounced french-like, with stress on the last syllable. The last vowel (here, a or e) designates the type of symmetry considered.

$\mathcal{H}^{\Omega}$) into the commutative unital algebra $\mathcal{A}$. An *alternal* (resp. *alternel*) mould is an infinitesimal character from the shuffle (Hopf) algebra $\mathcal{H}_0^{\Omega}$ (resp. the quasi-shuffle algebra $\mathcal{H}^{\Omega}$) into the commutative unital algebra $\mathcal{A}$, i.e. a mould M is alternal (resp. alternel) if and only if

$$M^{\boldsymbol{\omega} \text{ш} \boldsymbol{\omega}'} = \varepsilon^{\boldsymbol{\omega}} M^{\boldsymbol{\omega}'} + M^{\boldsymbol{\omega}} \varepsilon^{\boldsymbol{\omega}'},$$

resp.

$$M^{\boldsymbol{\omega} \text{Ш} \boldsymbol{\omega}'} = \varepsilon^{\boldsymbol{\omega}} M^{\boldsymbol{\omega}'} + M^{\boldsymbol{\omega}} \varepsilon^{\boldsymbol{\omega}'}.$$

The simple (resp. contracting) arborification of a symmetral (resp. symmetrel) mould is obviously a character of the (Hopf) algebra $\mathcal{H}_<^{\Omega}$, a *separative arborified mould* in J. Ecalle's terminology. If M and N are symmetrel moulds, so is $M \circ N$: we shall sketch the proof of this fact in Sect. 5 below. It can also be shown that the composition of two alternal moulds is alternal [15].

The product of moulds is nothing but the convolution product with respect to the deconcatenation coproduct. It immediately implies that the product respects symmetrality and symmetrelity. One can ask the question whether the mould composition admits such a Hopf-algebraic interpretation. Comodule-bialgebras enter into the game precisely here: this will be the purpose of the next two sections below.

4 B-Series: Composition and Substitution

Arborified moulds and partitioned B-series [6, 25] are closely related (see also [2]). We shall see that the Hopf-algebraic interpretation of arborified mould composition lies in B-series *substitution* [10], whereas arborified mould product accounts for B-series *composition* through the Hairer–Wanner theorem [25, Theorem III.1.10]. In the non-partitioned case (i.e. if the alphabet Ω is reduced to one element), it has been proved in [7] that the substitution product of B-series is obtained by dualizing a bialgebra coaction on $\mathcal{H}_<^{\Omega}$, making it a comodule-Hopf algebra over another bialgebra $\mathcal{B}_<^{\Omega}$ of trees. The latter is identical to $\mathcal{H}_<^{\Omega}$ as an algebra, but the coproduct, defined by means of extraction-contraction, is completely different:

$$\Gamma(F) = \sum_{F' \subseteq F} F' \otimes F/F'. \tag{13}$$

Here F' is a *covering subforest* of F, and F/F' is the forest obtained by contracting the connected components of F' on a point. The bialgebra $\mathcal{B}_<^{\Omega}$, graded by the number of edges, contains a lot of non-invertible degree zero elements, and therefore is not a Hopf algebra. The coaction Φ is nothing but the coproduct Γ once $\mathcal{B}_<^{\Omega}$ and $\mathcal{H}_<^{\Omega}$ have been identified as free commutative algebras.

The partitioned case (i.e. when the alphabet Ω contains several elements) is treated similarly, except that the forests are now decorated by Ω. The main issue with the extraction-contraction coproduct Γ is to decide how to decorate the vertices produced by contraction of subtrees. An obvious way consists in taking advantage of the semigroup structure of Ω: the decoration is thus the sum of the decorations of the vertices of the contracted subtree. One exactly recovers *arborified mould composition* [16, 30] that way, where arborified moulds are identified with linear maps from $\mathcal{B}_<^{\Omega}$ to the target algebra $\mathcal{A}$. This procedure is also at hand in the theory of ∞-B-series considered in [33]. Another way consists in simply putting the decoration of the root: it reveals to be the most natural one to deal with partitioned B-series (P-series), for which the number of decorations is usually finite.

5 A Comodule-Bialgebra Interpretation of Mould Composition

It is therefore natural to look for a comodule-bialgebra interpretation of ordinary (i.e. non-arborified) mould calculus: indeed, it turns out [13] that the quasi-shuffle Hopf algebra $(\mathcal{H}^{\Omega}, \text{ш}, \Delta)$ is a comodule-bialgebra over a bialgebra $(\mathcal{H}^{\Omega}, \text{ш}, \Gamma)$ where the definition of the "internal" coproduct Γ is directly inspired from mould composition:

$$\Gamma(\boldsymbol{\omega}) := \sum_{s\geq 1} \sum_{\boldsymbol{\omega}=\boldsymbol{\omega}^1.\cdots.\boldsymbol{\omega}^s} \|\boldsymbol{\omega}^1\| \cdots \|\boldsymbol{\omega}^s\| \otimes \boldsymbol{\omega}^1 \text{ш} \cdots \text{ш} \boldsymbol{\omega}^s. \tag{14}$$

Similarly to the arborified case, the coaction Φ is just given by the coproduct Γ. Recalling that moulds (resp. symmetrel moulds) are linear maps (resp. unital algebra morphisms) from $\mathcal{H}^{\Omega}$ to some unital commutative algebra $\mathcal{A}$, we have for any moulds M and N:

$$(M \times N)^{\boldsymbol{\omega}} = (M \otimes N)^{\Delta(\boldsymbol{\omega})}, \tag{15}$$

and for any *symmetrel* moulds M and N:

$$(M \circ N)^{\boldsymbol{\omega}} = (M \otimes N)^{\Gamma(\boldsymbol{\omega})}. \tag{16}$$

To be precise, another mould composition $\diamond$ is defined in [13] such that for any moulds M and N we have:

$$(M \diamond N)^{\boldsymbol{\omega}} = (M \otimes N)^{\Gamma(\boldsymbol{\omega})}, \tag{17}$$

and the composition $\diamond$ coincides with the composition $\circ$ when the two moulds are symmetrel. As the convolution product respects unital algebra morphisms, one can infer from this fact that the composition of two symmetrel moulds is symmetrel.

As well as the comodule-bialgebra structure on $\mathcal{H}^{\Omega}_{<}$ governs composition and substitution of (partitioned) B-series, the comodule-bialgebra structure on $\mathcal{H}^{\Omega}$ governs composition and substitution of word series.

6 Cycle-Free Graphs

The situation described in the previous section can be transposed from forests to other combinatorial objects. Let $\mathcal{H}^{\Omega}_{CF}$ be the linear span of oriented graphs without oriented cycles, with Ω-decorated vertices. The product is given by simple juxtaposition of graphs. Taking advantage of the poset structure on the set of vertices, we can define a coproduct Δ by a formula similar to (12), and a bialgebra coproduct Γ by a formula similar to (18), except that one has to avoid the formation of oriented cycles in the contracted graph:

$$\Gamma(G) = \sum_{G' \subseteq G,\, G/G' \text{cycle-free}} G' \otimes G/G'. \tag{18}$$

The rules for decorating vertices of G/G' are the same than in the previous section. Then $(\mathcal{H}^{\Omega}_{CF}, \cdot, \Delta)$ is a comodule-Hopf algebra over $(\mathcal{H}^{\Omega}_{CF}, \cdot, \Gamma)$ [29].

7 Finite Topological Spaces, Posets

The set of topologies on a finite set X is in bijection with quasi-orders, i.e. reflexive transitive relations. The open sets are the upper ideals for the associated quasi-order. This quasi-order is a partial order (i.e. is antisymmetric) if and only if the corresponding topology is T_0. For any finite set X one can consider the vector space $\mathbb{T}_X$ spanned by the topologies on X. This gives rise to a linear species in the sense of A. Joyal [1], denoted by $\mathbb{T}$, which is endowed with an external coproduct $\Delta : \mathbb{T}_X \to \bigoplus_{Y \subset X} \mathbb{T}_Y \otimes \mathbb{T}_{X \setminus Y}$ and an internal coproduct $\Gamma : \mathbb{T}_X \to \mathbb{T}_X$. They are given for any topology $\mathcal{T}$ on X by:

$$\Delta(\mathcal{T}) = \sum_{Y \in \mathcal{T}} \mathcal{T}|_{X \setminus Y} \otimes \mathcal{T}|_{Y},$$

$$\Gamma(\mathcal{T}) = \sum_{\mathcal{T}' \circledS \mathcal{T}} \mathcal{T}' \otimes \mathcal{T}/\mathcal{T}'.$$

Here $\mathcal{T}' \circledS \mathcal{T}$ means that $\mathcal{T}'$ is finer than $\mathcal{T}$ and verifies a technical condition reminiscent of the absence of oriented cycles in contracted graphs. The "quotient" topology $\mathcal{T}/\mathcal{T}'$ is still a topology on X, the quasi-order of which is the transitive

closure of the relation $\mathcal{R}$ defined by:

$$x \mathcal{R} y \Longleftrightarrow x \leq_{\mathcal{T}} y \text{ or } y \leq_{\mathcal{T}'} x.$$

The two coproducts Δ and Γ are compatible in the sense that they give rise, by forgetting the labels, to a commutative graded Hopf algebra $\mathcal{H} = \bigoplus_{n \geq 0} \mathcal{H}_n$ endowed with an extra coproduct Γ internal to each homogeneous component $\mathcal{H}_n$, such that $(\mathcal{H}, \cdot, \Delta)$ is a comodule-Hopf algebra on the bialgebra $(\mathcal{H}, \cdot, \Gamma)$ [18]. The same construction can be done verbatim for the species $\mathbb{T}^{\Omega}$ of Ω-decorated finite topological spaces.[4]

Sticking to posets, i.e. to T_0-topologies, we can repeat the construction, except that $\mathcal{T}/\mathcal{T}'$ may be not T_0 even if $\mathcal{T}$ (and therefore $\mathcal{T}'$) is a T_0 topology on X. We have then to consider the T_0 quotient of the topological space $(X, \mathcal{T}/\mathcal{T}')$, and the coproduct Γ thus obtained is not internal anymore strictly speaking. The contracting rules for decorations must then be applied again in the Ω-decorated case. Forgetting the labels again, one recovers still another comodule-Hopf algebra.

Characters of the Hopf algebra obtained from finite posets are called *ormoulds* by J. Ecalle. Characters of the Hopf algebra obtained from finite topological spaces are accordingly called *quasi-ormoulds* in [18]. Ormoulds and quasi-ormoulds can be multiplied according to the external coproduct, and composed according to the comodule-Hopf algebra structure. For an operadic approach to cycle-free graphs, posets and quasi-posets, see [19, 22].

Ormoulds do not seem to have applications to analysis as arborified moulds do. This seems to be due to the *coarborification process* [15, 17], which has no counterpart for arbitrary finite posets. Coarborification is closely related to elementary differentials, and ultimately relies on the fact that the linear span of Ω-decorated rooted trees is the free pre-Lie algebra generated by Ω [9]: no such simple algebraic seem to exist on the linear span of Ω-decorated finite connected posets.

8 Regularity Structures: Structure Group and Renormalization Group

A *regularity structure* [23, 24] is a triple (A, T, G) where A is a set of real numbers bounded from below and without accumulation point (the *homogeneities*), T is a direct sum:

$$T = \bigoplus_{\alpha \in A} T_{\alpha}$$

[4] Some confusion can arise in the literature around the words "label" and "decoration". Here the decorations in Ω should not be confused with the labels, which are elements of the finite set X. Two distinct elements can bear the same decoration, but never the same label.

where each T_α is a Banach space, and G is a group acting by continuous linear operators on T, such that for any $\Gamma \in G$ and any $\alpha \in A$:

$$(\Gamma - \mathrm{Id}_T)(T_\alpha) \subset \bigoplus_{\beta<\alpha} T_\beta.$$

The group G is the *structure group* and is not necessarily Abelian. Regularity structures, roughly speaking, allow to write Taylor formulas at any point for irregular functions or even distributions. As such they give a rigorous meaning for a wide class of ill-posed irregular (for example stochastic) partial differential equations. Finding a solution for those needs a *renormalization group* $\mathcal{R}$ attached to the regularity structure. The crucial notion of *local subcriticality* ensures that this group is finite-dimensional. Both groups are obtained from character groups of Hopf algebras obtained from rooted forests, in which both the vertices and the edges are decorated by $\Omega = \mathbb{N}^d$, where d is a fixed dimension [3]. The two coproducts are given by formulas like (12) and (18), but with a subtler play with decorations.

It turns out that, at the cost of an extra decoration of the vertices by $\mathbb{Z}^{d'}$ with some $d' > d$, one of the Hopf algebras is a comodule-bialgebra over the other[5] [3, Remark 4.2.9]. These two Hopf algebras $\mathcal{H}_+^{\mathrm{ex}}$ and $\mathcal{H}_-^{\mathrm{ex}}$ co-act on a space $\mathcal{H}^{\mathrm{ex}}$ (by two coactions Ψ_+ and Ψ_- respectively) in a way compatible with the comodule-bialgebra structure Φ of $\mathcal{H}_+^{\mathrm{ex}}$ on $\mathcal{H}_-^{\mathrm{ex}}$, in the sense that the following diagram commutes:

$$\begin{array}{ccc}
\mathcal{H}^{\mathrm{ex}} & \xrightarrow{\Psi_+} & \mathcal{H}_+^{\mathrm{ex}} \otimes \mathcal{H}^{\mathrm{ex}} \\
\downarrow{\scriptstyle \Psi_-} & & \downarrow{\scriptstyle \Phi\otimes\mathrm{Id}} \\
\mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}^{\mathrm{ex}} & & \mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}_+^{\mathrm{ex}} \otimes \mathcal{H}^{\mathrm{ex}} \\
\downarrow{\scriptstyle \Phi\otimes\Psi_-} & & \uparrow{\scriptstyle m_-\otimes\mathrm{Id}\otimes\mathrm{Id}} \\
\mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}_+^{\mathrm{ex}} \otimes \mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}^{\mathrm{ex}} & \xrightarrow{\tau_{23}} & \mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}_-^{\mathrm{ex}} \otimes \mathcal{H}_+^{\mathrm{ex}} \otimes \mathcal{H}^{\mathrm{ex}}
\end{array}$$

This yields by dualization an action of the extended renormalization group G_-^{ex} on an *extended regularity structure* $(A^{\mathrm{ex}}, T^{\mathrm{ex}}, G_+^{\mathrm{ex}})$, where T^{ex} is the dual of $\mathcal{H}^{\mathrm{ex}}$, and where $G_\pm^{\mathrm{ex}}$ is the character group of the Hopf algebra $\mathcal{H}_\pm^{\mathrm{ex}}$. In particular, the group G_-^{ex} acts on the space T^{ex} and acts by automorphisms on the extended structure group G_+^{ex} in a compatible way, in the sense that for any $\alpha \in G_-^{\mathrm{ex}}$, $g \in G_+^{\mathrm{ex}}$ and $m \in T^{\mathrm{ex}}$ we have:

$$\alpha.(g.m) = (\alpha.g).(\alpha.m), \tag{19}$$

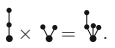
[5] These are not Hopf algebras anymore strictly speaking because, due to the extra $\mathbb{Z}^{d'}$-decoration of the vertices, the coproducts are given by infinite linear combinations. This issue can be handled by working in the symmetric monoidal category of *bi-graded vector spaces* [4, Paragraph 2.3].

see [4, Section 5] for details. The finite-dimensional renormalization group $\mathcal{R}$ is a quotient of the extended renormalization group G_-^{ex} [4, Remark 6.27].

The fact that $\mathcal{H}^{\mathrm{ex}}$ is not an algebra but only a comodule reflects the fact that, contrarily to functions, two irregular distributions cannot in general be multiplied. For a comprehensive account of regularity structures, the reader is referred to [3–5, 8, 23, 24].

9 A Cascade of Coproducts and Coactions on Trees

Inspired by [4], particularly Paragraphs 3.1 to 3.7, we give a rather simple example, due to Y. Bruned,[6] of a doubly infinite family $(\mathcal{T}_{ij})_{1\leq i<j}$ of comodule-bialgebras over the same bialgebra. Let $\mathcal{F}$ (resp. $\mathcal{T}$) be the linear span of rooted forests (resp. rooted trees), without any decoration. The product $\times$ on $\mathcal{T}$ is the *merging product*, i.e. for any two rooted trees t_1 and t_2, the product $t_1 \times t_2$ is obtained by merging the two roots. For example:



The unit element is the single-vertex tree $\bullet$. The usual commutative product $\cdot$ on $\mathcal{F}$ is given by the disjoint union of forests. We use the traditional notation $B_+ : \mathcal{F} \xrightarrow{\sim} \mathcal{T}$ for the grafting of all connected components of the forest on a common root. Its inverse is denoted by $B_- : \mathcal{T} \xrightarrow{\sim} \mathcal{F}$. A *planted tree* is the image of a tree by B_+, i.e. a tree with univalent root. One obviously has for two trees t_1 and t_2:

$$t_1 \times t_2 = B_+\big(B_-(t_1) \cdot B_-(t_2)\big),$$

hence the algebras $(\mathcal{F}, \cdot)$ and $(\mathcal{T}, \times)$ are isomorphic. The first is the free commutative algebra generated by the set of rooted trees, and the second is the free commutative algebra generated by the set of planted trees.

The *height* of a vertex is its distance from the root. The height of a subtree $s \subset t$ is the height of its root inside t. A *covering subforest of t of height i* is a partition s of a subset $\mathcal{E}$ of the set $\mathcal{V}(t)$ of vertices of t such that:

$$\mathcal{V}_i(t) \subset \mathcal{E} \subset \mathcal{V}_{\geq i}(t)$$

into connected blocks, where $\mathcal{V}_i(t)$ (resp. $\mathcal{V}_{\geq i}(t)$) is the set of vertices of t of height i (resp. $\geq i$), such that each subtree thus obtained is of height i. Thus each vertex in $\mathcal{V}_i$ is the root of a connected component of s. By convention, the empty forest $\mathbf{1}$ will be considered as the only covering subforest of height i if $\mathcal{V}_{\geq i}(t) = \emptyset$. Now

[6] Private communication.

we define for any $j \geq 1$ the linear map $\widetilde{B}_+^j : \mathcal{F} \to \mathcal{T}$ as the unique unital algebra morphism such that $\widetilde{B}_+^j|_{\mathcal{T}} = B_+^j|_{\mathcal{T}}$. For example,

$$\widetilde{B}_+^1(\mathbf{1}) = \bullet, \qquad \widetilde{B}_+^2(\bullet\bullet) = \text{  }.$$

by convention we set $\widetilde{B}^0 = \mathrm{Id} : \mathcal{F} \to \mathcal{F}$. We are now ready to define a family of coproducts $\Delta_i : \mathcal{T} \otimes \mathcal{T} \to \mathcal{T}$ for $i \geq 1$:

$$\Delta_i(t) := \sum_{s \subset t,\, h(s) = i-1} \widetilde{B}_+^{i-1}(s) \otimes t/s, \tag{20}$$

where the sum runs over all covering subforests of t of height $i - 1$. For example we have:

$$\Delta_1(\text{}) = \bullet \otimes \text{  } + \text{  } \otimes \text{  } + \text{} \otimes \text{  } + \text{} \otimes \bullet,$$
$$\Delta_2(\text{}) = \text{} \otimes \text{} + \text{} \otimes \text{},$$
$$\Delta_3(\text{}) = \text{} \otimes \text{},$$
$$\Delta_i(\text{}) = \bullet \otimes \text{} \text{ for } i \geq 4.$$

Theorem 2 *The coproducts Δ_i defined by (20) are coassociative and multiplicative with respect to the merging product $\times$.*

Proof This is a straightforward computation:

$$(\Delta_i \otimes \mathrm{Id})\Delta_i(t) = \sum_{\substack{u \subset s \subset t \\ h(u) = h(s) = i-1 \text{ in } t}} \widetilde{B}_+^{i-1}(u) \otimes \widetilde{B}_+^{i-1}(s/u) \otimes t/s.$$

Here we have noticed that a covering subforest $\overline{u}$ of $\widetilde{B}_+^{i-1}(s)$ of height $i - 1$ is obtained from a unique covering subforest u of s of height zero, and we have used the obvious identification of $\widetilde{B}_+^{i-1}(s)/\overline{u}$ with $\widetilde{B}_+^{i-1}(s/u)$. On the other hand,

$$\begin{aligned}(\mathrm{Id} \otimes \Delta_i)\Delta_i(t) &= \sum_{h(u) = i-1 \text{ in } t} \; \sum_{h(\widetilde{s}) = i-1 \text{ in } t/u} \widetilde{B}_+^{i-1}(u) \otimes \widetilde{B}_+^{i-1}(\widetilde{s}) \otimes (t/u)\big/\widetilde{s} \\ &= \sum_{\substack{u \subset s \subset t \\ h(u) = h(s) = i-1 \text{ in } t}} \widetilde{B}_+^{i-1}(u) \otimes \widetilde{B}_+^{i-1}(s/u) \otimes t/s.\end{aligned}$$

For $i \geq 2$, the multiplicativity with respect to $\times$ is a direct consequence of the fact that $\widetilde{B}_+^{i-1}$ is an algebra morphism from $(\mathcal{F}, \cdot)$ to $(\mathcal{T}, \times)$. For $i = 1$, it is a direct consequence of the remark below:

Remark 3 The coproduct Δ_1 is an avatar of the Butcher-Connes-Kreimer coproduct on rooted forests, namely:

$$\Delta_1 = (B_+ \otimes B_+) \circ \Delta_{\rm BCK}^{\rm op} \circ B_-. \tag{21}$$

The verification of (21) is straightforward and left to the reader. The bialgebras $(\mathcal{T}, \times, \Delta_i)$ are not co-unital for $i \geq 2$.

Let us now consider the extraction contraction coproduct $\Gamma : \mathcal{F} \to \mathcal{F} \otimes \mathcal{F}$ defined by:

$$\Gamma(t) := \sum_{s \subset t} s \otimes t/s, \tag{22}$$

which makes $(\mathcal{F}, \cdot, \Gamma)$ a bialgebra [7]. Here the sum runs over all covering subforests of t without indication of the height, i.e. partitions of $\mathcal{V}(t)$ into connected blocks. The co-unit is the unique unital algebra morphism such that $\varepsilon(\bullet) = 1$. Let us consider the maps $\Phi_i : \mathcal{T} \to \mathcal{F} \otimes \mathcal{T}$ defined by:

$$\Phi_i(t) := \sum_{h(s) \geq i-1} s \otimes t/s. \tag{23}$$

Here, the sum runs over *covering subforests of height* $\geq i-1$, i.e. partitions s of the set $\mathcal{V}_{\geq i-1}(t)$ into connected blocks.[7] By convention, the empty forest $\mathbf{1}$ will be considered as the only covering subforest of height $\geq i-1$ if $\mathcal{V}_{\geq i-1}(t) = \emptyset$. One obviously has $\Phi_1 = \Gamma_{|_{\mathcal{T}}}$

Theorem 4 *For any $1 \leq i < j$, the coproduct Δ_i and the map Φ_j defined above make $\mathcal{T}$ a comodule-bialgebra $\mathcal{T}_{ij}$ over the bialgebra $(\mathcal{F}, \cdot, \Gamma)$.*

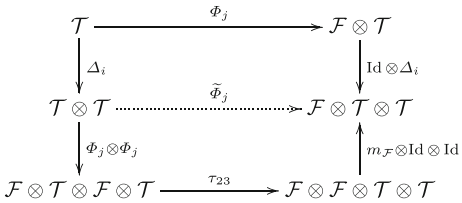
Proof We already know by Theorem 2 that $(\mathcal{T}, \times, \Delta_i)$ is a (not necessarily co-unital) bialgebra. We have now to prove that Φ_j is a coaction, namely:

$$(\mathrm{Id} \otimes \Phi_j) \circ \Phi_j = (\Gamma \otimes \mathrm{Id}) \circ \Phi_j, \tag{24}$$

$$(\varepsilon_{\mathcal{F}} \otimes \mathrm{Id}) \circ \Phi_j = \mathrm{Id}, \tag{25}$$

[7] There is a subtle point here: strictly speaking, a covering subforest of height j with $j > i-1$, which is a partition of a subset of $\mathcal{V}_{\geq j}$, cannot be considered as a covering subforest of height $\geq i-1$, which is a partition of $\mathcal{V}_{\geq i-1}$. Informally speaking, a covering subforest of height $\geq i-1$ covers $\mathcal{V}_{\geq i-1}$ whereas a covering subforest of height j only covers a set $\mathcal{E}$ such that $\mathcal{V}_j \subset \mathcal{E} \subset \mathcal{V}_{\geq j}$, where $\mathcal{V}_j$ stands for the set of height j vertices of t.

and that the comodule-bialgebra property is verified, i.e. the following diagram commutes:



and every map in this diagram is a unital algebra morphism. The verification of (24) is straightforward:

$$\begin{aligned}(I\otimes\Phi_j)\circ\Phi_j(t) &= \sum_{h(s)\geq j-1}(\mathrm{Id}\otimes\Phi_j)(s\otimes t/s)\\ &= \sum_{h(s)\geq j-1}\ \sum_{h(\overline{u})\geq j-1\,\mathrm{in}\,t/s} s\otimes\overline{u}\otimes t/s\Big/\overline{u}\\ &= \sum_{\substack{s\subset u\subset t,\\ h(s)\geq j-1,\ h(u)\geq j-1}} s\otimes u/s\otimes t/u\\ &= \sum_{\substack{s\subset u\subset t,\\ h(s)\geq 0\,\mathrm{in}\,u,\ h(u)\geq j-1\,\mathrm{in}\,t}} s\otimes u/s\otimes t/u\\ &= (\Gamma\otimes\mathrm{Id})\circ\Phi_j(t).\end{aligned}$$

The verification of (25) is left to the reader. The maps in the diagram are unital algebra morphisms. It remains to show the following identity for any $1\leq i<j$:

$$(\mathrm{Id}\otimes\Delta_i)\circ\Phi_j = m^{13}\circ(\Phi_j\otimes\Phi_j)\circ\Delta_i, \tag{26}$$

where m^{13} stands for $(m_{\mathcal{F}}\otimes\mathrm{Id}\otimes\mathrm{Id})\circ\tau_{23}$. This is once again a direct check:

$$\begin{aligned}(\mathrm{Id}\otimes\Delta_i)\circ\Phi_j(t) &= (\mathrm{Id}\otimes\Delta_i)\sum_{h(s)\geq j-1} s\otimes t/s\\ &= \sum_{h(s)\geq j-1,h(\overline{u})=i-1} s\otimes\widetilde{B}_+^{i-1}(\overline{u})\otimes t/s\Big/\overline{u}\\ &= \sum_{\substack{s\subset u\subset t\\ h(u\setminus s)=i-1,\ h(s)\geq j-i\,\mathrm{in}\,u}} s\otimes\widetilde{B}_+^{i-1}(u/s)\otimes t/u,\end{aligned}$$

whereas

$$
\begin{aligned}
m^{13}(\Phi_j \otimes \Phi_j)\Delta_i(t) &= m^{13}(\Phi_j \otimes \Phi_j) \sum_{h(u)=i-1} \widetilde{B}_+^{i-1}(u) \otimes t/u \\
&= \sum_{h(u)=i-1} \sum_{\substack{h(s')\geq j-i \text{ in } u \\ h(s'')\geq j-1 \text{ in } t/u}} s'.\overline{s''} \otimes \widetilde{B}_+^{i-1}(u/s') \otimes t/u \Big/ \overline{s''} \\
&= \sum_{\substack{s'\subset u\subset s''\subset t,\ h(u)=i-1 \\ h(s')\geq j-i \text{ in } u,\ h(s''\setminus u)\geq j-1}} s'.(s''/u) \otimes \widetilde{B}_+^{i-1}(u/s') \otimes t/s'' \\
&= \sum_{\substack{s'\subset u\subset s''\subset t,\ h(u)=i-1 \\ h(s')\geq j-i \text{ in } u,\ h(s''\setminus u)\geq j-1}} s'.(s''\setminus u) \otimes \widetilde{B}_+^{i-1}\Big(u.(s''\setminus u)\Big/ s'.(s''\setminus u)\Big) \otimes t/s'' \\
&= \sum_{\substack{s\subset v\subset t \\ h(s)\geq j-1,\ h(v\setminus s)=i-1}} s \otimes \widetilde{B}_+^{i-1}(v/s) \otimes t/v.
\end{aligned}
$$

The last line is obtained by the change of indices $s'.s'' \setminus u \mapsto s$ and $s'' \mapsto v$. □

Acknowledgements I thank Kurusch Ebrahimi–Fard for his encouragements, as well as Yvain Bruned, Martin Hairer and Lorenzo Zambotti for introducing me to regularity structures. Special thanks to Yvain for illuminating discussions and for providing me the example in Sect. 9.

References

1. Aguiar, M., Mahajan, S.: Monoidal Functors, Species and Hopf Algebras. CRM Monograph Series, vol. 29. American Mathematical Society, Providence (2010)
2. Brouder, Ch.: Runge-Kutta methods and renormalization. Eur. Phys. J. C Part. Fields **12**, 512–534 (2000)
3. Bruned, Y.: Equations singulières de type KPZ, Ph.D. Thesis, Univerity Paris 6, Dec 2015
4. Bruned, Y., Hairer, M., Zambotti, L.: Algebraic renormalisation of regularity structures. arXiv:1610.08468 (2016)
5. Bruned, Y., Chevyrev, I., Friz, P., Preiss, R.: A rough paths perspective on renormalization. arXiv:1701.01152 (2017)
6. Butcher, J.C.: An algebraic theory of integration methods. Math. Comput. **26**, 79–106 (1972)
7. Calaque, D., Ebrahimi-Fard, K., Manchon, D.: Two interacting Hopf algebras of trees. Adv. Appl. Math. **47**(2), 282–308 (2011)
8. Chandra, A., Weber, H.: An analytic BPHZ theorem for regularity structures. arXiv:1612.08138 (2016)
9. Chapoton, F., Livernet, M.: Pre-Lie algebras and the rooted trees operad. Int. Math. Res. Not. **2001**, 395–408 (2001)
10. Chartier, Ph., Hairer, E., Vilmart, G.: Numerical integrators based on modified differential equations. Math. Comput. **76**, 1941–1953 (2007)
11. Connes, A., Kreimer, D.: Hopf algebras, renormalization and noncommutative geometry. Commun. Math. Phys. **199**, 203–242 (1998)
12. Dür, A.: Möbius Functions, Incidence Algebras and Power Series Representations. Lecture Notes in Mathematics, vol. 1202. Springer, Berlin (1986)

13. Ebrahimi-Fard, K., Fauvet, F., Manchon, D.: A comodule-bialgebra structure for word-series substitution and mould composition. J. Algebra **489**, 552–581 (2017)
14. Ecalle, J.: Les fonctions résurgentes, vol. 1. Publications Mathématiques d'Orsay (1981). Available at http://portail.mathdoc.fr/PMO/feuilleter.php?id=PMO_1981
15. Ecalle, J.: Singularités non abordables par la géométrie. Ann. Inst. Fourier **42**(1–2), 73–164 (1992)
16. Ecalle, J., Vallet, B.: Prenormalization, correction, and linearization of resonant vector fields or diffeomorphisms. Prepub. Math. Orsay **95–32**, 90 (1995)
17. Ecalle, J., Vallet, B.: The arborification-coarborification transform: analytic, combinatorial, and algebraic aspects. Ann. Fac. Sci. Toulouse **XIII**(4), 575–657 (2004)
18. Fauvet, F., Foissy, L., Manchon, D.: The Hopf algebra of finite topologies and mould composition. Ann. Inst. Fourier. arXiv:1503.03820 (2015, to appear)
19. Fauvet, F., Foissy, L., Manchon, D.: Operads of finite posets. Electron. J. Comb. **25**(1), 29 pp. (2018)
20. Fauvet, F., Menous, F.: Ecalle's arborification-coarborification transforms and the Connes–Kreimer Hopf algebra. Ann. Sci. Éc. Norm. Sup. **50**(1), 39–83 (2017)
21. Foissy, L.: Les algèbres de Hopf des arbres enracinés décorés I,II. Bull. Sci. Math. **126**, 193–239, 249–288 (2002)
22. Foissy, L.: Algebraic structures associated to operads. arXiv:1702.05344 (2017)
23. Hairer, M.: A theory of regularity structures. Invent. Math. **198**(2), 269–504 (2014)
24. Hairer, M.: Introduction to regularity structures. Braz. J. Probab. Stat. **29**, 175–210 (2015)
25. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration Structure-Preserving Algorithms for Ordinary Differential Equations. Springer Series in Computational Mathematics, vol. 31. Springer, Berlin (2002)
26. Hoffman, M.E.: Quasi-shuffle products. J. Algebraic Combin. **11**, 49–68 (2000)
27. Kassel, Chr.: Quantum Groups. Graduate Texts in Mathematics, vol. 155. Springer, New York (1995)
28. Kreimer, D.: On the Hopf algebra structure of perturbative quantum field theories. Adv. Theor. Math. Phys. **2**, 303–334 (1998)
29. Manchon, D.: On bialgebra and Hopf algebra of oriented graphs. Confluentes Math. **4**(1), 1240003 (10 pp.) (2012)
30. Menous, F.: An example of local analytic q-difference equation: analytic classification. Ann. Fac. Sci. Toulouse **XV**(4), 773–814 (2006)
31. Molnar, R.K.: Semi-direct products of Hopf algebras. J. Algebra **45**, 29–51 (1977)
32. Murua, A.: The Hopf algebra of rooted trees, free Lie algebras, and Lie series. Found. Comput. Math. **6**, 387–426 (2006)
33. Murua, A., Sanz-Serna, J.-M.: Order conditions for numerical integrators obtained by composing simpler integrators. Philos. Trans. R. Soc. Lond. A **357**, 1079–1100 (1999)
34. Murua, A., Sanz-Serna, J.-M.: Word series for dynamical systems and their numerical integrators. arXiv:1502.05528 [math.NA] (2015)
35. Murua, A., Sanz-Serna, J.-M.: Hopf algebra techniques to handle dynamical systems and numerical integrators. arXiv:1702.08354 [math.DS] (2017)

Renormalization: A Quasi-shuffle Approach

Frédéric Menous and Frédéric Patras

Abstract In recent years, the usual BPHZ algorithm for renormalization in perturbative quantum field theory has been interpreted, after dimensional regularization, as a Birkhoff decomposition of characters on the Hopf algebra of Feynman graphs, with values in a Rota-Baxter algebra of amplitudes. We associate in this paper to any such algebra a universal semigroup (different in nature from the Connes-Marcolli "cosmical Galois group"). Its action on the physical amplitudes associated to Feynman graphs produces the expected operations: Bogoliubov's preparation map, extraction of divergences, renormalization. In this process a key role is played by commutative and noncommutative quasi-shuffle bialgebras whose universal properties are instrumental in encoding the renormalization process.

1 Introduction

In the early 2000s, the usual BPHZ algorithm for renormalization in perturbative quantum field theory has been interpreted, after dimensional regularization, as a Birkhoff decomposition of characters on the Hopf algebra of Feynman graphs, with values in a Rota-Baxter algebra of amplitudes [6, 7, 12]. This idea was later shown to be meaningful in a broad variety of contexts: in the theory of dynamical systems, in analysis and numerical analysis (Rayleigh-Schrödinger series) or, more recently, in the theory of regularity structures and the study of very irregular

F. Menous (✉)
Laboratoire de Mathmatiques d'Orsay, University of Paris-Sud, CNRS, Universit Paris-Saclay, Orsay, France
e-mail: Frederic.Menous@u-psud.fr

F. Patras
Laboratoire J.A. Dieudonné, Université de la Côte d'Azur, CNRS, UMR 7531, Nice Cedex 2, France
e-mail: Frederic.PATRAS@unice.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_21

stochastic differential equations or stochastic partial differential equations, see e.g. [4, 21, 28, 29, 31].

In this context, P. Cartier suggested the existence of a hidden universal symmetry group (the "cosmical Galois group") that would underlie renormalization. Using geometrical tools such as universal singular frames, Connes and Marcolli constructed a candidate group in 2004 [8]. Their construction was translated in the langage of Hopf algebras in [13] and the group shown to coincide with the prounipotent group of group-like elements in the completion with respect to the grading of the descent algebra -a Hopf algebra that, as an algebra, is the free associative algebra generated by the Dynkin operators [34].

However, the action of this group or of the descent algebra on the Hopf algebras of Feynman diagrams showing up in pQFT does not actually perform renormalization. It captures nicely certain phenomena related to Lie theory and the behaviour of the Dynkin operators: for example, the structure of certain renormalization group equations and the algebraic properties of beta functions (see the original article by Connes and Marcolli [8] and the detailed algebraic and combinatorial analysis of these phenomena in [35]. Further insights on the role of (generalized) Dynkin operators in the theory of differential equations can be found in [32]). However, the group and the descent algebra act on Feynman diagrams and do not encode operations that occur at the level of the target algebra of amplitudes. They fail therefore to capture typical renormalization operations such as projections on divergent or regular components of amplitudes. Substraction maps, for example, cannot be encoded in it, and neither are more advanced operations such as the construction of the counterterm.

In the present article, we follow a different approach that complements Connes-Marcolli's and its Hopf algebraic and combinatorial interpretation by showing show how a semigroup of operators can be associated to the algebra of coefficients of a given regularization and renormalization scheme in pQFT. Its construction relies heavily on the universal properties of commutative and noncommutative quasi-shuffle algebras. This semigroup acts in a natural way on regularized amplitudes and perform the expected operations: preparation map, extraction of counterterms, renormalization. Notice that many of our results and constructions do not require the algebra of coefficients to be commutative.

Let us sketch up the ideas and results. Concretely we deal with conilpotent bialgebras $H = k \oplus H^+$. These bialgebras are Hopf algebras and the coalgebra structure on H induces a convolution product on the space $\mathscr{L}(H, A)$ of linear morphisms from H to an associative algebra A. If A is unital, then the subset $\mathscr{U}(H, A)$ of linear morphisms that send the unit 1_H of H on the unit 1_A of A is a group for the convolution and, if A is commutative, the subset $\mathscr{C}(H, A)$ of characters (i.e. algebra morphisms) is a subgroup of $\mathscr{U}(H, A)$.

In pQFT, the algebra A is often called the algebra of (regularized) amplitudes, and we will often use this terminology. In this context, the renormalization process equips the target unital algebra A with a projection operator p_+ such that

$$A = \mathrm{Im}\, p_+ \oplus \mathrm{Im}\, p_- = A_+ \oplus A_-,$$

where $p_- = \mathrm{Id} - p_+$ and A_+ and A_- are subalgebras. Here, p_- should be thought of as a projection on the "divergent part", so that p_+ substract divergences. For example, in dimensional regularization, A identifies with the algebra of Laurent series, $\mathbb{C}[[\varepsilon, \varepsilon^{-1}]$, and p_- (resp. p_+) is the projection on $\varepsilon^{-1}\mathbb{C}[\varepsilon^{-1}]$ (resp. $\mathbb{C}[[\varepsilon]]$). As was first observed by Ebrahimi-Fard, building on previous results by Brouder and Kreimer, these data define a Rota-Baxter algebra structure on A and $\mathscr{L}(H, A)$.

The choice of the subtraction operator is not always unique – for example when using momentum subtraction schemes. How this phenomenon impacts the combinatorics and Rota-Baxter structures was investigated in [10]. Although we do not investigate it further here, the tools we develop in the present article should be useful in that context since they put forward the idea that one should study for its own the combinatorial structure of the target algebra of amplitudes A, independently of the choice of a particular subtraction map p_+.

It is then well-know that, given p_+, there exists a unique Birkhoff decomposition of any morphism $\varphi \in \mathscr{U}(H, A)$

$$\varphi_- * \varphi = \varphi_+ \qquad \varphi_+, \varphi_- \in \mathscr{U}(H, A)$$

where $\varphi_+(H^+) \subset A_+$ and $\varphi_-(H^+) \subset A_-$. Moreover, if A is commutative, this decomposition is defined in the subgroup $\mathscr{C}(H, A)$. The classical proofs of this result are recursive, using the filtration on H (they rely ultimately on the Bogoliubov recursion [14]).

We propose to develop here a "universal" framework to handle the combinatorics of renormalization and to give in this framework explicit, and in some sense universal, formulas for φ_+ and φ_-. To do so, we consider the quasi-shuffle Hopf algebra $QSh(A)$ over an algebra A, that is, the standard tensor coalgebra over A equipped with the quasi-shuffle (or stuffle) product. Using the properties of the functor QSh (including the surprising property, for any Hopf algebra H to be canonically embedded into $QSh(H^+)$), we compute then the inverse and the Birkhoff decomposition of a fundamental element $j \in \mathscr{U}(QSh(A), A)$ defined by

$$j(1) = 1_A, \quad j(a_1) = a_1, \quad j(a_1 \otimes \ldots \otimes a_s) = 0 \quad \text{if} \quad s \geq 2.$$

We show then the existence of an action of $\mathscr{U}(QSh(A), A)$ on $\mathscr{U}(H, A)$. More precisely we define a map

$$\mathscr{U}(QSh(A), A) \times \mathscr{U}(H, A) \to \mathscr{U}(H, A)$$

$$(f, \varphi) \mapsto f \odot \varphi,$$

such that

$$j \odot \varphi = \varphi \quad \text{and} \quad (f * g) \odot \varphi = (f \odot \varphi) * (g \odot \varphi),$$

and obtain explicit formulas such as:

1. If j^{*-1} is the inverse of j, then $\varphi^{*-1} = j^{*-1} \odot \varphi$.
2. If $j_- * j = j_+$ (Birkhoff decomposition), then $\varphi_- * \varphi = \varphi_+$ where $\varphi_\pm = j_\pm \odot \varphi$.

The article is organized as follows. After a preliminary section fixing notations and recalling general properties of Hopf algebras, Sect. 3 analyses the algebraic properties of algebras of regularized amplitudes and explains how they give rise to quasi-shuffle algebra structures. Section 4 introduces Hoffman's quasi-shuffle functor (i.e. the notion of quasi-shuffle algebra over an algebra -in the commutative case, it is the left adjoint to the forgetful functor from quasi-shuffle algebras to commutative algebras). Section 5 investigates its categorical properties, including a surprising right adjoint property (Theorem 1). Section 6 studies, using these techniques, the map j (mapping a cofree coalgebra to its cogenerating vector space). This is the key to latter applications to renormalization which are the purpose of Sect. 7, as well as the construction, for each algebra of amplitudes, of a "universal semigroup" in which the operations characteristic of renormalization are encoded. The last two sections survey various applications, in particular to Dynamics and Analysis.

2 Notation and Hopf Algebra Fundamentals

Everywhere in the article, algebraic structures are defined over a fixed ground field k of characteristic 0. We fix here the notations relative to bialgebras and Hopf algebras, following [17] (see also [5, 26] and [37]) and refer to these articles and surveys for details and generalities on the subject. Recall that a bialgebra B is an associative algebra with unit and a coassociative coalgebra with counit such that the product is a morphism of coalgebras (or, equivalently, the coproduct is a morphism of algebras). We will usually write m the product, Δ the coproduct, $u : k \to B$ the unit and $\eta : B \to k$ the counit. When ambiguities might arise we put an index (and denote e.g. m_B the product instead of m).

We use freely the Sweedler notation and write

$$\Delta h = \sum h_{(1)} \otimes h_{(2)}. \tag{1}$$

Thanks to coassociativity, we can define recursively and without any ambiguity the linear morphisms $\Delta^{[n]} : B \to B^{\otimes n}$ $(n \geq 1)$ by $\Delta^{[1]} = \mathrm{Id}$ and, for $n \geq 1$,

$$\Delta^{[n+1]} = (\mathrm{Id} \otimes \Delta^{[n]}) \circ \Delta = (\Delta^{[n]} \otimes \mathrm{Id}) \circ \Delta = (\Delta^{[k]} \otimes \Delta^{[n+1-k]}) \circ \Delta \quad (1 \leq k \leq n) \tag{2}$$

and write

$$\Delta^{[n]} h = \sum h_{(1)} \otimes \ldots \otimes h_{(n)} \tag{3}$$

In the same way, for $n \geq 1$, we define $m^{[n]} : B^{\otimes^n} \to B$ by $m^{[1]} = \mathrm{Id}$ and

$$m^{[n+1]} = m \circ (\mathrm{Id} \otimes m^{[n]}) = m \circ (m^{[n]} \otimes \mathrm{Id}) \tag{4}$$

The reduced coproduct Δ' on $H^+ := Ker\ \eta$ is defined by

$$\Delta' h = \Delta h - 1 \otimes h - h \otimes 1 \tag{5}$$

Its iterates (defined as for Δ) are written $\Delta'^{[n]}$. A bialgebra is conilpotent (or, more precisely, locally conilpotent) is for any $h \in H^+$ there exists a $n \geq 1$ (depending on h) such that $\Delta'^{[n]}(h) = 0$.

A bialgebra H is a Hopf algebra if there exists an antipode S, that is to say a linear map $S : H \to H$ such that:

$$m \circ (\mathrm{Id} \otimes S) \circ \Delta = m \circ (S \otimes \mathrm{Id}) \circ \Delta = u \circ \eta : H \to H \tag{6}$$

In this article, we will consider only conilpotent bialgebras, which are automatically Hopf algebras.

Given a connected bialgebra H and an algebra A with product m_A and unit u_A, the coalgebra structure of H induces an associative convolution product on the vector space $\mathscr{L}(H, A)$ of k–linear maps:

$$\forall (f, g) \in \mathscr{L}(H, A) \times \mathscr{L}(H, A), \quad f * g = m_A \circ (f \otimes g) \circ \Delta \tag{7}$$

with a unit given by $u_A \circ \eta$, such that $(\mathscr{L}(H, A), *, u_A \circ \eta)$ is an associative unital algebra.

Lemma 1 *Let H be a conilpotent bialgebra (and therefore a Hopf algebra) and set*

$$\mathscr{U}(H, A) = \{f \in \mathscr{L}(H, A) \quad ; \quad f(1_H) = 1_A\} \tag{8}$$

then $\mathscr{U}(H, A)$ is a group for the convolution product.

Proof $\mathscr{U}(H, A)$ is obviously stable for the convolution product. Following [17], we will remind why any element $f \in \mathscr{U}(H, A)$ as a unique inverse f^{*-1} in $\mathscr{U}(H, A)$. One can write formally

$$f^{*-1} = (u_A \circ \eta - (u_A \circ \eta - f))^{*-1} = u_A \circ \eta + \sum_{k \geq 1} (u_A \circ \eta - f)^{*k} \tag{9}$$

This series seems to be infinite but, because of the conilpotency assumption, for any $h \in H'$

$$(u_A \circ \eta - f)^{*k}(h) = (-1)^k m_A^{[k]} \circ f^{\otimes k} \circ \Delta'^{[k]}(h) \tag{10}$$

vanishes for k large enough.

When this result is applied to $\mathrm{Id} : H \to H \in \mathscr{U}(H, H)$, then its convolution inverse is the antipode S (this is the usual way of proving that any conilpotent bialgebra is a Hopf algebra).

Notation 1 If $B \subset A$ is a subalgebra of A which is not unital, then we write

$$\mathscr{U}(H, B) = \{f \in \mathscr{L}(H, A) \quad ; \quad f(1_H) = 1_A \quad \text{and} \quad f(H^+) \subset B\}$$

This is a subgroup of $\mathscr{U}(H, A)$.

Let now $\mathscr{C}(H, A)$ be the subset of $\mathscr{L}(H, A)$ whose elements are algebra morphisms (also called characters over A). Of course,

$$\mathscr{C}(H, A) \subset \mathscr{U}(H, A)$$

but this shall not be a subgroup: if A is not commutative, there is no reason why it should be stable for the convolution product. Nonetheless if A is commutative, the product from $A \otimes A$ to A is an algebra map: it follows that the convolution of algebra morphisms is an algebra morphism and $\mathscr{C}(H, A)$ is a subgroup of $\mathscr{U}(H, A)$.

Moreover if $f \in \mathscr{U}(H, A)$ is an algebra map, then its inverse f^{*-1} in $\mathscr{U}(H, A)$ is an antialgebra map given by $f^{*-1} = f \circ S$:

$$\begin{aligned} f * f \circ S &= m_A \circ (f \otimes f \circ S) \circ \Delta \\ &= m_A \circ (f \otimes f) \circ (\mathrm{Id} \otimes S) \circ \Delta \\ &= f \circ m \circ (\mathrm{Id} \otimes S) \circ \Delta \\ &= f \circ u \circ \eta \\ &= u_A \circ \eta, \end{aligned} \tag{11}$$

where we recall that the antipode is an antialgebra morphism:

$$S(gh) = S(h)S(g).$$

3 From Renormalization to Quasi-shuffle Algebras

The fundamental ideas of renormalization in pQFT were already alluded at in the introduction, we recall them very briefly and refer to textbooks for details (this first paragraph is mainly motivational, we will move immediately after to an algebraic framework that can be understood without mastering the quantum field theoretical background). Starting from a given quantum field theory, one expands perturbatively the quantities of interest (such as Green's functions). This expansion is indexed by Feynman diagrams, and to each of these diagrams is associated a quantity computed by means of certain integrals. Very often, these integrals are divergent and need to

be regularized and renormalized. Typically, a quantity such as

$$\phi(c) := \int_0^\infty \frac{dy}{y+c}$$

is divergent, but becomes convergent up to the introduction of an arbitrary small regularizing parameter ε (for dimensional reasons, one also introduces a mass term μ)

$$\phi(c;\varepsilon) := \int_0^\infty \frac{\mu^\varepsilon dy}{(y+c)^{1+\varepsilon}} = \frac{1}{\varepsilon} + \log(\mu/c) + O(\varepsilon).$$

In that toy model case, close to the dimensional regularization method, the "regularized amplitude" $\phi(c;\varepsilon)$ lives in $A = \mathbb{C}[[\varepsilon, \varepsilon^{-1}]$ and is renormalized by removing the divergency $\frac{1}{\varepsilon}$ (the component of the expansion in $\varepsilon^{-1}\mathbb{C}[\varepsilon^{-1}]$).

These ideas are axiomatized using the notion of Rota–Baxter algebras as follows. Following [11], let p_+ an idempotent of $\mathscr{L}(A, A)$ where A is a unital algebra (in our toy model example, p_+ would stand for the projection on $\mathbb{C}[[\varepsilon]]$). If we have for x, y in A:

$$p_+(x)p_+(y) + p_+(xy) = p_+(xp_+(y)) + p_+(p_+(x)y), \tag{12}$$

then p_+ is a Rota-Baxter operator, (A, p_+) is a Rota-Baxter algebra and if $p_- = \mathrm{Id} - p_+$, $A_+ = \mathrm{Im}\, p_+$ and $A_- = \mathrm{Im}\, p_-$ then

- $A = A_+ \oplus A_-$.
- p_- satisfies the same relation.
- A_+ and A_- are subalgebras.

Conversely if $A = A_+ \oplus A_-$ and A_+ and A_- are subalgebras, then the projection p_+ on A_+ parallel to A_- defines a Rota-Baxter algebra (A, p_+).

The idempotency condition is not required to define a Rota–Baxter algebra. In general:

Definition 1 A Rota–Baxter (RB) algebra is an associative algebra A equipped with a linear endomorphism R such that

$$\forall x, y \in A,\ R(x)R(y) = R(R(x)y + xR(y) - xy).$$

It is an idempotent RB algebra if R is idempotent (in that case we will set $p_+ := R$ to emphasize that we are in the framework typical for renormalization). It is a commutative Rota–Baxter algebra if it is commutative as an algebra.

The notion of Rota–Baxter algebra is actually slightly more general: a Rota–Baxter algebra of weight θ is defined by the identity

$$\forall x, y \in A,\ R(x)R(y) = R(R(x)y + xR(y) + \theta xy).$$

We restrict here the definition to the weight -1 case, which is the one meaningful for renormalization.

Using Rota–Baxter algebras of amplitudes, the principle of renormalization in physics can be formulated algebraically in the following way.

Proposition 1 *Let H be a conilpotent bialgebra and (A, p_+) an idempotent Rota-Baxter algebra (so that $A = A_- \oplus A_+$). Then for any $\varphi \in \mathscr{U}(H, A)$ there exists a unique pair $(\varphi_+, \varphi_-) \in \mathscr{U}(H, A_+) \times \mathscr{U}(H, A_-)$ such that*

$$\varphi_- * \varphi = \varphi_+ \tag{13}$$

Moreover, if A is commutative and φ is a character, then φ_+ and φ_- are also characters. This factorization is called the Birkhoff decomposition of φ.

Proof Let us postpone the assertion on characters and prove the existence and unicity -notions such as the one of Bogoliubov's preparation map will be useful later. As A_+ and A_- are subalgebras of A, $\mathscr{U}(H, A_+)$ and $\mathscr{U}(H, A_-)$ are subgroups of $\mathscr{U}(H, A)$. If such a factorization exists, then it is unique : If $\varphi = \varphi_-^{*-1} * \varphi_+ = \psi_-^{*-1} * \psi_+$, then

$$\phi = \psi_+ * \varphi_+^{*-1} = \psi_- * \varphi_-^{*-1} \in \mathscr{U}(H, A_+) \cap \mathscr{U}(H, A_-)$$

thus for $h \in H^+$, $\phi(h) \in A_+ \cap A_- = 0$. We finally get that

$$\psi_+ * \varphi_+^{*-1} = \psi_- * \varphi_-^{*-1} = u_A \circ \eta$$

and $\varphi_+ = \psi_+$, $\varphi_- = \psi_-$.

Let us prove now that the factorization exists. Let $\varphi \in \mathscr{U}(H, A)$, we must have $\varphi_+(1_H) = \varphi_-(1_H) = 1_A$. Let $\bar{\varphi} \in \mathscr{U}(H, A)$ the Bogoliubov preparation map defined recursively on the increasing sequence of vector spaces $H_n^+ := Ker\, \Delta'^{[n]}$ $(n \geq 1)$ by

$$\bar{\varphi}(h) = \varphi(h) - m_A \circ (p_- \otimes \mathrm{Id}) \circ (\bar{\varphi} \otimes \varphi) \circ \Delta'(h) \tag{14}$$

(since H is conilpotent, $H^+ = \cup_n H_n^+$). Now if φ_+ and φ_- are the elements of $\mathscr{U}(H, A)$ defined on H^+ by

$$\varphi_+(h) = p_+ \circ \bar{\varphi}(h) \quad , \quad \varphi_-(h) = -p_- \circ \bar{\varphi}(h) \quad (\bar{\varphi}(h) = \varphi_+(h) - \varphi_-(h)),$$

then

$$\varphi_+ \in \mathscr{U}(H, A_+) \quad , \quad \varphi_- \in \mathscr{U}(H, A_-) \quad , \quad \varphi_- * \varphi = \varphi_+$$

We turn now to another algebraic structure, induced by the one of RB algebras, but weaker – the one we will be concerned later on: quasi-shuffle algebras.

Concretely, the target algebras of amplitudes (such as the algebra of Laurent series) happen to be quasi-shuffle algebras, whereas the algebras of linear forms on Feynman diagrams with values in a commutative RB algebra of amplitudes happen to be noncommutative quasi-shuffle algebras.

Indeed, a RB algebra is always equipped with an associative product, the RB double product $\star$, defined by:

$$x \star y := R(x)y + xR(y) - xy \tag{15}$$

so that: $R(x)R(y) = R(x \star y)$. Setting $x \prec y := xR(y),\ x \succ y := R(x)y$, one gets

$$(xy) \prec z = xyR(z) = x(y \prec z),$$

$$(x \prec y) \prec z = xR(y)R(z) = x \prec (y \star z),$$

$$(x \succ y) \prec z = R(x)yR(z) = x \succ (y \prec z),$$

and so on. These observations give rise to the axioms of noncommutative quasi-shuffle algebras (NQSh, also called tridendriform, algebras). On an historical note, we learned recently from K. Ebrahimi-Fard that the following axioms and relations seem to have first appeared in the context of stochastic calculus, namely in the work of Karandikar in the early 1980s on matrix semimartingales, see e.g. [24]. See also [18] for details and other references.

Definition 2 A noncommutative quasi-shuffle algebra (NQSh algebra) is a nonunital associative algebra (with product written $\bullet$) equipped with two other products $\prec, \succ$ such that, for all $x, y, z \in A$:

$$(x \prec y) \prec z = x \prec (y \star z),\ (x \succ y) \prec z = x \succ (y \prec z) \tag{16}$$

$$(x \star y) \succ z = x \succ (y \succ z),\ (x \prec y) \bullet z = x \bullet (y \succ z) \tag{17}$$

$$(x \succ y) \bullet z = x \succ (y \bullet z),\ (x \bullet y) \prec z = x \bullet (y \prec z). \tag{18}$$

where $x \star y := x \prec y + x \succ y + x \bullet y$.

Notice that $(x \bullet y) \bullet z = x \bullet (y \bullet z)$ and (16) + (17) + (18) imply the associativity of $\star$:

$$(x \star y) \star z = x \star (y \star z). \tag{19}$$

When the RB algebra is commutative, the relations between the three products $\prec, \succ, \bullet$ simplify (since $x \prec y = xR(y) = y \succ x$) and one arrives at the definition:

Definition 3 A quasi-shuffle (QSh) algebra A is a nonunital commutative algebra (with product written $\bullet$) equipped with another product $\prec$ such that

$$(x \prec y) \prec z = x \prec (y \star z) \tag{20}$$

$$(x \bullet y) \prec z = x \bullet (y \prec z). \tag{21}$$

where $x \star y := x \prec y + y \prec x + x \bullet y$.

We also set for further use $x \succ y := y \prec x$ (this makes a QSh algebra a NQSh algebra). The product $\star$ is automatically associative and commutative and defines another commutative algebra structure on A.

It is sometimes convenient to equip NQSh and QSh algebras with a unit. The phenomenon is exactly similar to the case of shuffle algebras [36]. Given a NQSh algebra, one sets $B := k \oplus A$, and the products $\prec, \succ, \bullet$ have a partial extension to B defined by, for $x \in A$:

$$1 \bullet x = x \bullet 1 := 0,\ 1 \prec x := 0,\ x \prec 1 := x,\ 1 \succ x := x,\ x \succ 1 := 0.$$

The products $1 \prec 1$, $1 \succ 1$ and $1 \bullet 1$ cannot be defined consistently, but one sets $1 \star 1 := 1$, making B a unital commutative algebra for $\star$. The categories of NQSh/QSh and unital NQSh/QSh algebras are equivalent (under the operation of adding or removing a copy of the ground field).

Formally, the relations between RB algebras and NQSh algebras are encoded by the Lemma:

Lemma 2 *The identities $x \prec y := xR(y),\ x \succ y := R(x)y, x \bullet y := xy$ induce a forgetful functor from RB algebras to NQSh algebras, resp. from commutative RB algebras to QSh algebras.*

We already alluded to the fact that, in a given quantum field theory, the set of linear forms from the linear span of Feynman diagrams (or equivalently algebra maps from the polynomial algebra they generate) to a commutative RB algebra of amplitudes carries naturally the structure of a noncommutative RB algebra. In the context of QSh algebras, this result generalizes as follows:

Proposition 2 *Let C be a (coassociative) coalgebra with coproduct Δ and A be a NQSh algebra. Then the set of linear maps $Hom(C, A)$ is naturally equipped with the structure of a NQSh algebra by the products:*

$$f \prec g(c) := f(c^{(1)}) \prec g(c^{(2)}),$$

$$f \succ g(c) := f(c^{(1)}) \succ g(c^{(2)}),$$

$$f \bullet g(c) := f(c^{(1)}) \bullet g(c^{(2)}),$$

where we used Sweedler's notation $\Delta(c) = c^{(1)} \otimes c^{(2)}$.

The proposition follows from the fact that the relations defining NQSh algebras are non-symmetric (in the sense that they do not involve permutations: for example, in the equation $(x \prec y) \prec z = x \prec (y \star z)$, the letters x, y, z appear in the same order in the left and right hand side, and similarly for the other defining relations).

4 The Quasi-shuffle Hopf Algebra $QSh(A)$

For details on the constructions in this section, we refer the reader to [18, 22, 23]. Let A be an associative algebra. We write $QSh(A)$ for the graded vector space $QSh(A) = \bigoplus_{n\geq 0} QSh(A)_n = k \oplus \bigoplus_{n\geq 1} QSh(A)_n =: k \oplus QSh^+(A)$ where, for $n \geq 1$, $QSh(A)_n = A^{\otimes n}$ and $QSh(A)_0 = k$ (notice that when A is unital, one has to distinguish between $1 \in k = QSh(A)_0$ and $1_A \in A \subset QSh(A)_1$). We denote $l(\boldsymbol{a}) = n$ the length of an element $\boldsymbol{a}$ of $QSh(A)_n$.

For convenience, an element $\boldsymbol{a} = a_1 \otimes \ldots \otimes a_n$ of $QSh(A)$ will be called a word and will be written $a_1 \ldots a_n$ (it should not be confused with the product of the a_i in A). We will reserve the tensor product notation for the tensor product of elements of $QSh(A)$ (so that for example, $a_1a_2 \otimes a_3 \in QSh(A)_2 \otimes QSh(A)_1$). Also, we distinguish between the concatenation product of words (written $\cdot$: $a_1a_2a_3 \cdot b_1b_2 = a_1a_2a_3b_1b_2$) and the product in A by writing $a \cdot_A b$ the product of a and b in A (whereas $a \cdot b$ would stand for the word ab of length 2).

The graded vector space $QSh^+(A)$ (resp. $QSh(A)$) is given a graded (resp. unital) NQSh algebra structure by induction on the length of tensors such that for all $a, b \in A$, for all $v, w \in QSh(A)$:

$$av \prec bw = a(v \star bw),$$

$$av \succ bw = b(av \star w),$$

$$av \bullet bw = (a._A b)(v \star w),$$

where $\sqcup\!\sqcup := \star = \prec + \succ + \bullet$ is usually called the quasi-shuffle (or stuffle) product (by definition: $\forall v \in QSh(A), 1 \sqcup\!\sqcup v = v = v \sqcup\!\sqcup 1$). Notice that this product $\sqcup\!\sqcup$ can be defined directly by the two equivalent inductions

$$av \sqcup\!\sqcup bw := a(v \sqcup\!\sqcup bw) + b(av \sqcup\!\sqcup w) + a \cdot_A b(v \sqcup\!\sqcup w)$$

or

$$va \sqcup\!\sqcup wb := (v \sqcup\!\sqcup bw)a + (av \sqcup\!\sqcup w)b + (v \sqcup\!\sqcup w)a \cdot_A b.$$

When A is commutative, $QSh(A)$ is a unital quasi-shuffle algebra
For example:

$$a_1a_2 \sqcup\!\sqcup b = a_1a_2b + a_1ba_2 + ba_1a_2 + a_1(a_2 \cdot_A b) + (a_1 \cdot_A b)a_2 \tag{22}$$

Notice at last that, under the action of the four products $\prec, \succ, \star, \bullet$, the image of $QSh(A)_r \otimes QSh(A)_s$ is contained in $\bigoplus_{t=\max(r,s)}^{r+s} QSh(A)_t$

One can also define:

- a counit $\eta : QSh(A) \to k$ by $\eta(1) := 1$ and for $s \geq 1$, $\eta(a_1 \dots a_s) = 0$,
- a coproduct (called deconcatenation coproduct) $\Delta : QSh(A) \to QSh(A) \otimes QSh(A)$ such that $\Delta(1) = 1 \otimes 1$ and for $s \geq 1$ and $\boldsymbol{a} = a_1 \dots a_s \in QSh(A)_s$,

$$\Delta(\boldsymbol{a}) = \boldsymbol{a} \otimes 1 + 1 \otimes \boldsymbol{a} + \sum_{r=1}^{s-1} (a_1 \dots a_r) \otimes (a_{r+1} \dots a_s) \tag{23}$$

making $QSh(A)$ a graded coalgebra. It is a matter of fact to check that $QSh(A)$ is a unital conilpotent bialgebra (and thus a Hopf algebra, see e.g. [5]), which is called the quasi-shuffle or stuffle Hopf algebra on A (this terminology, that we adopt, is convenient, usual, but slightly misleading because when A is only associative, $QSh(A)$ is a unital noncommutative quasi-shuffle algebra).

5 Operations and Universal Properties

Let us focus now in the first part of this section on the case relevant to renormalization, that is when A is commutative but not necessarily unital. It follows then from standard arguments in universal algebra that, given a quasi-shuffle algebra B, morphisms of quasi-shuffle algebras from $QSh^+(A)$ to B are naturally in bijection with morphisms of (non unital) algebras from A to B:

$$Hom_{QSh}(QSh^+(A), B) \cong Hom_{Alg}(A, B).$$

In categorical terms (see [18] for a direct and elementary proof):

Proposition 3 (Quasi-shuffle PBW theorem) *The left adjoint U of the forgetful functor from the category of quasi-shuffle algebras QSh to the category of non unital commutative algebras Com, or "quasi-shuffle enveloping algebra" functor from Com to QSh, is Hoffman's quasi-shuffle algebra functor $A \longmapsto QSh^+(A)$.*

It is interesting to analyse the concrete meaning of this Proposition. Let us consider first the counit of the adjunction, that is the quasi-shuffle algebra map from $QSh^+(A)$ to A, when A is a quasi-shuffle algebra. By definition of $\prec$, the element $a_1 \dots a_n \in QSh(A)_n$ can be rewritten (in $QSh(A)$) $a_1 \prec (a_2 \prec \dots (a_{n-1} \prec a_n))$. The trick goes back to Schützenberger who used it in his seminal but not enough acknowledged study of shuffle algebras [36]. It follows that the counit of the adjunction maps $a_1 \dots a_n \in QSh(A)_n$ to $a_1 \prec (a_2 \prec \dots (a_{n-1} \prec a_n))$ (computed now in A).

Let us move now to the case when A is a commutative RB algebra. Then, A is in particular a quasi-shuffle algebra with $a \prec b := aR(b)$. The counit of the same adjunction is then the map that sends $a_1 \dots a_n \in QSh(A)_n$ to $a_1 R(a_2 R(a_3 \dots R(a_{n-1} R(a_n)))$. In particular, a^n is mapped to $aR(aR(a \dots R(aR(a)))$ -a term that is known to play a key role in renormalization, see in particular [14].

This relatively standard adjunction analysis can be completed in the case we are interested in (maps from $QSh^+(A)$ to B, when B is a quasi-shuffle algebra), due to the existence of a Hopf algebra structure on $QSh(A)$. According to Proposition 2, we have first that

Lemma 3 *Let A be an associative algebra and B a NQSh algebra, the vector space of linear morphisms $\mathscr{L}(QSh(A), B)$ is a NQSh algebra.*

Furthermore, by properties that hold for arbitrary maps from a conilpotent Hopf algebra to an algebra, if B is unital, the set of linear maps that map the unit of $QSh(A)$ to the unit of B, $\mathscr{U}(QSh(A), B)$ is a group for the product $\star$. Moreover, when B is commutative, the subset of algebra maps from $QSh(A)$ to B, $\mathscr{C}(QSh(A), B)$, is a subgroup.

Next, notice that the functor QSh is compatible with Hopf algebra structures: an algebra map l from A to B induces a map $QSh(l)$ of quasi-shuffle algebras from $QSh(A)$ to $QSh(B)$ defined by

$$QSh(l)(1) = 1 \quad \text{and} \quad QSh(l)(a_1 \dots a_r) = l(a_1) \dots l(a_r) \quad (r \geq 1)$$

and therefore $\Delta \circ QSh(l) = (QSh(l) \otimes QSh(l)) \circ \Delta$. In particular, $QSh(l)$ is a Hopf algebra morphism.

The last universal property of the QSh functor that we would like to emphasize is more intriguing and does not seem to have been noticed before. Whereas QSh is naturally a left adjoint, it also happens indeed to be a right adjoint, a property that will prove essential in our later developments.

Theorem 1 *Let H be a conilpotent Hopf algebra and A be a unital associative algebra, then we have a natural isomorphism between (unital) algebra maps from H to A and Hopf algebra maps from H to $QSh(A)$:*

$$Hom_{Alg}(H, A) \cong Hom_{Hopf}(H, QSh(A)).$$

Indeed, $QSh(A)$ is, as a coalgebra, the cofree coalgebra over A (viewed as a vector space) in the category of conilpotent coalgebras. These properties are dual to the ones of tensor algebras (more familiar, but equivalent up to the fact that the dual of a coalgebra is an algebra but the converse is not always true -this is the reason for the conilpotency hypothesis): the tensor algebra over a vector space V is, when equipped with the concatenation product, the free associative algebra over V. There is therefore a natural isomorphism between linear maps from the kernel C^+ of the counit of a coaugmented conilpotent coalgebra C to A and coalgebra maps from C

to $QSh(A)$

$$\mathscr{L}(C^+, A) \cong Hom_{Coalg}(C, QSh(A)).$$

Coaugmented means that there is a coalgebra map from the ground field to C, insuring that C decomposes as the direct sum of k and of the kernel of the counit (as happens for a Hopf algebra, for which the composition of the unit and the counit is a projection on the ground field orthogonally to the kernel of the counit).

The isomorphism is given explicitly as follows: it maps $\phi \in \mathscr{L}(C^+, A)$ to $\tilde{\phi} := \sum_{i=0}^{\infty} \phi^{\otimes n} \circ \Delta'^{[n]}$ (where $\phi^{\otimes 0} \circ \Delta'^{[0]}$ stands for the composition of the counit of C with the unit of $QSh(A)$). In particular, the map ϕ factorizes as (the restriction to C^+ of) $j \circ \tilde{\phi}$, where $j \in \mathscr{L}(QSh(A), A)$ is defined by $j(1) = 1_A$, $j(a_1) = a_1$ and $j(a_1 \dots a_r) = 0$ if $r \geq 2$.

To prove the Theorem, it is therefore enough to show that, when a linear map ϕ from H^+ to A is the restriction to H^+ of an algebra map from H to A, then the induced map $\tilde{\phi}$ is also an algebra map (since we already know it is a coalgebra map). Concretely, we have to prove that, for $h, h' \in H^+$, $\tilde{\phi}(hh') = \tilde{\phi}(h) \text{ш} \tilde{\phi}(h')$. The Theorem will then follow if we prove that

$$\sum_{n=1}^{\infty} \phi^{\otimes n} \circ \Delta'^{[n]}(hh') = \sum_{p=1}^{\infty} \phi^{\otimes p} \circ \Delta'^{[p]}(h) \text{ш} \sum_{q=1}^{\infty} \phi^{\otimes p} \circ \Delta'^{[q]}(h').$$

Using that ϕ and that Δ are algebra maps, this follows from the following Lemma (where, to avoid ambiguities, we use the notation $\Delta'^{[p]}(h) = h'_{(1,p)} \otimes \cdots \otimes h'_{(p,p)}$) by identification of the terms in the left and right hand side.

Lemma 4 *We have, for the iterated coproduct and $h \in H^+$,*

$$\Delta^{[n]}(h) = \sum_{i=1}^{n} \sum_{f \in Inj(i,n)} f_*(h'_{(1,i)} \otimes \cdots \otimes h'_{(i,i)}),$$

where $Inj(i, n)$ stands for the set of increasing injections from $[i] := \{1, \dots, i\}$ to $[n]$ and

$$f_*(h'_{(1,i)} \otimes \cdots \otimes h'_{(i,i)}) = l_{(1)} \otimes \cdots \otimes l_{(n)}$$

with $l_{(q)} := h'_{(p,i)}$ if $q = f(p)$ and $l_{(q)} := 1$ if q is not in the image of f.

For example, $\Delta^{[1]}(h) = \Delta'^{[1]}(h) = h = h'_{(1,1)}$,

$$\Delta^{[2]}(h) = \Delta(h) = h'_{(1,1)} \otimes 1 + 1 \otimes h'_{(1,1)} + h'_{(1,2)} \otimes h'_{(2,2)}$$

and

$$\begin{aligned}\Delta^{[2]}(hk) &= \Delta^{[2]}(h)\Delta^{[2]}(k)\\ &= (h'_{(1,1)} \otimes 1 + 1 \otimes h'_{(1,1)} + h'_{(1,2)} \otimes h'_{(2,2)})\\ &\quad \times (k'_{(1,1)} \otimes 1 + 1 \otimes k'_{(1,1)} + k'_{(1,2)} \otimes k'_{(2,2)}),\end{aligned}$$

so that

$$\Delta'^{[2]}(hk) = h'_{(1,1)} \otimes k'_{(1,1)} + k'_{(1,1)} \otimes h'_{(1,1)} + h'_{(1,1)}k'_{(1,2)} \otimes k'_{(2,2)} + k'_{(1,2)} \otimes h'_{(1,1)}k'_{(2,2)}$$

$$+ h'_{(1,2)}k'_{(1,1)} \otimes h'_{(2,2)} + h'_{(1,2)} \otimes h'_{(2,2)}k'_{(1,1)} + h'_{(1,2)}k'_{(1,2)} \otimes h'_{(2,2)}k'_{(2,2)},$$

where one recognizes the tensor degree 2 component of

$$(\Delta'^{[1]}(h) + \Delta'^{[2]}(h)) \text{ш} (\Delta'^{[1]}(k) + \Delta'^{[2]}(k)).$$

The Theorem has an important corollary, that we state also as a Theorem in view of its importance for our approach to renormalization.

Theorem 2 *Let H be a conilpotent bialgebra, then, the unit, written ι, of the adjunction in the previous Theorem, ($\iota(1) := 1$ and $\forall h \in H^+, \iota(h) = \sum_{k\geq 1} \Delta'^{[k]}(h)$) defines an injective Hopf algebra morphism from H to $QSh(H^+)$. In particular, any conilpotent (resp. conilpotent commutative) Hopf algebra embeds into a noncommutative quasi-shuffle (resp. a quasi-shuffle) Hopf algebra.*

We let the reader check the following Lemma, that will be important later in the article and makes Theorem 1 more precise:

Lemma 5 *The map $j \in \mathscr{L}(QSh(A), A)$ is a morphism of algebras.*

6 The Map $j \in \mathscr{U}(QSh(A), A)$

We shall now illustrate the ideas of the previous section on the map $j \in \mathscr{U}(QSh(A), A)$ (recall it is defined by $j(1) = 1_A$, $j(a_1) = a_1$ and $j(a_1 \dots a_r) = 0$ if $r \geq 2$). In a sense, this will be the only computation of inverse and of Birkhoff decomposition we will need. This map j plays a fundamental role. We already saw that it appears in the adjunction $\mathscr{L}(C^+, A) \cong Hom_{Coalg}(C, QSh(A))$. It will also appear later to be the unit of a semigroup structure on $\mathscr{U}(QSh(A), A)$ to be introduced in the next section.

For the inverse, we get j^{*-1}:

$$j^{*-1} = u_A \circ \eta + \sum_{k\geq 1} (u_A \circ \eta - j)^{*k}$$

Which means that $j^{*-1}(1) = 1_A$ and for $\boldsymbol{a} = a_1 \dots a_s \in QSh(A)^+$,

$$\begin{aligned} j^{*-1}(\boldsymbol{a}) &= \sum_{k\geq 1}(-1)^k m_A^{[k]} \circ j^{\otimes k} \circ \Delta'^{[k]}(\boldsymbol{a}) \\ &= \sum_{k\geq 1}(-1)^k \sum_{\substack{\boldsymbol{a}^1\dots\boldsymbol{a}^k=\boldsymbol{a} \\ \boldsymbol{a}^i\in QSh(A)^+}} m_A^{[k]} \circ j^{\otimes k}(\boldsymbol{a}^1 \otimes \dots \otimes \boldsymbol{a}^k) \\ &= (-1)^s m_A^{[s]}(a_1 \otimes \dots \otimes a_s) \\ &= (-1)^s a_1 \cdot_A \dots \cdot_A a_s = j \circ S(\boldsymbol{a}) \end{aligned} \tag{24}$$

where

$$\begin{aligned} S(\boldsymbol{a}) &= \sum_{k\geq 1}(-1)^k m^{[k]} \circ \Delta'^{[k]}(\boldsymbol{a}) \\ &= \sum_{k\geq 1}(-1)^k \sum_{\substack{\boldsymbol{a}^1\dots\boldsymbol{a}^k=\boldsymbol{a} \\ \boldsymbol{a}^i\in QSh(A)^+}} \boldsymbol{a}^1 \sqcup\!\sqcup \dots \sqcup\!\sqcup \boldsymbol{a}^k. \end{aligned} \tag{25}$$

Note that the previous sums run over all the possible factorizations in nonempty words of $\boldsymbol{a}$ for the concatenation product.

If (A, p_+) is a Rota-Baxter algebra then the Bogoliubov preparation map $\bar{j}$ associated to j, see Eq. (14), is such that $\bar{j}(1) = 1_A$ and can be defined recursively on vector spaces $QSh(A)_n$ $(n \geq 1)$ by

$$\bar{j}(h) = j(h) - m_A \circ (p_- \otimes \mathrm{Id}) \circ (\bar{j} \otimes j) \circ \Delta'(h) \tag{26}$$

Let us begin the recursion on the length of the sequence. If $\boldsymbol{a} = a_1$ then $\bar{j}(a_1) = j(a_1) = a_1$. Now, if $\boldsymbol{a} = a_1 \cdot a_2 = a_1 a_2$,

$$\bar{j}(a_1a_2) = j(a_1a_2) - m_A \circ (p_- \otimes \mathrm{Id}) \circ (\bar{j} \otimes j)((a_1) \otimes (a_2)) = -p_-(a_1) \cdot_A a_2 \tag{27}$$

and

$$\begin{aligned} \bar{j}(a_1a_2a_3) &= -m_A \circ (p_- \otimes \mathrm{Id}) \circ (\bar{j} \otimes j)((a_1a_2) \otimes (a_3)) \\ &= p_-(p_-(a_1) \cdot_A a_2) \cdot_A a_3 \end{aligned} \tag{28}$$

Thus, for $r \geq 2$,

$$\bar{j}(a_1 \dots a_r) = -p_-(\bar{j}(a_1 \dots a_{r-1})) \cdot_A a_r \tag{29}$$

It is then easy to prove that in general (see e.g. [14] for a systematic study of combinatorial approaches and closed solutions to the Bogoliubov recursion)

Proposition 4 *The Birkhoff decomposition*

$$(j_+, j_-) \in \mathscr{U}(QSh(A), A_+) \times \mathscr{U}(QSh(A), A_-)$$

such that

$$j_- * j = j_+$$

is given by the formula: for $r \geq 1$ *and* $\boldsymbol{a} = a_1 \otimes \ldots \otimes a_r \in QSh(A)^+$,

$$\begin{cases} j_+(\boldsymbol{a}) = p_+(\bar{j}(\boldsymbol{a})) & = (-1)^{r-1} p_+(p_-(\ldots(p_-(a_1) \cdot_A a_2) \ldots \cdot_A a_{r-1}) \cdot_A a_r) \\ j_-(\boldsymbol{a}) = -p_-(\bar{j}(\boldsymbol{a})) & = (-1)^r p_-(p_-(\ldots(p_-(a_1) \cdot_A a_2) \ldots \cdot_A a_{r-1}) \cdot_A a_r) \end{cases} \tag{30}$$

Moreover, if A *is commutative then* $\mathscr{C}(QSh(A), A)$ *is a group and* j_+ *and* j_- *are characters.*

Proof Let us prove the last assumption, when A is commutative. Since j is a character it is sufficient to prove that j_- is a character. By induction on $t \geq 0$ we will show that for two tensors $\boldsymbol{a}$ and $\boldsymbol{b}$ in $QSh(A)$, if $l(\boldsymbol{a}) + l(\boldsymbol{b}) = t$, then

$$j_-(\boldsymbol{a} \sqcup\!\sqcup \boldsymbol{b}) = j_-(\boldsymbol{a}) j_-(\boldsymbol{b}) \tag{31}$$

This identity is trivial for $t = 0$ and $t = 1$ since at least one of the sequences is the empty sequence. This also trivial for any t if one of the sequences is empty. Now suppose that $t \geq 2$ and that $\boldsymbol{a} = a_1 \ldots a_r \in QSh(A)_r$ and $\boldsymbol{b} = b_1 \ldots b_s \in QSh(A)_s$ with $r \geq 1$, $s \geq 1$ and $r + s = t$. Let $\tilde{\boldsymbol{a}} = a_1 \ldots a_{r-1} \in QSh(A)_{r-1}$ ($\tilde{\boldsymbol{a}} = 1$ if $r = 1$) and $\tilde{\boldsymbol{b}} = b_1 \ldots b_{s-1} \in QSh(A)_{s-1}$ ($\tilde{\boldsymbol{b}} = 1$ if $s = 1$), then:

$$\boldsymbol{a} \sqcup\!\sqcup \boldsymbol{b} = (\tilde{\boldsymbol{a}} \sqcup\!\sqcup \boldsymbol{b}) \cdot a_r + (\boldsymbol{a} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot b_s + (\tilde{\boldsymbol{a}} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot (a_r \cdot_A b_s)$$

Now we have

$$j_-(\boldsymbol{a}) = -p_-(j_-(\tilde{\boldsymbol{a}}) \cdot_A a_r) = -p_-(x) \quad \text{and} \quad j_-(\boldsymbol{b}) = -p_-(j_-(\tilde{\boldsymbol{b}}) \cdot_A b_s) = -p_-(y),$$

where $x := j_-(\tilde{\boldsymbol{a}}) \cdot_A a_r$ and $y := j_-(\tilde{\boldsymbol{b}}) \cdot_A b_s$. Thanks to the Rota-Baxter identity, and omitting $\cdot_A$ in the following computations in A,

$$\begin{aligned} j_-(\boldsymbol{a}) j_-(\boldsymbol{b}) &= p_-(x) p_-(y) \\ &= p_-(x p_-(y)) + p_-(p_-(x) y) - p_-(xy) \\ &= p_-(j_-(\tilde{\boldsymbol{a}}) a_r \, p_-(j_-(\tilde{\boldsymbol{b}}) b_s)) + p_-(p_-(j_-(\tilde{\boldsymbol{a}}) a_r) j_-(\tilde{\boldsymbol{b}}) b_s) \\ &\qquad - p_-(j_-(\tilde{\boldsymbol{a}}) a_r \, j_-(\tilde{\boldsymbol{b}}) b_s) \end{aligned}$$

but as A is commutative, by induction we get

$$\begin{aligned} j_-(\boldsymbol{a})j_-(\boldsymbol{b}) &= -p_-(j_-(\tilde{\boldsymbol{a}})j_-(\boldsymbol{b})a_r) - p_-(j_-(\boldsymbol{a})j_-(\tilde{\boldsymbol{b}})b_s) - p_-(j_-(\tilde{\boldsymbol{a}})j_-(\tilde{\boldsymbol{b}})a_r b_s) \\ &= -p_-(j_-(\tilde{\boldsymbol{a}} \sqcup\!\sqcup \boldsymbol{b})a_r) - p_-(j_-(\boldsymbol{a} \sqcup\!\sqcup \tilde{\boldsymbol{b}})b_s) - p_-(j_-(\tilde{\boldsymbol{a}} \sqcup\!\sqcup \tilde{\boldsymbol{b}})a_r b_s) \\ &= j_-((\tilde{\boldsymbol{a}} \sqcup\!\sqcup \boldsymbol{b}) \cdot a_r) + j_-((\boldsymbol{a} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot b_s) + j_-((\tilde{\boldsymbol{a}} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot (a_r b_s)) \\ &= j_-((\tilde{\boldsymbol{a}} \sqcup\!\sqcup \boldsymbol{b}) \cdot a_r + (\boldsymbol{a} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot b_s + (\tilde{\boldsymbol{a}} \sqcup\!\sqcup \tilde{\boldsymbol{b}}) \cdot (a_r b_s)) \\ &= j_-(\boldsymbol{a} \sqcup\!\sqcup \boldsymbol{b}) \end{aligned}$$

In the sequel, when there is no ambiguity, we shall omit the notation $\cdot_A$ when applying formula (30).

As we will see these formulas are almost sufficient to compute the Birkhoff decomposition for any conilpotent bialgebra.

7 The Universal Semigroup and Renormalization

Let A be a unital algebra. Then, by adjunction we know that

$$\mathscr{U}(QSh(A), A) \cong Hom_{Coalg}(QSh(A), QSh(A)).$$

In particular, the composition of coalgebra endomorphisms of $QSh(A)$ equips $\mathscr{U}(QSh(A), A)$ with a semigroup structure.

Definition 4 The universal semigroup associated to a unital algebra A is the set $\mathscr{U}(QSh(A), A)$ equipped with the associative unital product induced by composition of coalgebra endomorphisms of $QSh(A)$: for f and g in $\mathscr{U}(QSh(A), A)$

$$f \odot g := f \circ QSh(g) \circ \iota.$$

Its unit is the map j:

$$f \odot j = f \circ QSh(j) \circ \iota = f \circ Id = f.$$

This semigroup structure generalizes to an action on linear maps from a Hopf algebra to A as follows.

Definition 5 Let H be a conilpotent bialgebra. For $\varphi \in \mathscr{U}(H, A)$ and $f \in \mathscr{U}(QSh(A), A)$ we set

$$f \odot \varphi := f \circ QSh(\varphi) \circ \iota.$$

This morphism $f \odot \varphi$ is linear from H to A and unital:

$$f \odot \varphi(1_H) = f \circ QSh(\varphi) \circ \iota(1_H) = f \circ QSh(\varphi)(1) = f(1) = 1_A.$$

We get a left action of $\mathscr{U}(QSh(A), A)$ on $\mathscr{U}(H, A)$:

$$\odot : \mathscr{U}(QSh(A), A) \times \mathscr{U}(H, A) \to \mathscr{U}(H, A).$$

Moreover, when A is commutative, if $\varphi \in \mathscr{C}(H, A)$ and $f \in \mathscr{C}(QSh(A), A)$ it is clear, by composition of algebra morphisms, that $f \odot \varphi \in \mathscr{C}(H, A)$.

That j acts as the identity map on $\mathscr{U}(H, A)$ follows from: for $h \in H^+$,

$$\begin{aligned} j \odot \varphi(h) &= j \circ QSh(\varphi) \left(h + \sum_{k \geq 2} \sum h'_{(1)} \otimes \ldots \otimes h'_{(k)} \right) \\ &= \quad j \left(\varphi(h) + \sum_{k \geq 2} \varphi(h'_{(1)}) \cdot \ldots \cdot \varphi(h'_{(k)}) \right) \\ &= \qquad \varphi(h) \end{aligned} \tag{32}$$

Proposition 5 *The action $\odot$ and the convolution product $*$ (recall that $QSh(A)$ is a Hopf algebra) satisfy the distributivity relation: For f and g in $\mathscr{U}(QSh(A), A)$ and φ in $\mathscr{U}(H, A)$,*

$$(f * g) \odot \varphi = (f \odot \varphi) * (g \odot \varphi).$$

Indeed,

$$\begin{aligned} (f * g) \odot \varphi &= m_A \circ (f \otimes g) \circ \Delta \circ QSh(\varphi) \circ \iota \\ &= m_A \circ (f \otimes g) \circ (QSh(\varphi) \otimes QSh(\varphi)) \circ \Delta \circ \iota \\ &= m_A \circ (f \otimes g) \circ (QSh(\varphi) \otimes QSh(\varphi)) \circ (\iota \otimes \iota) \circ \Delta \\ &= m_A (f \odot \varphi \otimes g \odot \varphi) \circ \Delta \\ &= (f \odot \varphi) * (g \odot \varphi) \end{aligned} \tag{33}$$

Note that, in the case $H = QSh(A)$, $\mathscr{U}(QSh(A), A)$ is equipped with two products $*$ and $\odot$ that look similar, in their interactions, to the product and composition of power series.

Remark 1 These constructions generalize as follows. Let B be another unital algebra. For $\varphi \in \mathscr{U}(H, A)$ and $f \in \mathscr{U}(QSh(A), B)$ we define

$$f \odot \varphi = f \circ QSh(\varphi) \circ \iota.$$

The morphism $f \odot \varphi$ is linear from H to B and

$$f \odot \varphi(1_H) = f \circ QSh(\varphi) \circ \iota(1_H) = f \circ QSh(\varphi)(1) = f(1) = 1_B.$$

thus $f \odot \varphi \in \mathscr{U}(H, B)$. Moreover, when A and B are commutative, if $\varphi \in \mathscr{C}(H, A)$ and $f \in \mathscr{C}(QSh(A), B)$ it is clear, by composition of algebra morphisms that $f \odot \varphi \in \mathscr{C}(H, B)$.

Corollary 1 *Let $\varphi \in \mathscr{U}(H, A)$, then its convolution inverse if given by*

$$\varphi^{*-1} = j^{*-1} \odot \varphi.$$

Indeed, since $j \odot \varphi = \varphi$, if $\psi := j^{*-1} \odot \varphi$, then

$$\psi * \varphi = (j^{*-1} \odot \varphi) * (j \odot \varphi) = (j^{*-1} * j) \odot \varphi = (u_A \circ \eta) \odot \varphi = u_A \circ \eta.$$

For example, if $h \in H^+$ with $\Delta'^{[4]}(h) = 0$, then

$$\iota(h) = h + \sum h'_{(1)} \otimes h'_{(2)} + \sum h'_{(1)} \otimes h'_{(2)} \otimes h'_{(3)}$$

so,

$$QSh(\varphi) \circ \iota(h) = \varphi(h) + \sum \varphi(h'_{(1)}) \cdot \varphi(h'_{(2)}) + \sum \varphi(h'_{(1)}) \cdot \varphi(h'_{(2)}) \cdot \varphi(h'_{(3)})$$

and finally

$$\varphi^{*-1}(h) = j^{*-1} \circ QSh(\varphi) \circ \iota(h) = -\varphi(h) + \sum \varphi(h'_{(1)}) \varphi(h'_{(2)}) - \sum \varphi(h'_{(1)}) \varphi(h'_{(2)}) \varphi(h'_{(3)})$$

We recover the usual formula for the inverse.

Theorem 3 *Assume now that A is an idempotent Rota–Baxter algebra. Let $\varphi \in \mathscr{U}(H, A)$. Then the Birkhoff-Rota-Baxter decomposition of φ is given by*

$$\varphi_- = j_- \odot \varphi, \quad \varphi_+ = j_+ \odot \varphi.$$

Proof Indeed, since $j \odot \varphi = \varphi$, we have

$$\varphi_- * \varphi = (j_- \odot \varphi) * (j \odot \varphi) = (j_- * j) \odot \varphi = j_+ \odot \varphi = \varphi_+$$

and, of course, $\varphi_\pm \in \mathscr{U}(H, A_\pm)$.

For example, if $h \in H'$ with $\Delta'^{[4]}(h) = 0$, then

$$\varphi_+(h) = \; p_+(\varphi(h)) - \sum p_+(p_-(\varphi(h'_{(1)}))\varphi(h'_{(2)})) + \sum p_+(p_-(p_-(\varphi(h'_{(1)}))\varphi(h'_{(2)}))\varphi(h'_{(3)}))$$

$$\varphi_-(h) = -p_-(\varphi(h)) + \sum p_-(p_-(\varphi(h'_{(1)}))\varphi(h'_{(2)})) - \sum p_-(p_-(p_-(\varphi(h'_{(1)}))\varphi(h'_{(2)}))\varphi(h'_{(3)}))$$

Needless to say that if A is commutative, these computations works in the subgroup $\mathscr{C}(H, A)$.

Once these formulas are given, we get formulas in the different contexts where renormalization, or rather Birkhoff decomposition, is needed. We end this paper with two sections that illustrate how these formulas could be used:

- to perform inversion and Birkhoff decomposition of diffeomorphisms that correspond to characters on the Fa di Bruno Hopf algebra,
- to perform the Birkhoff decomposition with the same formula in various cofree Hopf algebras that differ by their algebra structures, but for which the map ι is the same as these Hopf algebras are tensor coalgebras.

8 Renormalizing Diffeomorphisms in pQFT and Dynamics

Let us focus in this section on the example of the Fa di Bruno Hopf algebra $\mathscr{H}_{\mathrm{FdB}}$ (see [3, 17, 19, 28]) whose group of characters corresponds to the group of formal identity-tangent diffeomorphisms. We will first express the reduced coproduct and then the map ι from this Hopf algebra to its associated quasi-shuffle Hopf algebra and then focus on the Birkhoff decomposition of characters with values in the Laurent series that appear in several areas, as a factorisation of diffeomorphisms for the composition.

Recall that the decomposition is unique: the same results could be obtained by induction using the classical renormalization process (the Bogoliubov recursion). One advantage of the present approach is to encode the combinatorics of renormalization into a universal framework, probably similar to the one P. Cartier suggested when advocating the existence of a "Galois group" underlying renormalization. Compare in particular our approach with [6, 12, 14].

Consider the group of formal identity tangent diffeomorphisms with coefficients in a commutative $\mathbb{C}$–algebra A:

$$G(A) = \{f(x) = x + \sum_{n\geq 1} f_n x^{n+1} \in A[[x]]\}$$

with its product $\mu : G(A) \times G(A) \to G(A)$:

$$\mu(f, g) = f \circ g.$$

For $n \geq 0$, the functionals on $G(A)$ defined by

$$a_n(f) = \frac{1}{(n+1)!}(\partial_x^{n+1} f)(0) = f_n \quad a_n : G(A) \to A$$

are called de Fa di Bruno coordinates on the group $G(A)$ and $a_0 = 1$ being the unit, they generates a graded unital commutative algebra

$$\mathscr{H}_{\mathrm{FdB}} = \mathbb{C}[a_1, \ldots, a_n, \ldots] \quad (\mathrm{gr}(a_n) = n)$$

The action of these functionals on a product in $G(A)$ defines a coproduct on $\mathscr{H}_{\mathrm{FdB}}$ that turns to be a graded connected Hopf algebra (see [17] for details). For $n \geq 0$, the coproduct is defined by

$$a_n \circ \mu = m \circ \Delta(a_n) \tag{34}$$

where m is the usual product in A, and the antipode reads

$$S \circ a_n = a_n \circ \mathrm{inv}$$

where $\mathrm{inv}(\varphi) = \varphi^{\circ -1}$ is the composition inverse of φ.

For example if $f(x) = x + \sum_{n\geq 1} f_n x^{n+1}$ and $g(x) = x + \sum_{n\geq 1} g_n x^{n+1}$ then if $h(x) = f \circ g(x) = x + \sum_{n\geq 1} h_n x^{n+1}$,

$$\begin{array}{lll} a_0(h) = 1 = a_0(f)a_0(g) & \to \Delta a_0 = & a_0 \otimes a_0 \\ a_1(h) = f_1 + h_1 & \to \Delta a_1 = & a_1 \otimes a_0 + a_0 \otimes a_1 \\ a_2(h) = f_2 + 2f_1g_1 + g_2 & \to \Delta a_2 = & a_2 \otimes a_0 + 2a_1 \otimes a_1 + a_0 \otimes a_2. \end{array}$$

More generally, using classical formulas on the composition of diffeomorphisms (see [3, 9, 19, 30]), we have

$$\Delta(a_n) = \sum_{k=0}^{n} \sum_{\substack{l_0+\ldots l_k = n-k \\ l_i \geq 0}} a_k \otimes a_{l_0} \ldots a_{l_k} \tag{35}$$

Let us consider sequences of positive integers

$$\mathscr{N} = \{\boldsymbol{n} = (n_1, \ldots, n_s) \in (\mathbb{N}^*)^s, \quad s \geq 1\}$$

For $\boldsymbol{n} = (n_1, \ldots, n_s) \in \mathscr{N}$, we denote

$$\|\boldsymbol{n}\| = n_1 + \ldots + n_s, \quad l(\boldsymbol{n}) = s, \quad a_{\boldsymbol{n}} = a_{n_1} \ldots a_{n_s}$$

and, if $n \geq 1$,

$$\mathscr{N}_n = \{\boldsymbol{n} \in \mathscr{N} \ ; \quad \|\boldsymbol{n}\| = n\}$$

With these notations, the reduced coproduct (with $a_0 = 1$) reads

$$\Delta'(a_n) = \sum_{k=1}^{n-1} \sum_{\boldsymbol{n} \in \mathscr{N}_{n-k}} \binom{k+1}{l(\boldsymbol{n})} a_k \otimes a_{\boldsymbol{n}} \tag{36}$$

and when iterating the coproduct, we get,

Proposition 6 *For $n \geq 1$,*

$$\iota(a_n) = \sum_{\boldsymbol{n} \in \mathcal{N}_n} \sum_{\substack{\boldsymbol{n}^1 \ldots \boldsymbol{n}^t = \boldsymbol{n} \\ t \geq 1, l(\boldsymbol{n}^1)=1}} \lambda(\boldsymbol{n}^1, \ldots, \boldsymbol{n}^t) a_{\boldsymbol{n}^1} \otimes \cdots \otimes a_{\boldsymbol{n}^t} \tag{37}$$

where the sums run over all the decompositions in non empty sequences $\boldsymbol{n}^1 \ldots \boldsymbol{n}^t = \boldsymbol{n}$ and

$$\lambda(\boldsymbol{n}^1, \ldots, \boldsymbol{n}^t) = \prod_{i=2}^{t} \binom{\|\boldsymbol{n}^1 \ldots \boldsymbol{n}^{i-1}\| + 1}{l(\boldsymbol{n}^i)}$$

Note that we kept in formula (37) the tensor product notation to avoid confusion since we deal with words whose letters are monomials. The proof is simply based on the recursive definition of reduced iterated coproduct and already provides a formula for the composition inverse of a diffeomorphism in $G(A)$.

Corollary 2 *Let $f(x) = x + \sum_{n \geq 1} f_n x^{n+1} \in G(A)$, we can consider its associated character defined by $\varphi(a_n) = f_n$ and then, using our previous formulas, the coefficients of the composition inverse g of f are given by*

$$g_n = \varphi^{*-1}(a_n) = \sum_{\boldsymbol{n} = (n_1, \ldots, n_s) \in \mathcal{N}_n} \left(\sum_{\substack{\boldsymbol{n}^1 \ldots \boldsymbol{n}^t = \boldsymbol{n} \\ t \geq 1, l(\boldsymbol{n}^1)=1}} (-1)^t \lambda(\boldsymbol{n}^1, \ldots, \boldsymbol{n}^t) \right) f_{n_1} \ldots f_{n_s}$$

This result, as the following one, uses the obvious isomorphism between $G(A)$ and $\mathscr{C}(\mathscr{H}_{\mathrm{FdB}}, A)$. One can also compute the Birkhoff decomposition in the group of formal identity-tangent diffeomorphism with coefficients in the a Rota-Baxter algebra of Laurent series $A = \mathbb{C}[[\varepsilon, \varepsilon^{-1}]$ with its usual projections p_+ and p_- on the regular and polar parts of such series. Any element

$$f(x) = x + \sum_{n \geq 1} f_n(\varepsilon) x^{n+1} \quad , \quad f_n(\varepsilon) \in \mathbb{C}[[\varepsilon, \varepsilon^{-1}]$$

can be decomposed as $f_- \circ f = f_+$ with

$$f_-(x) = x + \sum_{n \geq 1} f_{-,n}(\varepsilon) x^{n+1} \quad f_{-,n}(\varepsilon) \in \varepsilon^{-1}\mathbb{C}[\varepsilon^{-1}]$$
$$f_+(x) = x + \sum_{n \geq 1} f_{+,n}(\varepsilon) x^{n+1} \qquad f_{+,n}(\varepsilon) \in \mathbb{C}[[\varepsilon]].$$

Using Proposition 6, we get for $n \geq 1$,

Proposition 7 *The coefficients of the Birkhoff decomposition of a formal identity-tangent diffeomorphism are given by*

$$\begin{aligned}
\varphi_+(a_n) &= \sum_{\boldsymbol{n}\in\mathcal{N}_n} \sum_{\substack{\boldsymbol{n}^1\ldots\boldsymbol{n}^t=\boldsymbol{n}\\ t\geq 1, l(\boldsymbol{n}^1)=1}} \lambda(\boldsymbol{n}^1,\ldots,\boldsymbol{n}^t)(-1)^{t-1} p_+(p_-(\ldots(p_-(\varphi(a_{\boldsymbol{n}^1}))\varphi(a_{\boldsymbol{n}^2}))\ldots)\varphi(a_{\boldsymbol{n}^t}))\\
\varphi_-(a_n) &= \sum_{\boldsymbol{n}\in\mathcal{N}_n} \sum_{\substack{\boldsymbol{n}^1\ldots\boldsymbol{n}^t=\boldsymbol{n}\\ t\geq 1, l(\boldsymbol{n}^1)=1}} \lambda(\boldsymbol{n}^1,\ldots,\boldsymbol{n}^t)(-1)^{t} p_-(p_-(\ldots(p_-(\varphi(a_{\boldsymbol{n}^1}))\varphi(a_{\boldsymbol{n}^2}))\ldots)\varphi(a_{\boldsymbol{n}^t}))
\end{aligned} \tag{38}$$

where φ, φ_+ and φ_- are the characters associated to f, f_+ and f_- ($\varphi(a_n) = f_n$).

Let us explain how such diffeomorphisms appear in various area, where there Birkhoff decomposition makes sense.

Such a factorization appears first classically in quantum field theory: after dimensional regularization, the unrenormalized effective coupling constants are the image by a formal identity-tangent diffeomorphism of the coupling constants of the theory (see [7, 9] for a Hopf algebraic approach). Moreover, the coefficients of this diffeomorphism are Laurent series in the parameter ε associated to the dimensional regularization process and the Birkhoff decomposition of this diffeomorphism gives directly the bare coupling constants and the renormalized coupling constants.

As proved in [7], in the case of the massless ϕ_6^3 theory, the unrenormalized effective coupling constant can be written as a diffeomorphism $f(x) = x + \sum_{n\geq 1} f_n(\varepsilon)x^{n+1}$ where x is the initial coupling constant. From the physical point of view, the decomposition $f_- \circ f = f_+$ is such that, $x + \sum_{n\geq 1} f_{+,n}(0)x^{n+1}$ is the renormalized effective constant of the theory.

Such diffeomorphisms (and the need for renormalization) also appear in the classification of dynamical systems, especially when dealing with dynamical systems that cannot be analytically of formally linearized. Let us illustrate this on a very simple example (see [31] for a general approach). The following autonomous analytic dynamical system

$$\begin{cases} \dot{x} = \alpha x \\ \dot{z} = \beta z + b(x)z^2 \end{cases}$$

can be considered as a perturbation of the linear system

$$\begin{cases} \dot{x} = \alpha x \\ \dot{y} = \beta y \end{cases}$$

so that one could expect that a change of coordinate $(x, y) = \psi(x, z) = \big(x, f(x, z)\big)$ allows to go from one system to the other one, that is to linearize the first system. In this simple case (see [31] for details) the solution should be

$f(x,z)=\frac{z}{1-a(x)z}$ where

$$\alpha x a'(x)+\beta a(x)+b(x)=0$$

that yields formally, if $b(x)=\sum_{n\geq 0} b_n x^n$,

$$a(x)=-\sum_{n\geq 0}\frac{b_n}{\alpha n+\beta}x^n.$$

This series could be ill-defined whenever there exists n_0 such that $\alpha n_0+\beta=0$. This happens for example with $n=0$ for $(\alpha,\beta)=(1,0)$ and, in this case, we could regularize by considering the system with linear part $(\alpha,\beta)=(1+\varepsilon,\varepsilon)$. As a function of z, $f(x,z)$ is then an identity-tangent diffeomorphism whose coefficients are in $\mathbb{C}[[x]][[\varepsilon,\varepsilon^{-1}]$:

$$f(x,z)=\frac{z}{1-a(x)z}=z+\sum_{n\geq 1}a(x)^n z^{n+1},\quad a(x)=-\frac{b(0)}{\varepsilon}-\sum_{n\geq 1}\frac{b_n}{n(1+\varepsilon)+\varepsilon}x^n.$$

This very simple case can be handled directly and, after Birkhoff decomposition, the regular part in ε is

$$f_+(x,z)=\frac{z}{1-a_+(x)z}=z+\sum_{n\geq 1}a_+(x)^n z^{n+1},\quad a_+(x)=-\sum_{n\geq 1}\frac{b_n}{n(1+\varepsilon)+\varepsilon}x^n$$

and, for $\varepsilon=0$, the corresponding change of coordinate conjugates the system

$$\begin{cases}\dot{x}=x\\ \dot{z}=b(x)z^2\end{cases}$$

to

$$\begin{cases}\dot{x}=x\\ \dot{y}=b(0)y^2\end{cases}.$$

This approach can be generalized to more general systems for which the Birkhoff decomposition is not so obvious, so that formula (38) could be useful. For instance, the same process of regularization/factorization allows to conjugate the system

$$\begin{cases}\dot{x}=x\\ \dot{z}=\sum_{k\geq 1}b_k(x)z^{k+1}\end{cases}$$

to a system

$$\begin{cases} \dot{x} = x \\ \dot{y} = \sum_{k\geq 1} c_k y^{k+1} \end{cases}$$

which is called a "normal form", with coefficients c_k that do not depend any more on x.

Diffeomorphisms in higher dimension (and thus the corresponding Hopf algebra) appear as well in physics (with more than one coupling constant) and in dynamics: let us consider vector fields given by ν series $\boldsymbol{u}(\boldsymbol{x}) = (u_1(\boldsymbol{x}), \ldots, u_\nu(\boldsymbol{x})) \in \mathbb{C}_{\geq 2}\{\boldsymbol{x}\}$ of ν variables $\boldsymbol{x} = (x_1, \ldots, x_\nu)$ that can be seen as "perturbations" of linear vector fields $(\lambda_1 x_1, \ldots, \lambda_\nu x_\nu)$:

$$\frac{dx_i}{dt} = \lambda_i x_i + u_i(\boldsymbol{x}) = X_i(\boldsymbol{x}), \quad i = 1, \ldots, \nu. \tag{39}$$

The linearization problem consists in finding an identity-tangent diffeomorphism φ in dimension ν such that the change of coordinates $\boldsymbol{x} = \varphi(\boldsymbol{y})$ transforms the previous object into its linear part. For differential equations, this reads, for $i = 1, \ldots, \nu$:

$$\frac{dx_i}{dt} = \sum_{j=1}^{\nu} \frac{dy_j}{dt} \frac{\partial \varphi_i}{\partial y_j}(\boldsymbol{y}) = \sum_{j=1}^{\nu} \lambda_j y_j \frac{\partial \varphi_i}{\partial y_j}(\boldsymbol{y}) = \lambda_i \varphi_i(\boldsymbol{y}) + u_i(\varphi(\boldsymbol{y})) = \lambda_i x_i + u_i(\boldsymbol{x}). \tag{40}$$

When trying to solve these so-called "homological equations", some obstructions can occur, independently on any assumption on the analycity of φ. These equations cannot be formally systematically solved when some combinations $m_1\lambda_1 + \ldots m_\nu \lambda_\nu - \lambda_i$ vanish (here $i \in \{1, \ldots, \nu\}$, $m_j \geq 0$, $\sum m_j \geqslant 2$):

Such cancellations, which are called *resonances*, prevent from linearizing the differential and one can once again use regularization of the linear part and Birkhoff decomposition to get a change of coordinate that conjugate the vector field to a so-called normal form, see [31].

9 Tensor Coalgebras, MZVs, Analysis

If X be an alphabet (that is a set), its associated tensor vector space $T(X)$ inherits a coalgebra structure related to the concatenation. If we note tensors products as words $\boldsymbol{x} = x_1 \otimes \cdots \otimes x_s = x_1 \ldots x_s$,

$$\Delta(\boldsymbol{x}) = 1 \otimes \boldsymbol{x} + \sum_{\boldsymbol{x}^1 \boldsymbol{x}^2 = \boldsymbol{x}} \boldsymbol{x}^1 \otimes \boldsymbol{x}^2 + \boldsymbol{x} \otimes 1$$

where the central sum, that corresponds to the reduced coproduct, is over nonempty words $\boldsymbol{x}^1, \boldsymbol{x}^2$ whose concatenation is $\boldsymbol{x}$.

The quasi-shuffle Hopf algebras $QSh(A)$ are examples of such coalgebras (choose simply a linear basis X of A!). There are however many Hopf algebras with such a coalgebra structure that differ as algebras – but the associated map ι and the associated formula for the Birkhoff decomposition of characters, does not depend on the algebra structure. For the map ι, we obviously get:

$$\iota(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{x}^1\boldsymbol{x}^2\dots\boldsymbol{x}^t=\boldsymbol{x} \\ t\geq 1\ ;\ \boldsymbol{x}^i\neq\emptyset}} \boldsymbol{x}^1 \otimes \boldsymbol{x}^2 \otimes \cdots \otimes \boldsymbol{x}^t \tag{41}$$

and if φ is a character from a Hopf algebra with such a coalgebra structure, with values in a commutative Rota-Baxter algebra (A, p_+), the factorization $\varphi_- * \varphi = \varphi_+$ is given for any $\boldsymbol{x} \in T(X)$ by

$$\begin{aligned}
\varphi_+(\boldsymbol{x}) &= \sum_{\substack{\boldsymbol{x}^1\boldsymbol{x}^2\dots\boldsymbol{x}^t=\boldsymbol{x} \\ t\geq 1\ ;\ \boldsymbol{x}^i\neq\emptyset}} (-1)^{t-1} p_+(p_-(\dots(p_-(\varphi(\boldsymbol{x}^1))\varphi(\boldsymbol{x}^2))\dots)\varphi(\boldsymbol{x}^t)) \\
\varphi_-(\boldsymbol{x}) &= \sum_{\substack{\boldsymbol{x}^1\boldsymbol{x}^2\dots\boldsymbol{x}^t=\boldsymbol{x} \\ t\geq 1\ ;\ \boldsymbol{x}^i\neq\emptyset}} (-1)^{t} p_-(p_-(\dots(p_-(\varphi(\boldsymbol{x}^1))\varphi(\boldsymbol{x}^2))\dots)\varphi(\boldsymbol{x}^t))
\end{aligned} \tag{42}$$

Let us list some example where this formula appear or can be used.

Example 1 (Renormalization af Multiple Zeta Values (MZV)) In [20, Section 3] Guo and Zhang consider regularized MZV as characters on a quasi-shuffle algebra $\mathscr{H}_{\mathfrak{M}} = T(\mathfrak{M})$ whose quasi-shuffle product stems from the additive semigroup structure of the alphabet

$$\mathfrak{M} = \left\{ \left[\begin{smallmatrix} s \\ r \end{smallmatrix}\right] ;\ (s,r) \in \mathbb{Z} \times \mathbb{R}^{+*} \right\}.$$

They propose then a directional regularization of MZV defined on words

$$Z(\left[\begin{smallmatrix} s_1 \\ r_1 \end{smallmatrix}\right] \dots \left[\begin{smallmatrix} s_k \\ r_k \end{smallmatrix}\right]; \varepsilon) = \sum_{n_1>\dots>n_k>0} \frac{e^{n_1 r_1 \varepsilon} \dots e^{n_k r_k \varepsilon}}{n_1^{s_1} \dots n_k^{s_k}}$$

that defines a character on $\mathscr{H}_{\mathfrak{M}}$ with values in an algebra of Laurent series. The formula they give for the Birkhoff decomposition (Theorem 3.8 in [20]) coincide Eq. (42).

Example 2 (Rooted ladders) As a toy model for applications in physics [9, section 4.2] considers a character on the polynomial commutative Hopf algebra $\mathscr{H}^{\mathrm{lad}}$ of

ladder trees. If the ladder tree with n nodes is t_n, then

$$\Delta(t_n) = t_n \otimes 1 + \sum_{k=1}^{n-1} t_k \otimes t_{n-k} + 1 \otimes t_n.$$

It is a matter of fact to identify the coalgebra structure of $\mathscr{H}^{\mathrm{lad}}$ with the tensor deconcatenation coalgebra $T(\{x\})$ over an alphabet with one letter, where t_n corresponds to the word $\underbrace{x \dots x}_{n}$. Formula (42) can be applied to the character mapping the tree t_n to an n-fold Chen's iterated integral defined recursively by

$$\psi(p;\varepsilon,\mu)(t_n) = \mu^{\varepsilon} \int_p^{\infty} \psi(x;\varepsilon,\mu)(t_{n-1}) \frac{dx}{x^{1+\varepsilon}} = \frac{e^{-n\varepsilon \log(p/\mu)}}{n!\varepsilon^n} = f_n(\varepsilon)$$

with values in the Laurent series in ε. We get for the couterterms:

$$\psi_-(p;\varepsilon,\mu)(t_n) = \sum_{\substack{n_1+\dots+n_t=n \\ t\geq 1\ ,\ n_i>0}} (-1)^t (-1)^t p_-(p_-(\dots(p_-(f_{n_1}(\varepsilon))f_{n_2}(\varepsilon))\dots)f_{n_t}(\varepsilon)) \tag{43}$$

Example 3 (Differential equations) When dealing with differential equations and associated diffeomorphisms (flow, conjugacy map), characters on shuffle Hopf algebras appear almost naturally. For instance, such characters correspond to:

- the coefficients of word series in [33],
- "symmetral moulds" in mould calculus (see [15, 16])
- or Chen's iterated integrals (see for instance [25, 27]).

Let us just give the example of a simple differential equation related to mould calculus (see [29]). Let $b(x,y) = \sum_{n\geq 0} x^n b_n(y) \in y^2\mathbb{C}[[x,y]]$ and $d \in \mathbb{N}$. If one looks for a formal identity tangent diffeomorphism $\varphi(x,y)$ in y, with coefficients in $\mathbb{C}[[x]]$ such that, if y is a solution of

$$(E_{b,d}) \qquad x^{1-d}\partial_x y = b(x,y)$$

then $z = \varphi(x,y)$ is a solution of

$$(E_{0,d}) \qquad x^{1-d}\partial_x z = 0.$$

One can try to compute this diffeomorphism as a "mould series":

$$\varphi_d(x,y) = y + \sum_{s\geq 1} \sum_{n_1,\dots n_s\in\mathbb{N}} V_d(n_1,\dots,n_s)\mathbb{B}_{n_s}\dots\mathbb{B}_{n_1}.y \quad (\mathbb{B}_n = b_n(y)\partial_y) \tag{44}$$

where V_d is a character on the shuffle algebra $T(\mathbb{N})$, with values in $\mathbb{C}[[x]]$. Whenever d is a positive integer, this character can be computed and for any word $(n_1, \ldots, n_s)$

$$V_d(n_1, \ldots, n_s) = \frac{(-1)^s x^{n_1+\ldots+n_s+sd}}{(\check{n}_1 + d)(\check{n}_2 + 2d)\ldots(\check{n}_s + sd)} \quad (\check{n}_i = n_1+\ldots+n_i). \tag{45}$$

The map $\varphi_d(x, y) \in \mathbb{C}[[x, y]]$ is then well defined and conjugates $(E_{b,d})$ to $(E_{0,d})$. For $d = 0$, there may be divisions by 0 and, in this case, one can consider $d = \varepsilon$ as a real parameter and use the expansion $x^{\varepsilon} = \sum \frac{(\varepsilon \log x))^n}{n!}$ so that the character V_ε has its values in $\mathfrak{B}[[\varepsilon]][\varepsilon^{-1}]$ where $\mathfrak{B} = \mathbb{C}[[\log x, x]$. If one uses the same formula (42) to perform the Birkhoff decomposition, the regular character $V_{\varepsilon,+}$, evaluated at $\varepsilon = 0$ allows to find a diffeomorphism (as in Eq. (44)) that conjugates $x\partial_x y = b(x, y)$ to $x\partial_x z = 0$ with a price to pay: it contains monomials in x and $\log x$. See [29] for details.

Not also that the same ideas can be used for the the even-odd factorization of characters in combinatorial Hopf algebras (see [1, 2] and [12]).

Acknowledgements We acknowledge support from the CARMA grant ANR-12-BS01-0017, "Combinatoire Algébrique, Résurgence, Moules et Applications"and the CNRS GDR "Renormalisation". We thank warmly K. Ebrahimi-Fard, from whom we learned some years ago already the meaningfulness of Rota–Baxter algebras and their links with quasi–shuffle algebras.

References

1. Aguiar, M., Hsiao, S.K.: Canonical characters on quasi-symmetric functions and bivariate Catalan numbers. Electron. J. Combin. **11**(2) (2004/06). Research Paper 15, 34 pp. (electronic)
2. Aguiar, M., Bergeron, N., Sottile, F.: Combinatorial Hopf algebras and generalized Dehn-Sommerville relations. Compos. Math. **142**(1), 1–30 (2006)
3. Brouder, C., Frabetti, A., Krattenthaler, C.: Non-commutative Hopf algebra of formal diffeomorphisms. Adv. Math. **200**(2), 479–524 (2006)
4. Bruned, Y., Hairer, M., Zambotti, L.: Algebraic renormalisation of regularity structures. arXiv preprint arXiv:1610.08468 (2016)
5. Cartier, P.: A primer of Hopf algebras. In: Cartier, P.E., Julia, B., Moussa, P., Vanhove, P. (eds.) Frontiers in Number Theory, Physics, and Geometry II, pp. 537–615. Springer, Berlin/Heidelberg (2017)
6. Connes, A., Kreimer, D.: Renormalization in quantum field theory and the Riemann-Hilbert problem. I: The Hopf algebra structure of graphs and the main theorem. Commun. Math. Phys. **210**(1), 249–273 (2000)
7. Connes, A., Kreimer, D.: Renormalization in quantum field theory and the Riemann-Hilbert problem. II: The β-function, diffeomorphisms and the renormalization group. Commun. Math. Phys. **216**(1), 215–241 (2001)
8. Connes, A., Marcolli, M.: From physics to number theory via noncommutative geometry. In: Cartier, P.E., Julia, B., Moussa, P., Vanhove, P. (eds.) Frontiers in Number Theory, Physics, and Geometry. I, pp. 269–347. Springer, Berlin (2006)
9. Ebrahimi-Fard, K., Patras, F.: Exponential Renormalization Annales Henri Poincar **11**(5), 943–971 (2010)

10. Ebrahimi-Fard, K., Patras, F.: Exponential Renormalization II: Bogoliubov's R-operation and momentum subtraction schemes. J. Math. Phys. **53**(8), 15 (2012)
11. Ebrahimi-Fard, K., Guo, L., Kreimer, D.: Integrable renormalization. I: the ladder case. J. Math. Phys. **45**(10), 3758–3769 (2004)
12. Ebrahimi-Fard, K., Guo, L., Manchon, D.: Birkhoff type decompositions and the Baker-Campbell-Hausdorff recursion. Commun. Math. Phys. **267**(3), 821–845 (2006)
13. Ebrahimi-Fard, K., Gracia-Bondia, J., Patras, F.: A Lie theoretic approach to renormalization. Commun. Math. Phys. **276**, 519–549 (2007)
14. Ebrahimi-Fard, K., Manchon, D., Patras, F.: A noncommutative Bohnenblust-Spitzer identity for Rota-Baxter algebras solves Bogolioubov's recursion. J. Noncommutative Geom. **3**(2), 181–222 (2009)
15. Ecalle, J.: Singularités non abordables par la géométrie. (French) [Singularities that are inaccessible by geometry] Ann. Inst. Fourier **42**(1–2), 73–164 (1992)
16. Fauvet, F., Menous, F.: Ecalle's arborification-coarborification transforms and Connes-Kreimer Hopf algebra. Ann. Sci. Éc. Norm. Supér. (4) **50**(1), 39–83 (2017)
17. Figueroa, H., Gracia-Bondia, J.M.: Combinatorial Hopf algebras in quantum field theory. I. Rev. Math. Phys. **17**(8), 881–976 (2005)
18. Foissy, L., Patras, F.: Lie theory for quasi-shuffle bialgebras. In: Periods in Quantum Field Theory and Arithmetic. Springer Proceedings in Mathematics and Statistics (to appear)
19. Frabetti, A., Manchon, D.: Five interpretations of Fa Di Bruno's formula. In: Dyson-Schwinger Equations and Fa Di Bruno Hopf Algebras in Physics and Combinatorics, edited by European Mathematical Society, pp. 5–65. Strasbourg, France (2011)
20. Guo, L., Zhang, B.: Renormalization of multiple zeta values. J. Algebra **319**(9), 3770–3809 (2008)
21. Hairer, M.: A theory of regularity structures. Invent. Math. **198**(2), 269–504 (2014)
22. Hoffman, M.E.: Quasi-shuffle products. J. Algebraic Combin. **11**(1), 49–68 (2000)
23. Hoffman, M.E., Ihara, K.: Quasi-shuffle products revisited. J. Algebra **481**, 293–326 (2017)
24. Karandikar, R.L.: Multiplicative decomposition of non-singular matrix valued continuous semimartingales. Ann. Probab. **10**(4), 1088–1091 (1982)
25. Kreimer, D.: Chen's iterated integral represents the operator product expansion. Adv. Theor. Math. Phys. **3**(3), 627–670 (1999)
26. Majid, S.: Foundations of Quantum Group Theory. Cambridge University Press, Cambridge (1995)
27. Manchon, D., Paycha, S.: Shuffle relations for regularised integrals of symbols. Commun. Math. Phys. **270**, 13–51 (2007)
28. Menous, F.: On the stability of some groups of formal diffeomorphisms by the Birkhoff decomposition. Adv. Math. **216**(1), 1–28 (2007)
29. Menous, F.: Formal differential equations and renormalization. Connes, Alain (ed.) et al., Renormalization and Galois theories. European Mathematical Society, IRMA Lectures in Mathematics and Theoretical Physics 15, 229–246 (2009)
30. Menous, F.: Formulas for the Connes-Moscovici Hopf Algebra. In: Ebrahimi-Fard, K., et al. (eds.) Combinatorics and Physics. Contemporary Mathematics, vol. 539, pp. 269–285 (2011)
31. Menous, F.: From dynamical systems to renormalization. J. Math. Phys. **54**(9), 24 (2013)
32. Menous, F., Patras, F.: Logarithmic derivatives and generalized Dynkin operators. J. Algebraic Combin. **38**(4), 901–913 (2013)
33. Murua, A., Sanz-Serna, J.M.: Computing normal forms and formal invariants of dynamical systems by means of word series. Nonlinear Anal. Theory Methods Appl. **138**, 326–345 (2016)
34. Patras, F.: L'algèbre des descentes d'une bigèbre graduée. J. Algebra **170**(2), 547–566 (1994)
35. Patras, F.: Dynkin operators and renormalization group actions in pQFT. In: Bergvelt, M., Yamskulna, G., Zhao, W. (eds.) Vertex Operator Algebras and Related Areas. Contemporary Mathematics, vol. 497, pp. 169–184. American Mathematical Society, Providence (2009)
36. Schützenberger, M.-P.: Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un problème de mathématiques appliquées, Séminaire Dubreil–Jacotin Pisot (Algèbre et théorie des nombres) (1958/1959)
37. Sweedler, M.E.: Hopf algebras. W.A. Benjamin, Inc., New York (1969)

Hopf Algebra Techniques to Handle Dynamical Systems and Numerical Integrators

Ander Murua and Jesús M. Sanz-Serna

Abstract In a series of papers the present authors and their coworkers have developed a family of algebraic techniques to solve a number of problems in the theory of discrete or continuous dynamical systems and to analyze numerical integrators. Given a specific problem, those techniques construct an abstract, *universal* version of it which is solved algebraically; then, the results are transferred to the original problem with the help of a suitable morphism. In earlier contributions, the abstract problem is formulated either in the dual of the shuffle Hopf algebra or in the dual of the Connes-Kreimer Hopf algebra. In the present contribution we extend these techniques to more general Hopf algebras, which in some cases lead to more efficient computations.

1 Introduction

A series of papers [1, 5, 7–10, 25–28, 37] have developed a family of algebraic techniques to solve a number of problems in the theory of discrete or continuous dynamical systems and to analyze numerical integrators. Given a specific problem, those techniques construct an abstract, *universal* version of it which is solved algebraically; then, the result is transferred to the original problem with the help of a suitable morphism Ψ. The abstract problem is formulated either in the dual of the shuffle Hopf algebra of words [27] or in the dual of the Connes-Kreimer Hopf algebra of rooted trees [8]. Operations with elements of the relevant dual are mapped by Ψ into operations with formal series of differential operators. For the shuffle Hopf algebra, the solution of the original problem appears expressed

A. Murua (✉)
Konputazio Zientziak eta A. A. Saila, Informatika Fakultatea, UPV/EHU, Donostia–San Sebastián, Spain
e-mail: Ander.Murua@ehu.es

J. M. Sanz-Serna
Departamento de Matemáticas, Universidad Carlos III de Madrid, Leganés (Madrid), Spain

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_22

as a so-called *word series* [27]. In the Connes-Kreimer case, the resulting series for the original problem are *B-series;* the Butcher group (the group of characters of the Connes-Kreimer Hopf algebra) and B-series first appeared in the context of numerical analysis of differential equations (see [37] for a survey) decades before the Connes-Kreimer Hopf algebra was introduced in the context of renormalization in quantum field theory. Duals of Hopf algebras are useful in this setting because they provide rules for composing formal series.

In the present contribution we extend these techniques to more general Hopf algebras, which in some cases lead to more efficient computations (cf. [15]).

Problems that may be treated in this form include averaging of periodically or quasiperiodically forced differential systems [7–10, 25, 28], construction of formal invariants of motion [8–10, 26] computation of normal forms [26, 28], calculations on central manifolds [5], and error analysis of splitting integrators for deterministic [27] or stochastic [1] systems of differential equations. Of course it would be impossible to take up here each of those problems; the examples in this paper only refer to the computation of high-order averaged systems for periodically forced differential equations and to the analysis of the Strang splitting formula when applied to perturbations of integrable systems.

The techniques studied here go back to a number of earlier developments, in particular, mention has to be made of Ecalle's mould calculus [13, 14] (see [15, 29, 32, 33, 38] for more recent contributions), and of the algebraic theory of integrators [4, 22–24, 37].

An outline of this paper now follows. Section 2 reviews some well-known ideas on the reformulation of differential systems in Euclidean spaces as operator differential equations. Section 3 illustrates the algebraic approach in the series of papers mentioned at the beginning of this introduction. It does so by considering a concrete averaging problem in $\mathbb{R}^5$ and explicitly finding a high-order averaged system by first working abstractly in the group of characters of the shuffle Hopf algebra. The complexity of the computations grows very quickly with the order of the averaged system sought and this motivates the material in Sect. 4, where we show how to work with other Hopf algebras to increase the efficiency of the algorithms. The vector fields (derivations) appearing in the given problem $\mathscr{P}$ in Euclidean space are written as images by a Lie algebra homomorphism Ψ mapping a suitable graded Lie algebra $\tilde{\mathfrak{g}}$ into the Lie algebra of derivations. From $\tilde{\mathfrak{g}}$ we construct a graded, commutative Hopf algebra $\mathscr{H}$ in such a way that an 'abstact' version of $\mathscr{P}$ may be formally solved in the group of characters $\mathscr{G}$ of $\mathscr{H}$; finally the formal solution in $\mathscr{G}$ is translated into a formal solution of $\mathscr{P}$. For the concrete averaging problem in $\mathbb{R}^5$, we present a succession of alternative Hopf algebras that make it possible to compute approximations of increasingly higher order. The final section presents material where the ideas in the paper are suitably modified to cater for problems written in perturbation form, generalizing the notion of *extended word series* introduced in [27] and used in [26, 28].

Due to space constraints, we have not attempted to present the results in the most general conceivable scenario. For instance, it is possible to work with differential systems defined on differentiable manifolds rather than in Euclidean spaces and scalars could be complex rather real.

2 Algebraic Formulation of Differential Systems

It is well known that differential systems in $\mathbb{R}^D$ may be interpreted as differential equations that describe the evolution of suitably chosen linear operators. This section reviews that interpretation, which plays a key role in later developments. We use the following notation. The vector space $\mathscr{C} = C^\infty(\mathbb{R}^D)$ consists of all smooth $\mathbb{R}$-valued functions on $\mathbb{R}^D$. Functions $\chi \in \mathscr{C}$ are sometimes called observables. With respect to the pointwise multiplication of observables, the space $\mathscr{C}$ is an associative and commutative algebra. The symbol $\mathrm{End}(\mathscr{C})$ denotes the vector space of all linear operators $X : \mathscr{C} \to \mathscr{C}$. When operators are multiplied by composition, $(X_1 X_2)(\chi) = X_1(X_2(\chi))$, $\mathrm{End}(\mathscr{C})$ is an associative algebra with a unit: the identity operator $I : \chi \mapsto \chi$.

Consider the initial value problem in $\mathbb{R}^D$

$$\frac{d}{dt}x(t) = f(x(t), t), \quad x(0) = x_0, \tag{1}$$

with f smooth. For each frozen value of t, the vector field $f(\cdot, t)$ defines a first-order linear differential operator $F(t) \in \mathrm{End}(\mathscr{C})$ that associates with each observable χ the observable $F(t)\chi \in \mathscr{C}$ such that

$$F(t)\chi(x) = f(x, t)^T \cdot \nabla\chi(x) = \sum_{j=1}^{D} f_j(x, t)\frac{\partial}{\partial x_j}\chi(x)$$

for each $x = (x_1, \ldots, x_D) \in \mathbb{R}^D$. Actually, $F(t)$ is a derivation of the algebra $\mathscr{C}$, i.e.

$$F(t)(\chi_1\chi_2) = (F(t)\chi_1)\,\chi_2 + \chi_1\,(F(t)\chi_2).$$

The space $\mathrm{Der}(\mathscr{C}) \subset \mathrm{End}(\mathscr{C})$ consisting of all derivations in $\mathscr{C}$ is a Lie algebra with respect to the commutator $[F_1, F_2] = F_1F_2 - F_2F_1$.

Assuming for the time being that for each $x_0 \in \mathbb{R}^d$ the solution $x(t)$ of (1) exists for all $t \in \mathbb{R}$, we may define a one-parameter family $X(t)$, $t \in \mathbb{R}$, of elements of $\mathrm{End}(\mathscr{C})$ as follows: for each observable $\chi \in \mathscr{C}$ and each $t \in \mathbb{R}$, $X(t)\chi \in \mathscr{C}$ is such that $X(t)\chi(x(0)) = \chi(x(t))$ for each $x(0) \in \mathbb{R}^D$. Clearly each $X(t)$ is an automorphism of the algebra $\mathscr{C}$, i.e.

$$X(t)(\chi_1\chi_2) = X(t)(\chi_1)\,X(t)(\chi_2), \tag{2}$$

for any $\chi_1, \chi_2 \in \mathscr{C}$. The set $\mathrm{Aut}(\mathscr{C})$ of all algebra automorphisms is a group for the composition of operators.

Since given $\chi \in \mathscr{C}$,

$$\frac{d}{dt}\chi(x(t)) = \chi'(x(t)) \cdot f(x(t),t)^T \cdot \nabla\chi(x(t))$$

we have that

$$\frac{d}{dt}X(t) = X(t)F(t), \quad X(0) = I, \tag{3}$$

or equivalently

$$X(t) = I + \int_0^t X(s)F(s)\,ds. \tag{4}$$

In this way the solvability of (1) implies the solvability of the operator initial value problem (3). When comparing (3) with (1) we note that (3) is linear in X even when (1) is not linear in x; the multiplication of operators $X(t)F(t)$ in (3) corresponds to the composition of the maps $t \mapsto x(t)$, $(x,t) \mapsto f(x,t)$ that appears in (1).

Conversely assume that $F : \mathbb{R} \to \mathrm{Der}(\mathscr{C})$ is such that there exists a one-parameter family $X(t)$ of elements of $\mathrm{Aut}(\mathscr{C})$ satisfying (3). We then define, for each t, a vector field $f(\cdot,t)$ in R^D by setting $f^i(x,t) = F(t)\chi^i(x)$, $i = 1,\ldots,D$, where χ^i is the i-th coordinate function $\chi^i(x) = x^i$ (superscripts denote components of a vector), and consider the corresponding problem (1). Then, it is easily checked that (1) has, for each x_0, a solution $x(t)$ defined for all real t and the i-component of $x(t)$ may be found as $x^i(t) = (X(t)\chi^i)(x_0)$. We emphasize that for this construction to work it is essential that the operators $X(t)$ satisfy (2), i.e. they are automorphisms of $\mathscr{C}$.

We will present below algebraic frameworks where the initial value problem (3) is interpreted in a broader sense, admitting solution curves $X(t)$ that evolve in groups of *formal automorphisms* rather than in $\mathrm{Aut}(\mathscr{C})$. Roughly speaking such formal automorphisms will be *formal* series of linear maps that preserve multiplication of observables as in (2). Even in the case where $X(t)$ does not correspond to an actual curve in $\mathrm{Aut}(\mathscr{C})$, such formal solution curves $X(t)$ may be used to derive rigorous results on the solution $x(t)$ of (1).

3 An Example

In this section we illustrate the use of Hopf algebra techniques by means of an example: the construction of high-order averaged systems for a periodic differential system in $\mathbb{R}^5$.

3.1 A Highly-Oscillatory Differential System

The following system of differential equations arises in the study of vibrational resonance in an energy harvesting device [11]:

$$\begin{aligned}
\frac{dx}{dt} &= y,\\
\frac{dy}{dt} &= \frac{1}{2}x\left(1-x^2\right) - y + \frac{v}{20} + A\cos\left(\frac{t}{10}\right) + \omega^2\cos(\omega t),\\
\frac{dv}{dt} &= -\frac{v}{100} - \frac{y}{2}.
\end{aligned}$$

Here v is the voltage across the load resistor, x and y are auxiliary state variables, and $\omega \gg 1$ is the frequency of the environmental vibration. The aim is to investigate the effect that the value of the amplitude A of the low-frequency forcing has on the output v.

Averaging, i.e. reducing the time-periodic system to an autonomous system with a help of a periodic change of variables [2, 35], is a very helpful tool to study this kind of problem [25]. To average the vibrational resonance problem above, we begin by introducing new variables

$$x = X - \cos(t\omega), \quad y = Y + \omega\sin(t\omega) + \cos(t\omega), \quad v = V + \frac{1}{2}\cos(t\omega), \tag{5}$$

chosen to ensure that in the transformed system

$$\begin{aligned}
\frac{dX}{dt} &= Y + \cos(t\omega), \qquad (6)\\
\frac{dY}{dt} &= -\frac{X}{4} - \frac{X^3}{2} - Y + \frac{V}{20} + A\cos\left(\frac{t}{10}\right)\\
&\qquad + \left(\frac{3X^2}{2} - \frac{11}{10}\right)\cos(t\omega) - \frac{3}{4}X\cos(2t\omega) + \frac{1}{8}\cos(3t\omega),\\
\frac{dV}{dt} &= -\frac{V}{100} - \frac{Y}{2} - \frac{101}{200}\cos(t\omega),
\end{aligned}$$

the highly oscillatory terms have amplitudes of size $\mathcal{O}(1)$ as $\omega \to \infty$. Suppression of the terms that oscillate with high frequency then results in the averaged system

$$\begin{aligned}
\frac{dX}{dt} &= Y,\\
\frac{dY}{dt} &= -\frac{X}{4} - \frac{X^3}{2} - Y + \frac{V}{20} + A\cos\left(\frac{t}{10}\right),\\
\frac{dV}{dt} &= -\frac{V}{100} - \frac{Y}{2},
\end{aligned}$$

whose solutions approximate the solution $(X(t), Y(t), V(t))$ of (6) with errors of size $\mathcal{O}(1/\omega)$ in bounded intervals $0 \le t \le T < \infty$. Approximations with $\mathcal{O}(1/\omega)$ errors (first-order averaging) to the original state variables x, y v, are then obtained from (5). Approximations to x, y v, with errors $\mathcal{O}(1/\omega^2)$ (second-order averaging) may be obtained by changing variables in (6) so as to reduce the amplitude of the highly oscillatory terms from $\mathcal{O}(1)$ to $\mathcal{O}(1/\omega)$ and then discarding the highly oscillatory terms. The iteration of the procedure leads successively to approximations with errors $\mathcal{O}(1/\omega^n)$ for $n = 3, 4, \ldots$ (high-order averaging).

The averaged systems found in this way are nonautonomous since the low-frequency forcing is not averaged out. In order to deal with autonomous averaged problems we introduce two additional real-valued state variables C, S satisfying

$$\frac{dC}{dt} = -\frac{S}{10}, \qquad \frac{dS}{dt} = \frac{C}{10}$$

and with initial conditions $C(0) = 1$, $S(0) = 1$, so that $C(t) = \cos(t/10)$, and write problem (6) as

$$\begin{aligned}
\frac{dX}{dt} &= Y + \cos(t\omega), \qquad (7)\\
\frac{dY}{dt} &= -\frac{X}{4} - \frac{X^3}{2} - Y + \frac{V}{20} + AC\\
&\quad + \left(\frac{3X^2}{2} - \frac{11}{10}\right)\cos(t\omega) - \frac{3}{4}X\cos(2t\omega) + \frac{1}{8}\cos(3t\omega),\\
\frac{dV}{dt} &= -\frac{V}{100} - \frac{Y}{2} - \frac{101}{200}\cos(t\omega),\\
\frac{dC}{dt} &= -\frac{S}{10},\\
\frac{dS}{dt} &= \frac{C}{10}.
\end{aligned}$$

Note that this system in $\mathbb{R}^5$ is of the form (1) with

$$f(x,t) = f_a(x) + \cos(t\omega)\, f_b(x) + \cos(2t\omega)(x)\, f_c + \cos(3t\omega)\, f_d(x).$$

It is trivial to write down the derivations $F_a, \ldots, F_d$, associated with $f_a, \ldots, f_d$. For instance:

$$F_a = Y\partial_X + \left(-\frac{X}{4} - \frac{X^3}{2} - Y + \frac{V}{20} + AC\right)\partial_Y + \left(-\frac{V}{100} - \frac{Y}{2}\right)\partial_V - \frac{S}{10}\partial_C + \frac{C}{10}\partial_S.$$

Then the derivation corresponding to $f(x,t)$ is, for each t,

$$F_a + \cos(t\omega)\, F_b + \cos(2t\omega)\, F_c + \cos(3t\omega)\, F_d. \tag{8}$$

3.2 *Solving the Oscillatory Problem with Word Series*

We now introduce the alphabet $\mathscr{A} = \{a, b, c, d\}$, the corresponding (infinite) set $\mathscr{W}$ of all words $a, \ldots, d, aa, ab, \ldots, dd, aaa, \ldots$ (including the empty word 1) and the free associative algebra $\mathbb{R}\langle\mathscr{A}\rangle$ consisting of all the *linear combinations* of words with real coefficients. Multiplication $\star$ in $\mathbb{R}\langle\mathscr{A}\rangle$ is defined by concatenating words [34], which implies that 1 is the unit of this (noncommutative) algebra.

Furthermore we consider again the vector space of linear combinations of words but now endow it with the (commutative) shuffle product $\sqcup\!\sqcup$ and denote by $\mathscr{H}$ the resulting (shuffle) algebra. Actually $\mathscr{H}$ is a Hopf algebra for the deconcatenation coproduct. This algebra is *graded*; its graded component of degree n, $n = 0, 1, \ldots$, consists of the linear combinations of words with n letters. The dual vector space $\mathscr{H}^*$ may be identified with the set of all *formal series* α of the form $\sum_{\mathscr{W}} c_w w$ for real $c_w \in \mathbb{R}$ so that the image $\langle\alpha, w\rangle$ of the word w by the linear form α is the coefficient c_w. Thus $\mathscr{H}^*$ is a much larger space than $\mathbb{R}\langle\mathscr{A}\rangle$. Note that the concatenation product $\star$ may be extended from $\mathbb{R}\langle\mathscr{A}\rangle$ to $\mathscr{H}^*$ in an obvious way. We denote by $\mathscr{G} \subset \mathscr{H}^*$ the group of characters of $\mathscr{H}$ consisting of the elements $\gamma \in \mathscr{H}^*$ that satisfy the shuffle relations: $\langle\gamma, w \sqcup\!\sqcup w'\rangle = \langle\gamma, w\rangle\langle\gamma, w'\rangle$ for all words w, w'. The Lie algebra of infinitesimal characters $\mathfrak{g} \subset \mathscr{H}^*$ consists of those $\beta \in \mathscr{H}^*$ such that $\langle\beta, w \sqcup\!\sqcup w'\rangle = \langle\beta, w\rangle\langle 1, w'\rangle + \langle 1, w\rangle\langle\beta, w'\rangle$ for each pair of words. Characters and infinitesimal characters are related through the relations $\mathscr{G} = \exp(\mathfrak{g})$, $\mathfrak{g} = \log(\mathscr{G})$, i.e. each element γ in the group is the exponential $1 + \beta + (1/2)\beta \star \beta + \cdots$ of the element $\beta = (\gamma - 1) - (1/2)(\gamma - 1) \star (\gamma - 1) + \cdots$ See [27, Sec. 6.1] for a review of the constructions above.

To solve (7), we associate with each letter in $\mathscr{A}$ the corresponding derivation in the expression (8), i.e. we set

$$\Psi(a) = F_a, \quad \Psi(b) = F_b, \quad \Psi(c) = F_c, \quad \Psi(d) = F_d, \tag{9}$$

and extend the mapping Ψ to an algebra morphism from $\mathbb{R}\langle\mathscr{A}\rangle$ to the algebra $\mathrm{End}_{\mathbb{R}}(\mathscr{C})$, $D = 5$, by setting $\Psi(aa) = F_a F_a$, $\Psi(ab) = F_a F_b$, etc. The free Lie algebra $\mathscr{L}(\mathscr{A})$ is the linear subspace of $\mathbb{R}\langle\mathscr{A}\rangle$ consisting of linear combinations of iterated commutators such as $[a, b] = ab - ba$, $[a, [a, b]] = a[a, b] - b[a, b] = aab - aba - bab + bba, \ldots$ (the letters $a, \ldots, b$ are seen as iterated commutators of order $n = 1$). This Lie algebra is graded; its graded component of degree n, $n = 1, 2, \ldots$, consists of the linear combinations of iterated commutators involving words with n letters. The restriction of Ψ to $\mathscr{L}(\mathscr{A})$ is a Lie algebra morphism $\mathscr{L}(\mathscr{A}) \to \mathrm{Der}_{\mathbb{R}}(\mathscr{C}) \subset \mathrm{End}_{\mathbb{R}}(\mathscr{C})$. Note that, for fixed t, (8) is the image under

Ψ of the element

$$\beta(t) = a + \cos(t\omega)\, b + \cos(2t\omega)\, c + \cos(3t\omega)\, d \in \mathscr{L}(\mathscr{A}). \tag{10}$$

The 'abstract' initial value problem

$$\frac{d}{dt}\alpha(t) = \alpha(t) \star \beta(t), \quad \alpha(0) = 1, \tag{11}$$

where at the outset $\alpha(t)$ is sought as a curve in $\mathbb{R}\langle\mathscr{A}\rangle$ is such that the mapping Ψ transforms it into the operator initial value problem (3) corresponding to (7). We shall solve (11), and then the application of Ψ will lead to a solution of (7).

We recall that for integrable[1] curves $\beta(t)$ in $\mathfrak{g} \supset \mathscr{L}(\mathscr{A})$ (and in particular for $\beta(t)$ in (10)), the problem (11) possesses a unique formal solution $\alpha(t)$ that for each t lies in the space of formal series $\mathscr{H}^* \supset \mathbb{R}\langle\mathscr{A}\rangle$. This solution may be found by a Picard iteration and is given by a Chen series [34]

$$\alpha(t) = \sum_{w \in \mathscr{W}} \langle \alpha(t), w \rangle \; w,$$

where for each $w \in W$ the coefficient, $\langle \alpha(t), w \rangle$ has a known expression as an iterated integral (see e.g. [27, Sec. 2.1], [28, Sec. 2.1] for details). Furthermore, for each t, $\alpha(t)$ satisfies the shuffle relations and therefore belongs to the group of characters $\mathscr{G} \subset \mathscr{H}^*$. In other words, when seen as a nonautonomous initial value problem to determine a curve $\alpha(t)$ in the group $\mathscr{G}$ given a curve $\beta(t)$ in the algebra $\mathfrak{g}$, (11) is uniquely solvable (see e.g. [27, Sec 2.2.4]). For each fixed t, $\Psi(\alpha(t))$ (Ψ is applied in the obvious term by term way) is a formal series whose terms belong to $\mathrm{End}(\mathscr{C})$ (they are actually differential operators). Furthermore the fact that $\alpha(t) \in \mathscr{G}$ implies (see e.g. [27, Sec. 6.1.3]) that the formal series $\Psi(\alpha(t))$ satisfies (2), i.e. it is formally an automorphism, and by proceeding as in the preceding section we then find that the solutions of our problem in $\mathbb{R}^5$ may be represented as formal series

$$x(t) = \sum_{w \in \mathscr{W}} \langle \alpha(t), w \rangle f_w(x(0)), \qquad x(0) \in \mathbb{R}^5,$$

where the mappings $f_w : \mathbb{R}^5 \to \mathbb{R}^5$ are the so-called *word basis functions* [27]; the i-th component of f_w is obtained by applying to the i-coordinate function χ^i the endomorphism $\Psi(w)$. Series of this form are called *word series* [1, 26–28, 37].

[1] More precisely it is sufficient to ask that, for each word, the real-valued function $\langle \beta(t), w \rangle$ be locally integrable.

3.3 Averaging with Word Series

We now average (7) by first averaging its abstract version (10) and (11). We seek a $2\pi/\omega$-periodic map $\kappa : \mathbb{R} \to \mathscr{G}$ and a (time-independent) $\bar{\beta} \in \mathfrak{g}$ such that

$$\frac{d}{dt}\kappa(t) = \kappa(t) \star \beta(t) - \bar{\beta} \star \kappa(t). \tag{12}$$

It is easily checked that then $\alpha(t) = \exp(\bar{\beta}\, t) \star \kappa(t)$; in this way the formal solution $\alpha(t)$ of the periodic problem (10) and (11) is obtained, via the *periodic map* $\kappa(t)$, from the solution $\bar{\alpha}(t) = \exp(\bar{\beta}\, t)$ of the linear *autonomous* problem $(d/dt)\bar{\alpha} = \bar{\alpha}(t) \star \bar{\beta}$, $\bar{\alpha}(0) = 1$ (the averaged problem).

There is some freedom when solving (12). In *stroboscopic averaging* one imposes the additional condition $\kappa(0) = 1$, so that the averaged solution $\bar{\alpha}(t)$ coincides with $\alpha(t)$ at all stroboscopic times $t_k = k(2\pi/\omega)$, $k \in \mathbb{Z}$ [8]. Alternatively, it is also possible to impose the *zero-mean* condition

$$\int_0^{2\pi/\omega} \log(\kappa(t))dt = 0. \tag{13}$$

(Note that the stroboscopic condition demands that $\log(\kappa(t))$ vanishes at $t = 0$ rather than on average over a period as in (13).)

By proceeding recursively with respect to the number of letters in the words involved, the stroboscopic condition (respectively the zero-mean condition) and (12) uniquely determine all the coefficients of the formal series $\bar{\beta}$ and $\kappa(t)$.[2] We have implemented the corresponding recursions in a symbolic manipulation package. As an example, when truncating the series for $\bar{\beta}$ so as to only keep words with three or less letters, we find, in the zero-mean case:

$$\begin{aligned}\bar{\beta}^{[3]} = a + \frac{1}{\omega^2}\Big(&\tfrac{1}{4}\,abb - \tfrac{1}{2}\,bab + \tfrac{1}{4}\,bba + \tfrac{1}{16}\,acc - \tfrac{1}{8}\,cac + \tfrac{1}{16}\,cca \\ &+ \tfrac{1}{36}\,add - \tfrac{1}{18}\,dad + \tfrac{1}{36}\,dda - \tfrac{1}{8}\,bbc + \tfrac{1}{4}\,bcb - \tfrac{1}{8}\,cbb \\ &- \tfrac{1}{12}\,bcd + \tfrac{1}{8}\,bdc - \tfrac{1}{24}\,cbd + \tfrac{1}{8}\,cdb - \tfrac{1}{24}\,dbc - \tfrac{1}{12}\,dcb\Big) \in \mathfrak{g}.\end{aligned}$$

The corresponding result under the stroboscopic condition is similar but includes more terms (40 rather than 19).

Now that the problem (10) and (11) has been averaged, we apply the transformation Ψ to average our problem in the Euclidean space $\mathbb{R}^5$. From $\bar{\beta}$ we obtain the

[2] For stroboscopic averaging, the recursions that allow the simple computation of the coefficients of $\bar{\beta}$ and $\kappa(t)$ may be seen in [8] or [28], but in those references $\bar{\beta}$ and $\kappa(t)$ are found with the help of an auxiliary transport equation rather than via (12).

formal vector field given by the word-series

$$\bar{f}(x) = \sum_{w \in \mathscr{W}} \langle \bar{\beta}, w \rangle f_w(x),$$

and from $\kappa(t)$ we construct the formal periodic change of variables given by the word-series

$$U(x,t) = \sum_{w \in \mathscr{W}} \langle \kappa(t), w \rangle f_w(x),$$

such that the solutions $x(t)$ of (7) are formally given as $x(t) = U(\bar{x}(t), t)$ with $(d/dt)\bar{x} = \bar{f}(\bar{x})$.

To deal with *bona fide* vector fields and changes of variables, one has to truncate the corresponding formal series. In our example, the truncation $\bar{\beta}^{[3]}$ found above leads to a vector field in $\mathbb{R}^5$ which after eliminating the auxiliary variables C and S, reduces to the following time-dependent vector field in $\mathbb{R}^3$:

$$Y\,\partial_X + \left(-\frac{X}{4} - \frac{X^3}{2} - Y + \frac{V}{20} + A\,\cos(\frac{t}{10})\right)\partial_Y - \left(\frac{V}{100} + \frac{Y}{2}\right)\partial_V$$
$$+\frac{1}{\omega^2}\left(\frac{3X}{4}\,\partial_X + \left(-\frac{9X^3}{4} + \frac{51X}{640} - \frac{3Y}{4}\right)\partial_Y - \frac{3X}{8}\,\partial_V\right).$$

With the help of a truncated change of variables, the solutions of the corresponding differential system provides $\mathscr{O}(1/\omega^3)$ approximations to $X(t)$, $Y(t)$, $V(t)$ in (7). Truncations of this kind and their accuracy are discussed in detail in [9] and [10].

It is important to emphasize that the construction above is *universal:* $\bar{\beta}$ and $\kappa(t)$ would not change if the expressions for the vector fields $f_a, \ldots, f_d$ in $\mathbb{R}^5$ considered above were replaced by another set of four vector fields in $\mathbb{R}^D$ with arbitrary D. There is a price to be paid for this generality: in our case there are 4^n words with n letters and accordingly the complexity of the computations grows very quickly as n increases. In a laptop computer our computations had to be limited to $n \leq 8$. In what follows we show how to replace the shuffle Hopf algebra $\mathscr{H}$ by alternative Hopf algebras which may lead to simpler computations.

4 General Hopf Algebras

In preceding section we studied the operator initial value problem (3) with the help of a mapping Ψ whose restriction to the free Lie algebra $\mathscr{L}(\mathscr{A})$ is a Lie algebra morphism into the algebra of derivations Der($\mathscr{C}$). We now study the more general

situation where $\mathscr{L}(\mathscr{A})$ is replaced by a graded Lie algebra

$$\tilde{\mathfrak{g}} = \bigoplus_{n\geq 1} \mathfrak{g}_n, \tag{14}$$

with finite-dimensional homogeneous subspaces $\mathfrak{g}_n$,[3] and there are a Lie algebra homomorphism $\Psi : \tilde{\mathfrak{g}} \to \mathrm{Der}(\mathscr{C})$ and a curve $\beta : \mathbb{R} \to \tilde{\mathfrak{g}}$ such that $\Psi(\beta(t)) = F(t)$ for all $t \in \mathbb{R}$.

Note that Ψ can be uniquely extended to an associative algebra homomorphism from the universal enveloping algebra $U(\tilde{\mathfrak{g}})$ of $\tilde{\mathfrak{g}}$ to $\mathrm{End}(\mathscr{C})$, which we denote with the same symbol Ψ. We shall use the symbol $\star$ to denote the (associative) product in $U(\tilde{\mathfrak{g}})$ such that $[G_1, G_2] = G_1 \star G_2 - G_2 \star G_1$ for all $G_1, G_2 \in U(\tilde{\mathfrak{g}})$. In the particular case where (14) is the free Lie algebra generated by a finite alphabet $\mathscr{A}$, $U(\tilde{\mathfrak{g}})$ coincides with $\mathbb{R}\langle\mathscr{A}\rangle$ and $\star$ is the concatenation product.

4.1 Solving the Operator Initial Value Problem

We denote by

$$\{G_i \ : i \in \mathscr{I}\} \tag{15}$$

a homogeneous basis of the graded Lie algebra (14), where $\mathscr{I}$ is some set of indices, $\mathscr{I} = \bigcup_{n\geq 1} \mathscr{I}_n$, and $\{G_i \ : i \in \mathscr{I}_n\}$ is a basis of $\mathfrak{g}_n$ for each $n \geq 1$. If $\beta(t) = \sum_{i\in\mathscr{I}} \lambda_i(t) G_i$, we rewrite (4) as

$$X(t) = I + \sum_{i\in\mathscr{I}} \int_0^t \lambda_i(t) X(t) \Psi(G_i)\, dt,$$

an equation that may be solved by the following Picard iteration,

$$\begin{aligned}
X^{[0]}(t) &= I \\
X^{[1]}(t) &= I + \sum_{i\in\mathscr{I}} \int_0^t \lambda_i(t) X^{[0]}(t) \Psi(G_i)\, dt = I + \sum_{i\in\mathscr{I}} \left(\int_0^t \lambda_i(t)\, dt \right) \Psi(G_i), \\
X^{[2]}(t) &= I + \sum_{i\in\mathscr{I}} \int_0^t \lambda_i(t) X^{[1]}(t) \Psi(G_i)\, dt \\
&\cdots = \cdots
\end{aligned}$$

[3] It is not essential to assume that each $\mathfrak{g}_n$ is finite dimesional. The arguments below may be readily adapted to cover the general case under the proviso that the summation in (23) is well defined (cf. second paragraph after (24)).

In this way, one may construct a formal solution $X(t)$ of (3) of the form

$$X(t) = I + \sum_{m \geq 1} \sum_{(i_1,\ldots,i_m) \in \mathscr{I}^m} a_{i_1,\ldots,i_m}(t)\, \Psi(G_{i_1}) \cdots \Psi(G_{i_m}). \tag{16}$$

Unfortunately, this series is unnecessarily complicated as there are many linear dependencies among the endomorphisms of the form $\Psi(G_{i_1}) \cdots \Psi(G_{i_m})$. For instance, $\Psi(G_{i_1})\Psi(G_{i_2}) - \Psi(G_{i_2})\Psi(G_{i_1})$ has to coincide with $\Psi([G_{i_1}, G_{i_2}])$ and therefore must be a linear combination of endomorphisms $\Psi(G_i)$, $i \in \mathscr{I}$.[4]

If $<$ denotes a total order relation in $\mathscr{I}$, the Poincaré-Birkhoff-Witt (PBW) theorem ensures that the products

$$\{G_{i_1} \star \cdots \star G_{i_m} \ : \ i_1 \leq \cdots \leq i_m\} \tag{17}$$

(including the empty product equal to the unit element $\mathbb{1}$) provide a basis of $U(\tilde{\mathfrak{g}})$. Therefore, it is possible to simplify (16) by removing the linear dependencies in the right-hand side so as to end up with a formal series that only uses endomorphisms of the form $\Psi(G_{i_1}) \cdots \Psi(G_{i_m})$ with $i_1 \leq \cdots \leq i_m$. However the basis of $U(\tilde{\mathfrak{g}})$ given by the PBW theorem may not be the most convenient in practice[5] and in what follows we shall work with an arbitrary homogeneous basis of $U(\tilde{\mathfrak{g}})$

$$\{Z_j \ : \ j \in \mathscr{J}\}, \quad \mathscr{J} = \bigcup_{n \geq 0} \mathscr{J}_n \tag{18}$$

where $\mathscr{J}$ is some set of indices and $\{Z_i \ : i \in \mathscr{J}_n\}$ is, for each $n \geq 0$, a basis of the graded component of degree n. Note that the structure constants $\lambda^i_{i',i''}$ of the basis $\{G_i \ : \ i \in \mathscr{I}\}$ of the Lie algebra $\tilde{\mathfrak{g}}$,

$$[G_{i'}, G_{i''}] = \sum_i \lambda^i_{i',i''} G_i, \quad i', i'' \in \mathscr{I},$$

uniquely determine (see Sect. 4.5) the structure constants $\mu^j_{j',j''}$ of the basis $\{Z_j \ : \ i \in \mathscr{J}\}$ of $U(\tilde{\mathfrak{g}})$,

$$Z_{j'} \star Z_{j''} = \sum_j \mu^j_{j',j''} Z_j, \quad j', j'' \in \mathscr{J}. \tag{19}$$

[4] In the case where (14) is the free Lie algebra generated by a finite alphabet $\mathscr{A}$, we saw that it is possible to write the solution $X(t)$ as a series constructed from endomorphisms of the form $\Psi(G_{a_1}) \cdots \Psi(G_{a_m})$, with the $a_i \in \mathscr{A}$; this is far more compact than (16), which involves terms $\Psi(G_{i_1}) \cdots \Psi(G_{i_m})$ made of arbitrary elements G_i of the basis.

[5] This was illustrated in the preceding section, where we used the basis of $\mathbb{R}\langle\mathscr{A}\rangle$ consisting of words.

4.2 Constructing the Hopf Algebra

We now construct a Hopf algebra $\mathscr{H}$ which will play in the present circumstances the role that the shuffle Hopf algebra had in the preceding section. The presentation that follows uses explicitly the choice of basis in (18); this is convenient for the computational purposes we have in mind. However the Hopf algebra $\mathscr{H}$ that we shall construct is in fact independent of the choice of basis, as shown in Sect. 4.5 below. In the particular case where $\tilde{\mathfrak{g}}$ is freely generated by the elements of a finite alphabet $\mathscr{A}$, the construction below results in the shuffle Hopf algebra of the preceding section.

For each $j \in \mathscr{J}$ we consider the linear form u_j on $U(\tilde{\mathfrak{g}})$ that takes the value 1 at the element Z_j and vanishes at each $Z_{j'}$, $j' \neq j$ and set $\mathscr{H}$ equal to the graded dual $\bigoplus_{n\geq 0} \mathscr{H}_n$ of $U(\tilde{\mathfrak{g}})$, i.e. each $\mathscr{H}_n$ is the subspace of the linear dual $U(\tilde{\mathfrak{g}})^*$ of $U(\tilde{\mathfrak{g}})$ spanned by the u_j, $j \in \mathscr{J}_n$.

We define a product in $\mathscr{H}$ as follows. The algebra $U(\tilde{\mathfrak{g}})$ possesses a canonical coalgebra structure whose coproduct $\Delta : U(\tilde{\mathfrak{g}}) \to U(\tilde{\mathfrak{g}}) \otimes U(\tilde{\mathfrak{g}})$ is uniquely determined by requiring that

- $\Delta(\beta) = \mathbb{1} \otimes \beta + \beta \otimes \mathbb{1}$, for all $\beta \in \tilde{\mathfrak{g}}$,
- Δ be an algebra homomorphism.

This coproduct is by construction cocommutative, i.e. if, for each $j \in \mathscr{J}$,

$$\Delta(Z_j) = \sum_{j',j'' \in \mathscr{J}} \eta^j_{j',j''} Z_{j'} \otimes Z_{j''}.$$

then $\eta^j_{j',j''} = \eta^j_{j'',j'}$. Through the duality between the vector spaces $U(\tilde{\mathfrak{g}})$ and $\mathscr{H}$, Δ induces the following commutative multiplication operation in $\mathscr{H}$:

$$u_{j'} u_{j''} = \sum_{j \in \mathscr{J}} \eta^j_{j',j''} u_j = \sum_{j \in \mathscr{J}} \langle \Delta(Z_j), u_{j'} \otimes u_{j''} \rangle \, u_j .$$

Similarly, the product $\star$ in $U(\tilde{\mathfrak{g}})$ with structure constants given in (19) induces by duality a coproduct $\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}$ given by

$$\Delta(u_j) = \sum_{j',j'' \in \mathscr{J}} \mu^j_{j',j''} \, u_{j'} \otimes u_{j''}, \qquad j \in \mathscr{J} \tag{20}$$

(our hypotheses ensure that the summation in (20) has finitely-many non-zero terms and is therefore well defined). In this way $\mathscr{H}$ is a connected, commutative, graded Hopf algebra.

We now turn to the dual vector space $\mathscr{H}^*$. Each element $\gamma \in \mathscr{H}^*$ may be represented as a formal series

$$\gamma = \sum_{j \in \mathscr{J}} \langle \gamma, u_j \rangle \, Z_j,$$

where $\langle \gamma, u_j \rangle$ is the image of $u_j \in \mathscr{H}$ by the linear form γ. Thus $\mathscr{H}^*$ may be seen as a superspace of $U(\tilde{\mathfrak{g}})$. The associative algebra structure of $U(\tilde{\mathfrak{g}})$ may be extended naturally to $\mathscr{H}^*$: for $\gamma', \gamma'' \in \mathscr{H}^*$, the series that represents their product $\gamma = \gamma' \star \gamma'' \in \mathscr{H}^*$ is given by

$$\begin{aligned}
\sum_{j\in\mathscr{J}} \langle \gamma, u_j \rangle Z_j &= \left(\sum_{j'\in\mathscr{J}} \langle \gamma', u_{j'} \rangle Z_{j'} \right) \star \left(\sum_{j''\in\mathscr{J}} \langle \gamma'', u_{j''} \rangle Z_{j''} \right) \\
&= \sum_{j',j''\in\mathscr{J}} \langle \gamma', u_{j'} \rangle \langle \gamma'', u_{j''} \rangle Z_{j'} \star Z_{j''} \\
&= \sum_{j',j''\in\mathscr{J}} \langle \gamma', u_{j'} \rangle \langle \gamma'', u_{j''} \rangle \sum_{j\in\mathscr{J}} \mu^j_{j',j''} Z_j \\
&= \sum_{j\in\mathscr{J}} \left(\sum_{j',j''\in\mathscr{J}} \mu^j_{j',j''} \langle \gamma', u_{j'} \rangle \langle \gamma'', u_{j''} \rangle \right) Z_j .
\end{aligned}$$

In other words

$$\langle \gamma, u_j \rangle = \sum_{j',j''\in\mathscr{J}} \mu^j_{j',j''} \langle \gamma', u_{j'} \rangle \langle \gamma'', u_{j''} \rangle$$

i.e. the product $\star$ in $\mathscr{H}^*$ corresponds via duality to the coproduct (20) in $\mathscr{H}$. The group of characters of $\mathscr{H}$ and the Lie algebra of infinitesimal characters are

$$\mathscr{G} = \left\{ \gamma \in \mathscr{H}^* \ : \ \langle \gamma, u_{j'} u_{j''} \rangle = \langle \gamma, u_{j'} \rangle \langle \gamma, u_{j''} \rangle \right\},$$

and

$$\mathfrak{g} = \left\{ \gamma \in \mathscr{H}^* \ : \ \langle \gamma, u_{j'} u_{j''} \rangle = \langle \gamma, u_{j'} \rangle \langle \mathbb{1}, u_{j''} \rangle + \langle \mathbb{1}, u_{j'} \rangle \langle \gamma, u_{j''} \rangle \right\},$$

respectively. These are related by a bijection $\exp : \mathfrak{g} \to \mathscr{G}$, as we saw in the particular case considered in the preceding section.

The abstract initial value problem

$$\frac{d}{dt}\alpha(t) = \alpha(t) * \beta(t), \quad \alpha(0) = \mathbb{1}, \tag{21}$$

with $\beta(t)$ any given integrable curve in $\mathfrak{g}$ possesses a solution that for each t is an element of $\mathscr{G}$. This solution may be computed by finding its coefficients by recursion with respect to the grading. In particular this is so for the curve such that $\Psi(\beta(t)) = F(t)$ for all $t \in \mathbb{R}$, whose existence we assumed at the beginning of this section. We next translate this result into a result for the operator problem.

4.3 Back to the Operator Initial Value Problem

The mapping Ψ may be defined on $\mathscr{H}^* \supset U(\tilde{\mathfrak{g}})$ as an algebra map from $\mathscr{H}^*$ to the direct product algebra $\prod_{n\geq 0} \mathrm{End}(\mathscr{C})$ sending each $\gamma \in \mathscr{H}^*$ to

$$\Psi(\gamma) = \sum_{n\geq 0} \sum_{j' \in \mathscr{J}_n} \langle \gamma, u_{j'} \rangle \, \Psi(Z_{j'}) = \sum_{j \in \mathscr{J}} \langle \gamma, u_j \rangle \, \Psi(Z_j).$$

For the product of two formal series of endomorphisms we have

$$\left(\sum_{j' \in \mathscr{J}} \langle \gamma', u_{j'} \rangle \, \Psi(Z_{j'}) \right) \left(\sum_{j'' \in \mathscr{J}} \langle \gamma'', u_{j''} \rangle \, \Psi(Z_{j''}) \right) = \sum_{j \in \mathscr{J}} \langle \gamma, u_j \rangle \, \Psi(Z_j),$$

where $\gamma = \gamma' \star \gamma'' \in \mathscr{H}^*$.

The solution $\alpha(t)$ of the abstract initial value problem leads to the following formal solution of (3) (a compact alternative to (16))

$$X(t) = \sum_{j \in \mathscr{J}} \langle \alpha(t), u_j \rangle \, \Psi(Z_j),$$

and we shall check presently that (2) is formally satisfied in order to obtain a formal solution of the initial value problem (1).

From the definition of Δ, for arbitrary $\chi_1, \chi_2 \in \mathscr{C}$,

$$\Psi(Z_j)(\chi_1 \chi_2) = \sum_{j', j'' \in \mathscr{J}} \eta^j_{j', j''} \, (\Psi(Z_{j'})\chi_1) \, (\Psi(Z_{j''})\chi_2)$$

and it follows by duality that, for each $\gamma \in \mathscr{H}^*$,

$$\sum_{j \in \mathscr{J}} \langle \gamma, u_j \rangle \, \Psi(Z_j)(\chi_1 \chi_2) = \sum_{j', j'' \in \mathscr{J}} \langle \gamma, u_{j'} u_{j''} \rangle \, (\Psi(Z_{j'})\chi_1) \, (\Psi(Z_{j''})\chi_2).$$

If, in particular, $\gamma \in \mathscr{G}$, then the right-hand side of the last equality coincides with

$$\left(\sum_{j' \in \mathscr{J}} \langle \gamma, u_{j'} \rangle \, \Psi(Z_{j'})\chi_1 \right) \left(\sum_{j'' \in \mathscr{J}} \langle \gamma, u_{j''} \rangle \, \Psi(Z_{j''})\chi_2 \right),$$

i.e. the formal series $\sum_{j \in \mathscr{J}} \langle \gamma, u_j \rangle \, \Psi(Z_j)$ is formally an automorphism. This is in particular true, for each t, for the series $X(t)$ above, since we know that $\alpha(t) \in \mathscr{G}$.

4.4 Averaging with More General Hopf Algebras

An abstract problem of the form (21) with $\beta(t)$ $2\pi/\omega$-periodic with values in $\mathfrak{g}$ may be averaged with the help of Eq. (12) exactly as we saw in the case of the shuffle Hopf algebra. The result may be then transferred, via the morphism Ψ, to average periodically forced systems (1) .

4.4.1 Averaging with Decorated Rooted Trees

As an illustration we take up again the task of averaging (7) but this time we work with the Grossman–Larson graded Lie algebra of rooted trees [17] with vertices decorated by letters of the alphabet $\mathscr{A} = \{a, b, c, d\}$. We use once more (9) and extend Ψ to a Lie algebra morphism from the Grossman-Larson Lie algebra to the Lie algebra of derivations $\mathrm{Der}(\mathscr{C})$. In the construction above, $\mathscr{H}$ is the Connes-Kreimer Hopf algebra of rooted trees and the group of characters $\mathscr{G}$ is the Butcher group.

As an example we find, under the zero-mean condition and truncating the contributions of trees with four or more vertices:

$$\begin{aligned}\bar\beta^{[3]} = \underline{a} + \frac{1}{\omega^2}\Big(&\tfrac14\,\underline{a[b[b]]} - \tfrac14\,\underline{b[ab]} + \tfrac14\,b[b[a]] + \tfrac14\,\underline{a[b^2]} - \tfrac12\,\underline{b[a[b]]}\\ &- \tfrac18\,\underline{c[a[c]]} + \tfrac{1}{16}\,a[c^2] - \tfrac{1}{16}\,c[ac] + \tfrac{1}{16}\,a[c[c]] + \tfrac{1}{16}\,c[c[a]]\\ &+ \tfrac{1}{36}\,a[d[d]] + \tfrac{1}{36}\,d[d[a]] - \tfrac{1}{36}\,d[ad] + \tfrac{1}{36}\,a[(d)^2] - \tfrac{1}{18}\,d[a[d]]\\ &- \tfrac18\,b[b[c]] + \tfrac18\,b[bc] - \tfrac18\,c[b[b]] + \tfrac14\,b[c[b]] - \tfrac18\,c[(b)^2] - \tfrac18\,c[b[b]]\\ &+ \tfrac14\,b[c[b]] - \tfrac18\,c[(b)^2] - \tfrac{1}{12}\,b[c[d]] + \tfrac{1}{12}\,c[bd] - \tfrac{1}{12}\,d[c[b]]\\ &+ \tfrac18\,b[d[c]] + \tfrac18\,c[d[b]] - \tfrac18\,d[bc] - \tfrac{1}{24}\,c[b[d]] + \tfrac{1}{24}\,b[cd] - \tfrac{1}{24}\,d[b[c]]\Big).\end{aligned}$$

Here the notation for rooted trees is as follows:

- a denotes the one-vertex rooted tree where the root is decorated with the symbol a,
- $d[b[c]]$ denotes the 'tall' rooted tree where the decoration d corresponds to the root, the vertex decorated by b is linked to the root, and the vertex decorated with c is linked to the vertex decorated with b,
- $d[bc]$ denotes the 'bushy' rooted tree where d is the decoration of the root and the vertices with decoration b and c are linked to the root, etc.

As in the case of words, results on the abstract problem are transferred to Euclidean space with the help of Ψ. We again find a system $(d/dt)\bar x = \bar f(\bar x)$, where $\bar f$ is a formal series of vector fields in $\mathbb{R}^5$ and a $2\pi/\omega$-periodic formal change of variables $x = U(\bar x, t)$ such that solutions $x(t)$ of (7) are formally given as

$x(t) = U(\bar{x}(t), t)$. Now the formal series are indexed by rooted trees rather than by words, i.e. they are B-series [7, 8, 37].

What is the advantage of using rooted trees rather than words? The expression for $\bar{\beta}^{[3]}$ displayed above, with 33 rooted trees, is obviously more involved than its counterpart with words involving 19 words. However the images by Ψ of many trees vanish. For instance, in the display above only the five rooted trees underlined have a nonzero image. This may be exploited by working in the quotient by $\ker(\Psi)$ of the Lie algebra of rooted trees, thereby decreasing the dimension of the graded components, which allows symbolic manipulation packages to take the expansions to higher order. A further reduction may be achieved by noting that $\bar{\beta}$ has to be a Lie element, i.e. it must be expressible in terms of commutators. We may then work in the Lie subalgebra generated by $a,\ldots,d$ of the previous quotient subalgebra. For instance for the display above we find the compact expression

$$a + \frac{1}{\omega^2}\left(\tfrac{1}{4}\,[b,[b,a]] - \tfrac{1}{8}\,[c,[a,c]]\right).$$

4.4.2 Averaging in a Lie Algebra Generated by Monomial Vector Fields

We have just seen how to work in a Lie algebra better suited to the concrete example at hand than the Lie algebra corresponding to words. Another possibility in this direction is to use a graded Lie algebra generated by monomial vector fields. In our example, we consider the graded Lie algebra $\tilde{\mathfrak{g}} = \bigoplus_{n\geq 1} \mathfrak{g}_n$ of vector fields generated by the monomial vector fields

$$U\,\partial_Y, U\,\partial_Z, V\,\partial_V, V\,\partial_Y, X^3\,\partial_Y, X^2\,\partial_Y, X\,\partial_Y, Y\,\partial_V, Y\,\partial_X, Y\,\partial_Y, Z\,\partial_U,\ \partial_V,\ \partial_X,\ \partial_Y,$$

each of them belonging to $\mathfrak{g}_1$. That is, we may multiply each of the monomial vector fields above by a bookkeeping parameter ϵ, so that monomial vector fields affected by a n-th power of ϵ belongs to $\mathfrak{g}_n$. In that case, the map $\Psi : \tilde{\mathfrak{g}} \to \mathrm{Der}(\mathscr{C})$ corresponds to replacing ϵ by 1. With the help of this graded Lie algebra a symbolic package in a laptop computer may carry the computations necessary to perform n-order averaging up to $n = 16$, while, as mentioned above, with words we could not go beyond $n = 8$.

4.4.3 Summary

The technique in [8] or [28] summarized in Sect. 3 averages oscillatory differential systems like (7) by first reformulating them in an abstract form (11) that is integrated in the group of characters $\mathscr{G}$ of the shuffle Hopf algebra. The solution of the abstract problem is then averaged and the result transferred back to the original system. While the technique is completely general, its computational complexity grows very quickly with the required accuracy. We have just seen that, by working with

alternative Hopf algebras, it is possible to diminish the computational cost and achieve substantially higher orders of accuracy in a given computing environment.

4.5 Explicit Construction of the Coproduct Δ

In this subsection, we focus on determining, in a form suitable for actual computations, the coproduct Δ of the Hopf algebra constructed in Sect. 4.2 from a given graded Lie algebra $\tilde{\mathfrak{g}}$.

In the particular case where $\tilde{\mathfrak{g}}$ is the free Lie algebra generated by an alphabet $\mathscr{A}$, $U(\tilde{\mathfrak{g}})$ is isomorphic to the algebra $\mathbb{R}\langle\mathscr{A}\rangle$. It therefore possesses a basis (18) with $\mathscr{J}$ given by the set of words on the alphabet $\mathscr{A}$ (the operation $\star$ corresponds to the concatenation of words). The coproduct Δ of $\mathscr{H}$ expressed in that basis indexed by words is then the deconcatenation coproduct, which has a particularly simple form.

For an arbitrary graded Lie algebra $\tilde{\mathfrak{g}}$, the coproduct $\Delta : \mathscr{H} \to \mathscr{H} \times \mathscr{H}$ can be uniquely determined from the structure constants $\lambda^{i}_{i',i''}$ of a basis (15) of $\tilde{\mathfrak{g}}$. Recall that the Poincaré-Birkhoff-Witt (PBW) basis of $U(\tilde{\mathfrak{g}})$ is a basis (18) indexed by the set

$$\mathscr{J} = \{e\} \cup \{(i_1, \dots, i_m) \in \mathscr{I}^m \ : \ m \geq 1, \ i_1 \leq \cdots \leq i_m\}, \tag{22}$$

where, as above, $\mathscr{I}$ is the set of indices for the homogeneous basis of the graded Lie algebra $\tilde{\mathfrak{g}}$ and $\mathscr{I}^m$ is the product $\mathscr{I} \times \cdots \times \mathscr{I}$ (m-times). The empty index e is associated with the unit $\mathbb{1}$ of $U(\tilde{\mathfrak{g}})$, that is $Z_e = \mathbb{1}$. For $j = (i_1, \dots, i_m) \in \mathscr{J}$, the elements Z_j are defined by (17) scaled by the inverse of the product of some factorials. More precisely, $Z_j = 1/j!\, G_{i_1} \star \cdots \star G_{i_m}$, where $j! = m!$ if $i_1 = i_2 = \cdots = i_m$, and $j! = k!(i_{k+1}, \dots, i_m)!$ if $i_1 = \cdots = i_k < i_{k+1}$. It is well known [3] that the basis $\{u_j \ : \ j \in \mathscr{J}\}$ of $\mathscr{H}$ dual to the PWB basis of $U(\tilde{\mathfrak{g}})$ satisfies that $u_j = v_{i_1} \cdots v_{i_m}$ for $j = (i_1, \dots, i_m) \in \mathscr{J}$, where $v_i := u_{(i)}$ for each $i \in \mathscr{I}$. We thus have that, as an algebra, $\mathscr{H}$ is a polynomial algebra on the commuting indeterminates $\{v_i \ : \ i \in \mathscr{I}\}$, that is, the symmetric algebra $S(V)$ over the vector space V spanned by $\{v_i \ : \ i \in \mathscr{I}\}$.

Since $\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}$ is an algebra map, it is enough to determine $\Delta(v_i) \in \mathscr{H} \otimes \mathscr{H}$ for $i \in \mathscr{I}$ from the structure constants $\lambda^{i}_{i',i''}$. However, existing algorithms for that task are rather involved. Fortunately, there are bases of $U(\tilde{\mathfrak{g}})$ that are computationally more convenient than the PWB basis for our purposes. This may be illustrated for the Grossman-Larson graded Lie algebra $\tilde{\mathfrak{g}}$ considered in Sect. 4.4: it has a basis (15) indexed by the set $\mathscr{I} := \mathscr{T}$ of rooted trees decorated by the letters of the alphabet $\mathscr{A} = \{a, b, c, d\}$ providing a simple description of the Lie bracket in terms of grafting of rooted trees [17]. In that case, the index set (22) can be identified with the set $\mathscr{F}$ of forest of rooted trees over $\mathscr{A}$. As noted before, the corresponding commutative Hopf algebra $\mathscr{H}$ is the Connes-Kreimer Hopf algebra over the alphabet $\mathscr{A}$. The construction of $\mathscr{H}$ described above in terms of the PWB basis of $U(\tilde{\mathfrak{g}})$ realizes the Hopf algebra $\mathscr{H}$ as the polynomial algebra

on the commuting indeterminates $\{v_i \ : \ i \in \mathscr{T}\} \subset U(\tilde{\mathfrak{g}})^*$. However, the expresions of $\Delta(v_i)$ for $i \in \mathscr{T}$ in that representation of $\mathscr{H}$ is rather cumbersome, and fails to reflect the nice combinatorial nature of the coproduct of the Connes-Kreimer Hopf algebra.

The task of determining the commutative graded Hopf algebra $\mathscr{H}$ from the structure constants $\lambda^i_{i',i''}$ of a basis (15) of $\tilde{\mathfrak{g}}$ can be reformulated in terms of a graded Lie coalgebra [21] structure (V, δ) related to the graded Lie algebra $\tilde{\mathfrak{g}}$ as follows. Let $V = \bigoplus_{n\geq 1} V_n$ be a graded vector space with a homogeneous basis $\{v_i \ : \ i \in \mathscr{I}\}$, and consider the graded linear map $\delta : V \to V \otimes V$ defined by

$$\delta(v_i) = \sum_{i',i'' \in \mathscr{I}} \lambda^i_{i',i''} \, v_{i'} \otimes v_{i''}, \ \text{for } i \in \mathscr{I}. \tag{23}$$

Since the coefficients $\lambda^i_{i',i''}$ are the structure constants with respect to a basis (15) of the graded Lie algebra $\tilde{\mathfrak{g}}$, (V, δ) is by construction a graded Lie coalgebra. The dual map $\delta^* : V^* \otimes V^* \to V^*$ endows the linear dual V^* with a structure of Lie algebra such that $\tilde{\mathfrak{g}}$ is isomorphic to a Lie subalgebra of the Lie algebra V^*.

Now, our original task can be formulated as follows: find an algebra map $\Delta : S(V) \to S(V) \otimes S(V)$ satisfying the following two conditions:

- the coproduct Δ endows the symmetric algebra $S(V)$ with a graded connected Hopf algebra structure $\mathscr{H}$,
- the linear map $\hat{\delta} : V \to V \otimes V$ such that, for each $v \in V$, $\hat{\delta}(v)$ is the projection of $\Delta(v)$ onto $V \otimes V$ satisfies the relation

$$\delta = \hat{\delta} - \tau \circ \hat{\delta}, \tag{24}$$

  where $\tau : V \otimes V \to V \otimes V$ is defined by $\tau(v \otimes v') = v' \otimes v$.

Such an algebra map Δ is not unique, but the corresponding coalgebra structure on $S(V)$ is unique up to isomorphisms.[6]

Observe that here there is no need to assume that the homogeneous vector subspaces $\tilde{\mathfrak{g}}_n$ are finite dimensional. One only needs to assume that the Lie coproduct (i.e. Lie cobracket) δ is well defined, or in other words, that the structure constants $\lambda^i_{i',i''}$ of the Lie bracket with respect to a basis (15) are such that, for each $i \in \mathscr{I}$, the sum in (23) is well defined.

Notice also that the choice of the basis for V plays no role in this formulation in terms of the Lie coalgebra (V, δ). It can be shown that, for a given basis (15) of the Lie algebra $\tilde{\mathfrak{g}}$, each choice of Δ gives rise, after dualization, to a different basis (18) (indexed by the set (22)) of $U(\tilde{\mathfrak{g}})$.

We will next make use of the concept of pre-Lie algebra. (We refer to [19] for a survey on pre-Lie algebras.) Assume now that there exists a graded linear map $\hat{\delta} : V \to V \otimes V$ satisfying (24). According to Proposition 3.5.2. in [16] (see also

[6] It is actually the universal coenvelopping coalgebra [21] of the Lie coalgebra V.

Theorem 5.8 in [18]), if $(V, \hat{\delta})$ is a graded pre-Lie coalgebra (that is, $(V^*, \hat{\delta}^*)$ is a pre-Lie algebra) then there exists a graded algebra map $\Delta : S(V) \to S(V) \otimes S(V)$ satisfying the following conditions:

- the coproduct Δ endows the symmetric algebra $S(V)$ with a graded connected Hopf algebra structure $\mathscr{H}$,
- for each $v \in V$, $\Delta(v) - 1 \otimes v - v \otimes 1 \in S(V) \otimes V$,
- for each $v \in V$, $\hat{\delta}(v)$ is the projection to $V \otimes V$ of $\Delta(v)$.

(Actually, the converse also holds [16].) Furthermore such an algebra map Δ is uniquely determined by $\hat{\delta}$. In [16], a recursive procedure to determine $\Delta(v)$ for each $v \in S(V)$ in terms of the pre-Lie coproduct $\hat{\delta}$ is presented. In Theorem 1 below, we suggest an alternative recursive procedure.

It is worth mentioning that the dual basis of the basis of monomials $v_{i_1} \cdots v_{i_m}$ of $S(V)$ corresponding to the coproduct Δ uniquely determined by $\hat{\delta}$ is precisely the basis of the universal enveloping algebra $U(\tilde{\mathfrak{g}})$ of the pre-Lie algebra $\tilde{\mathfrak{g}}$ considered in [30].

Coming back to the Grossman-Larson graded Lie algebra $\tilde{\mathfrak{g}}$ of rooted trees over an alphabet $\mathscr{A}$, it is known that it is the free pre-Lie algebra over the set $\mathscr{A}$ [6]. The Lie algebra morphism $\Psi : \tilde{\mathfrak{g}} \to \mathrm{Der}(\mathscr{C})$ considered in Sect. 4.4.1 is actually the unique extension of (9) to a pre-Lie algebra morphism from the Grossman-Larson Lie algebra over the alphabet $\{a, b, c, d\}$ to the Lie algebra of derivations $\mathrm{Der}(\mathscr{C})$. The corresponding pre-Lie coproduct $\hat{\delta} : V \to V \otimes V$ can be nicely described in terms of all the splittings of the rooted tree in two parts by successively removing each of the edges. The coproduct $\Delta : V \to S(V) \otimes S(V)$ uniquely determined in Proposition 3.5.2 of [16] coincides with the Connes-Kreimer coproduct defined in terms of the so called admissible cuts of rooted trees and forests.

This construction of the Hopf algebra $\mathscr{H}$ from a pre-Lie coalgebra structure $(V, \hat{\delta})$ may seem rather restrictive. However, any graded Lie coalgebra $V = \bigoplus_{n \geq 1} V_n$ with Lie coproduct δ admits at least one pre-Lie coproduct $\hat{\delta}$ satisfying (24), as we will show later on.

Let $\mathscr{H} = S(V)$ be the graded connected commutative Hopf algebra uniquely determined by a given pre-Lie coalgebra $(V, \hat{\delta})$, with coproduct $\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}$ and antipode $S : \mathscr{H} \to \mathscr{H}$. We define the grading operator $\rho : \mathscr{H} \to \mathscr{H}$ given by $\rho(u) = n\, v$ if u belongs to the graded component $\mathscr{H}_n$. Furthermore, we define the derivation $\partial : \mathscr{H}^* \to \mathscr{H}^*$ as the dual map of ρ, i.e. $\langle \partial(\gamma), u \rangle = \langle \gamma, \rho(u) \rangle$ for each $\gamma \in \mathscr{H}^*$ and each $u \in \mathscr{H}$.

We also use the (generalized) Dynkin operator D as considered in [12, 23] for graded connected commutative Hopf algebras. (See [20, 31] and references therein for the generalized Dynkin operator in the cocommutative case.) The Dynkin operator $D : S(V) \to V$ of $\mathscr{H} = S(V)$ is the convolution $D := \rho * S$ (in the references above, D is actually defined as $S * \rho$) of the antipode and the grading operator, that is,

$$D := \mu_{\mathscr{H}} \circ (\rho \otimes S) \circ \Delta,$$

where $\mu_{\mathscr{H}}: \mathscr{H} \otimes \mathscr{H} \to \mathscr{H}$ is the multiplication map of the algebra $\mathscr{H} = S(V)$. It is not difficult to check that the convolution of D with the identity $\mathrm{id}_{\mathscr{H}}$ in $\mathscr{H}$ coincides with ρ, that is,

$$\rho = \mu_{\mathscr{H}} \circ (D \otimes \mathrm{id}_{\mathscr{H}}) \circ \Delta. \tag{25}$$

The Dynkin operator has the property that

$$D(u) = 0 \quad \text{for all} \quad u \in V^2 S(V), \tag{26}$$

and we thus have that, for each $v \in V$,

$$(D \otimes \mathrm{id}_{\mathscr{H}}) \circ \Delta(v) = D(v) \otimes 1 + (D \otimes \mathrm{id}_V) \circ \hat{\delta}(v). \tag{27}$$

This in turn implies, together with (25) the following result, which allows to inductively determine $D(v)$ for each $v \in V$ in terms of the pre-Lie coproduct $\hat{\delta}$.

Lemma 1 *For each $v \in V$,*

$$D(v) = \rho(v) - \mu_{\mathscr{H}} \circ (D \otimes \mathrm{id}_V) \circ \hat{\delta}(v). \tag{28}$$

Theorem 1 *Let $(V, \hat{\delta})$ be a graded pre-Lie coalgebra, and consider the linear map $D : S(V) \to V$ determined by (26) and (28). The symmetric algebra $\mathscr{H} = S(V)$ becomes a graded connected Hopf algebra with the coproduct Δ determined as the unique graded algebra map $\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}$ such that*

$$\Delta(v) = 1 \otimes v + v \otimes 1 + \bar{\Delta}(v), \quad v \in V, \tag{29}$$

where the linear map $\bar{\Delta} : V \to \mathscr{H} \otimes V$ is uniquely determined by the identity

$$(\rho \otimes \mathrm{id}_V) \circ \bar{\Delta} = (D - \rho) \otimes 1 + (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D \otimes \Delta) \circ \hat{\delta}.$$

Proof From (29) one has that

$$(\rho \otimes \mathrm{id}_V) \circ \bar{\Delta} = -(\rho(v) \otimes 1) + (\rho \otimes \mathrm{id}_V) \circ \Delta$$

Application of (25), the coassociativity of Δ, and (27) lead to

$$\begin{aligned}
(\rho \otimes \mathrm{id}_V) \circ \Delta &= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (((D \otimes \mathrm{id}_{\mathscr{H}}) \circ \Delta) \otimes \mathrm{id}_V) \circ \Delta(v) \\
&= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D \otimes \mathrm{id}_{\mathscr{H}} \otimes \mathrm{id}_{\mathscr{H}}) \circ (\Delta \otimes \mathrm{id}_V) \circ \Delta(v) \\
&= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D \otimes \mathrm{id}_{\mathscr{H}} \otimes \mathrm{id}_{\mathscr{H}}) \circ (\mathrm{id}_{\mathscr{H}} \otimes \Delta) \circ \Delta(v) \\
&= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (\mathrm{id}_V \otimes \Delta) \circ (D \otimes \mathrm{id}_{\mathscr{H}}) \circ \Delta(v)
\end{aligned}$$

$$= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (\mathrm{id}_V \otimes \Delta) \circ (D(v) \otimes 1 + (D \otimes \mathrm{id}_V) \circ \hat{\delta}(v))$$
$$= (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D(v) \otimes 1 \otimes 1) + (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D \otimes \Delta) \circ \hat{\delta}(v)$$
$$= (D(v) \otimes 1) + (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (D \otimes \Delta) \circ \hat{\delta}(v).$$

□

Given a graded Lie coalgebra (V, δ), consider the graded linear map $\hat{\delta} : V \to V \otimes V$ determined in terms of δ by

$$\rho \circ \hat{\delta} = (\mathrm{id}_V \otimes \rho) \circ \delta, \tag{30}$$

that is,

$$\hat{\delta}(v_i) = \sum_{i', i'' \in \mathscr{I}} \frac{|i''|}{|i|} \lambda^i_{i', i''} \, v_{i'} \otimes v_{i''}, \quad \text{for } i \in \mathscr{I}.$$

Clearly, (24) holds, and it is not difficult to check that $(V, \hat{\delta})$ is a pre-Lie coalgebra, or equivalently, that the binary operation $\triangleright : V^* \otimes V^* \to V^*$ obtained by dualizing the coproduct $\hat{\delta}$ endows V^* with a structure of graded pre-Lie algebra. Indeed, for each $\beta', \beta'' \in V^*$,

$$\beta' \triangleright \beta'' = \partial^{-1}[\partial(\beta'), \beta''], \quad \beta', \beta'' \in \tilde{\mathfrak{g}},$$

where ∂^{-1} denotes the inverse of the restriction to V^* of ∂, so that $[\beta', \beta''] = \beta' \triangleright \beta'' - \beta'' \triangleright \beta'$.

Theorem 2 *Let (V, δ) be a graded Lie coalgebra. The symmetric algebra $\mathscr{H} = S(V)$ becomes a graded connected Hopf algebra with the coproduct Δ determined as the unique graded algebra map $\Delta : \mathscr{H} \to \mathscr{H} \otimes \mathscr{H}$ such that (29) holds and the linear map $\bar{\Delta} : V \to \mathscr{H} \otimes V$ is uniquely determined by the relation*

$$(\rho \otimes \mathrm{id}_V) \circ \bar{\Delta} = (\mu_{\mathscr{H}} \otimes \mathrm{id}_V) \circ (\rho \otimes \Delta) \circ \hat{\delta},$$

where $\hat{\delta} : V \to V \otimes V$ is determined by (30).

Proof The definition (30) of $\hat{\delta}$ and the assumption of (V, δ) being a Lie coalgebra implies that the identity (28) holds with $D(v) := \rho(v)$. The result then follows from Theorem 1. □

Finally, we provide the dual basis (18) (indexed by the set (22)) of the basis of monomials $v_{i_1} \cdots v_{i_m}$ of $S(V)$ corresponding to the graded connected commutive Hopf algebra structure on $S(V)$ determined in Theorem 2.

We first need some notation.

- Given $i \in \mathscr{I}$, we write $|i| = n$ if $i \in \mathscr{I}_n$. For $j = (i_1, \ldots, i_m) \in \mathscr{J}$, we set $|j| = |i_1| + \cdots + |i_m|$. We also write $|e| = 0$.

- Given $j = (i_1, \ldots, i_m) \in \mathscr{J}$ and $i \in \mathscr{I}$, we write $i \in j$ if $i \in \{i_1, \ldots, i_m\}$, and, in that case, we denote as $(j \backslash i)$ the element of $\mathscr{J}$ obtained by removing from $j = (i_1, \ldots, i_m)$ one occurrence of i. In particular, if $j = (i)$, then $(j \backslash i) = e$.

For $j = (i_1, \ldots, i_m) \in \mathscr{J}$, we set:

$$Z_j = \frac{|i|}{|j|} \sum_{i \in j} G_i \star Z_{(j \backslash i)}. \tag{31}$$

Observe that $Z_{(i)} = G_i$ for all $i \in \mathscr{I}$.

Theorem 3 *The set* (18) *of elements of* $U(\tilde{\mathfrak{g}})$ *given by* (31) *is a basis of* $U(\tilde{\mathfrak{g}})$ *dual to the basis of monomials* $u_j = v_{i_1} \cdots v_{i_m}$ *for* $j = (i_1, \ldots, i_m) \in \mathscr{J}$ *of the Hopf algebra determined in Theorem* 2.

Proof For each character $\alpha \in \mathscr{G}$ of $\mathscr{H}$, it holds [23]

$$\langle \partial(\alpha) \star \alpha^{-1}, u \rangle = \langle \alpha, D(u) \rangle \quad \text{for all} \quad u \in \mathscr{H}.$$

Since $D(v) = \rho(v)$ for all $v \in V$,

$$\langle \partial(\alpha) \star \alpha^{-1}, v \rangle = \langle \alpha, \rho(v) \rangle.$$

The later is equivalent to

$$\partial \left(\sum_{j \in \mathscr{J}} \langle \alpha, u_j \rangle \, Z_j \right) = \partial \left(\sum_{i \in \mathscr{I}} \langle \alpha, v_i \rangle \, G_i \right) \star \left(\sum_{j \in \mathscr{J}} \langle \alpha, u_j \rangle \, Z_j \right).$$

One finally arrives to (31) by expanding the right-hand side of that identity and equating terms. □

5 Perturbed Problems

As we saw in Sect. 3, the use of the shuffle Hopf algebra to average oscillatory problems in Euclidean space leads to word series expansions. *Extended word series*, introduced in [27], are a generalization of word series which appear in a natural way when solving some problems by means of the techniques we are studying. These include the reduction to normal form of continuous or discrete dynamical systems [26, 27], the analysis of splitting algorithms of perturbed integrable problems [27], the computation of formal invariants of perturbed Hamiltonian problems [26] and averaging of perturbed problems [28]. We now study the extension of these

techniques to scenarios where the shuffle Hopf algebra is replaced by other Hopf algebras.

We consider the situation where in the initial value problem (3), $F(t)$ is a perturbation $F(t) = F^0 + \tilde{F}(t)$ of a derivation $F^0 \in \mathrm{Der}(\mathscr{C})$ with a well defined exponential curve $\exp(t\, F^0)$ in $\mathrm{Aut}(\mathscr{C})$. If the solution $X(t)$ of the given problem

$$\frac{d}{dt}X(t) = X(t)(F_0 + \tilde{F}(t)), \quad X(0) = I, \tag{32}$$

exists, then it may be written as $X(t) = Y(t)\exp(t\, F_0)$, where the curve $Y : \mathbb{R} \to \mathrm{Aut}(\mathscr{C})$ is the solution of the initial value problem

$$\frac{d}{dt}Y(t) = Y(t)\exp(t\, F_0)\tilde{F}(t)\exp(-t\, F_0), \quad Y(0) = I. \tag{33}$$

5.1 Algebraic Framework for Perturbed Problems

We assume that there exist a graded Lie algebra

$$\bigoplus_{n\geq 0} \mathfrak{g}_n \tag{34}$$

with finite-dimensional homogeneous subspaces $\mathfrak{g}_n$, and a Lie group $\mathscr{G}_0$ with Lie algebra $\mathfrak{g}_0$ such that the exponential map $\exp : \mathfrak{g}_0 \to \mathscr{G}_0$ is bijective; we observe that $\mathfrak{g}_0$ and

$$\tilde{\mathfrak{g}} = \bigoplus_{n\geq 1} \mathfrak{g}_n,$$

are respectively a Lie subalgebra and a Lie ideal of (34).

Under such assumptions, one can prove that there exists an action $\cdot$ of the group $\mathscr{G}_0$ on the Lie algebra $\tilde{\mathfrak{g}}$ that is homogeneous of degree 0 (i.e. its restriction to each $\mathfrak{g}_n$ is an action on $\mathfrak{g}_n$), and such that, for arbitrary $\tilde{\beta} \in \bigoplus_{n\geq 1} \mathfrak{g}_n$ and $\beta_0 \in \mathfrak{g}_0$, $\alpha_0(t) = \exp(t\, \beta_0)$,

$$\frac{d}{dt}\left(\alpha_0(t) \cdot \tilde{\beta}\right) = [\beta_0, \alpha_0(t) \cdot \tilde{\beta}]. \tag{35}$$

We consider the commutative graded connected Hopf algebra $\mathscr{H} = \bigoplus_{n\geq 0} \mathscr{H}_n$ associated with the graded Lie algebra $\tilde{\mathfrak{g}}$, its group of characters $\mathscr{G} \subset \mathscr{H}^*$, and its Lie algebra of infinitesimal characters $\mathfrak{g} \subset \mathscr{H}^*$.

For $\beta \in \mathfrak{g}_0$, $\mathrm{ad}_\beta = [\beta, \cdot]$ is a derivation of (homogeneous degree 0 of) the graded Lie algebra (34). Its restriction to $\tilde{\mathfrak{g}}$ is also a derivation of the Lie subalgebra $\tilde{\mathfrak{g}}$. This derivation can be extended to a derivation of the Lie algebra $\mathfrak{g}$ of infinitesimal characters of $\mathscr{H}$. Hence, one can construct the semidirect sum Lie algebra

$$\bar{\mathfrak{g}} := \mathfrak{g} \oplus_S \mathfrak{g}_0 \supset \bigoplus_{n\geq 0} \mathfrak{g}_n.$$

More specifically, given $\bar{\beta} = \beta_0 + \beta \in \bar{\mathfrak{g}}$ and $\bar{\beta}' = \beta_0' + \beta' \in \bar{\mathfrak{g}}$ (where $\beta_0, \beta_0' \in \mathfrak{g}_0$ and $\beta, \beta' \in \tilde{\mathfrak{g}}$), then

$$[\bar{\beta}, \bar{\beta}'] = [\beta_0, \beta_0'] + (\mathrm{ad}_{\beta_0}\beta' - \mathrm{ad}_{\beta_0'}\beta + [\beta, \beta']),$$

The action of $\mathscr{G}_0$ on $\tilde{\mathfrak{g}}$ can be extended to an action of $\mathscr{G}_0$ on $U(\tilde{\mathfrak{g}})$ and from that to an action on $\mathscr{H}^*$. In particular, this defines an action of $\mathscr{G}_0$ on $\mathscr{G}$, which allows us to consider the semidirect product group

$$\bar{\mathscr{G}} := \mathscr{G} \ltimes \mathscr{G}_0.$$

More specifically, let $(\alpha, \alpha_0), (\alpha', \alpha_0') \in \bar{\mathscr{G}}$ (where $\alpha_0, \alpha_0' \in \mathscr{G}_0$ and $\alpha, \alpha' \in \mathscr{G}$), then the product law $\circ$ in $\bar{\mathscr{G}}$ is defined in terms of the action $\cdot$ and the product laws $\circ$ and $\star$ of $\mathscr{G}_0$ and $\mathscr{G}$ respectively as

$$(\alpha, \alpha_0) \circ (\alpha', \alpha_0') = (\alpha \star (\alpha_0 \cdot \alpha), \alpha_0 \circ \alpha_0').$$

We identify $(\mathbb{1}, \mathscr{G}_0)$ with $\mathscr{G}_0$, and $(\mathscr{G}, \mathrm{id}_0)$ with $\mathscr{G}$ (here id_0 denotes the neutral element in the Lie group $\mathscr{G}_0$); then we write the elements $(\alpha, \alpha_0) \in \bar{\mathscr{G}}$ as $\alpha \circ \alpha_0$. In particular, $\alpha_0 \cdot \alpha = \alpha_0 \circ \alpha \circ \alpha_0^{-1}$. We denote as id the identity element in $\bar{\mathscr{G}}$.

Given a smooth curve $\bar{\alpha} : \mathbb{R} \to \bar{\mathscr{G}}$ such that $\bar{\alpha}(0) = \mathrm{id}$, its derivative at $t = 0$ is

$$\left.\frac{d}{dt}\bar{\alpha}(t)\right|_{t=0} := \left.\frac{d}{dt}\alpha(t)\right|_{t=0} + \left.\frac{d}{dt}\alpha_0(t)\right|_{t=0},$$

with $\bar{\alpha}(t) = \alpha(t) \circ \alpha_0(t)$, where for all $t \in \mathbb{R}$, $\alpha(t) \in \mathscr{G}$, $\alpha_0(t) \in \mathscr{G}_0$. This can be used to define the adjoint representation $\mathrm{Ad} : \bar{\mathscr{G}} \to \mathrm{Aut}(\bar{\mathfrak{g}})$. In particular, for $\alpha_0 \in \mathscr{G}_0$, $\beta \in \mathfrak{g}$, $\mathrm{Ad}_{\alpha_0}\beta = \alpha_0 \cdot \beta$.

The exponential map $\exp : \bar{\mathfrak{g}} \to \bar{\mathscr{G}}$ is defined as follows: given $\bar{\beta} = \beta_0 + \beta \in \bar{\mathfrak{g}}$, then $\exp(\beta_0 + \beta) := \alpha(1) \circ \exp(\beta_0)$, where $\alpha(t) \in \mathscr{G}$ is the solution of (21) with $\beta(t)$ replaced by $\exp(t\beta_0) \cdot \beta$. With this definition, $\{\exp(t\,(\beta_0 + \beta)) \,:\, t \in \mathbb{R}\}$ is a one-parameter subgroup of $\bar{\mathscr{G}}$, and $(d/dt)\exp(t\,(\beta_0 + \beta))\big|_{t=0} = \beta_0 + \beta$.

In general $\exp : \bar{\mathfrak{g}} \to \bar{\mathscr{G}}$ is not surjective [26]. Given $\bar{\alpha} = \alpha \circ \exp(\beta_0) \in \bar{G}$, there exists $\beta \in \mathfrak{g}$ such that $\bar{\alpha} = \exp(\beta_0 + \beta)$ if, for each $n \geq 1$, the restriction to $\mathfrak{g}_n$ of $\int_0^1 \mathrm{Ad}_{\exp(t\beta_0)}dt$ is invertible. (The importance of this hypothesis will be illustrated in Sect. 5.3 below.)

5.2 Back to Perturbed Differential Equations

We now consider a perturbed operator differential equation (32), and assume that there exist a Lie algebra homomorphism

$$\Psi : \bigoplus_{n\geq 0} \mathfrak{g}_n \to \mathrm{Der}(\mathscr{C}),$$

an element $\beta_0 \in \mathfrak{g}_0$ and a curve $\tilde{\beta}(t)$ in $\tilde{\mathfrak{g}}$ with $\Psi(\beta_0) = F_0$ and $\Psi(\tilde{\beta}(t)) = \tilde{F}(t)$. This together with (35) implies that

$$\Psi(\exp(t\,\beta_0)\cdot\tilde{\beta}(t)) = \exp(F^0)\tilde{F}(t)\exp(-F^0).$$

Equation (33) now reads

$$\frac{d}{dt}Y(t) = Y(t)\Psi(\alpha_0(t)\cdot\tilde{\beta}(t)), \quad Y(0) = I, \tag{36}$$

where $\alpha_0(t) = \exp(t\beta_0)$. The problem (36) can be formally solved with the techniques in the preceding section as

$$Y(t) = \sum_{j\in\mathscr{J}} \langle \alpha(t), u_j\rangle\, \Psi(Z_j),$$

where $\alpha : \mathbb{R} \to \mathscr{G}$ is the solution of

$$\frac{d}{dt}\alpha(t) = \alpha(t) \star (\alpha_0(t)\cdot\tilde{\beta}(t)), \quad \alpha(0) = \mathbb{1}.$$

Hence, a formal solution $X(t)$ of (32) is given by

$$X(t) = \left(\sum_{j\in\mathscr{J}} \langle \alpha(t), u_j\rangle\, \Psi(Z_j)\right)\exp(t\,\Psi(\beta_0)).$$

As expected, the map that sends each $\bar{\alpha} = \alpha\circ\exp(\beta_0) \in \bar{\mathscr{G}}$ to the formal automorphism

$$\left(\sum_{j\in\mathscr{J}} \langle \alpha, u_j\rangle\, \Psi(Z_j)\right)\exp(t\,\Psi(\beta_0))$$

behaves as a group homomorphism. Similarly, the map that sends each $\bar{\beta} = \beta_0 + \beta \in \bar{\mathfrak{g}}$ to the formal derivation $\Psi(\beta_0) + \sum_{j \in \mathscr{J}} \langle \beta, u_j \rangle \Psi(Z_j)$ behaves as a Lie algebra homomorphism. In addition, if $\exp(\beta_0 + \beta) = \alpha \circ \exp(\beta_0)$, then

$$\exp\left(\Psi(\beta_0) + \sum_{j \in \mathscr{J}} \langle \beta, u_j \rangle \Psi(Z_j)\right) = \left(\sum_{j \in \mathscr{J}} \langle \alpha, u_j \rangle \Psi(Z_j)\right) \exp(t\, \Psi(\beta_0).$$

The adjoint representation $\mathrm{Ad} : \bar{\mathscr{G}} \to \mathrm{Aut}(\bar{\mathfrak{g}})$ also translates as expected through the map Ψ, so that it can be used to apply changes of variables in operator differential equations of the form (32).

5.3 *Application: Modified Equations for Splitting Methods*

The material just presented may be applied to analyze numerical integrators of differential equations. We refer to [27] for a detailed study of the application of splitting integrators to the solution of perturbations of integrable problem; that study is based on the use of the shuffle Hopf algebra/extended word series. Here we show how to proceed when the word series scenario is replaced by the more general framework developed in this section. For simplicity the attention is restricted to the well-known Strang splitting formula.

Assume that $\tilde{F}(t)$ is independent of t, and that $\exp(t\, \tilde{F})$ exists. Then, it is well known that the solution operator $X(t) = \exp(t\,(F_0 + \tilde{F}))$ can be approximated at $t = \tau, 2\tau, 3\tau, \ldots$ (τ is the time step) by $X(k\,\tau) \approx X_k \in \mathrm{Aut}(\mathscr{C})$, where $X_0 = I$ and

$$X_k = X_{k-1} \exp(\frac{\tau}{2} F_0) \exp(\tau\, \tilde{F}) \exp(\frac{\tau}{2} F_0), \quad k = 1, 2, 3, \ldots \tag{37}$$

If $F_0 = \Psi(\beta_0)$, $\tilde{F} = \Psi(\tilde{\beta})$ with $\beta_0 \in \mathfrak{g}_0$ and $\tilde{\beta} \in \tilde{\mathfrak{g}}$, then

$$X_k = X_{k-1} \left(\sum_{j \in \mathscr{J}} \langle \alpha^\tau, u_j \rangle \Psi(Z_j),\right),$$

where

$$\alpha^\tau = \exp(\tfrac{\tau}{2} \beta_0) \circ \exp(\tau\, \tilde{\beta}) \circ \exp(\tfrac{\tau}{2} \beta_0) = \exp(\tau\, \hat{\beta}^\tau) \circ \exp(\tau\, \beta_0),$$
$$\hat{\beta}^\tau = \exp(\tfrac{\tau}{2} \beta_0) \cdot \tilde{\beta}.$$

Let us assume that, for each $n \geq 1$, the restriction to $\mathfrak{g}_n$ of $\int_0^\tau \mathrm{Ad}_{\exp(t\,\beta_0)} dt$ is invertible.[7] In that case, there exists $\beta^\tau \in \mathfrak{g}$ such that $\alpha^\tau = \exp(\tau\,(\beta_0 + \beta^\tau))$, which back to operators, implies that $X_k \in \mathrm{Aut}(\mathscr{C})$ formally coincides with $X^\tau(k\tau)$, where $X^\tau(t)$ is the formal solution with $X^\tau(0) = I$ of the *modified equation*

$$\frac{d}{dt} X^\tau(t) = X^\tau(t) \left(F_0 + \sum_{i \in \mathscr{I}} \langle \beta^\tau, v_i \rangle \, \Psi(G_i) \right).$$

Modified equations are of course a powerful tool to analyze the performance of numerical integrators, see e.g. [36].

5.4 Hopf Algebraic Framework

To conclude the paper we shall briefly show how to cast the product group and product Lie algebra constructed in Sect. 5.1 as the group of characters and Lie algebra of infinitesimal characters of a suitable Hopf algebra. As in Sect. 4.5 we consider a (graded) subspace V of the commutative graded connected Hopf algebra $\mathscr{H} = \bigoplus_{n \geq 0} \mathscr{H}_n$ associated with the graded Lie algebra $\tilde{\mathfrak{g}} = \bigoplus_{n \geq 1} \mathfrak{g}_n$. Recall that V has a Lie coalgebra structure such that the Lie algebra $\mathfrak{g}$ of infinitesimal characters of $\mathscr{H}$ is isomorphic to the Lie algebra V^* dual to the Lie coalgebra V.

In addition to the hypotheses in Sect. 5.1, we assume that:

- $\mathscr{G}_0$ is an affine algebraic group with Lie algebra $\mathfrak{g}_0$. That is, $\mathscr{G}_0$ (respectively $\mathfrak{g}_0$) is the group of characters (respectively Lie algebra of infinitesimal characters) of a finitely generated commutative Hopf algebra $\bar{\mathscr{H}}_0$.
- The action $\cdot$ of the group $\mathscr{G}_0$ on $\mathfrak{g}$ can be obtained by dualizing a comodule map $\hat{\Delta}_0 : V \to \bar{\mathscr{H}}_0 \otimes V$. That is, given $\alpha_0 \in \mathscr{G}_0$, $\beta \in \mathfrak{g}$,

$$\langle \alpha_0 \cdot \beta, u \rangle = \langle \alpha_0 \otimes \beta, \hat{\Delta}_0(u) \rangle,$$

  for each $u \in \mathscr{H}$.

Then, $\bar{\mathscr{H}} := \bar{\mathscr{H}}_0 \oplus \mathscr{H}_1 \oplus \mathscr{H}_2 \oplus \cdots$ can be endowed with a commutative graded Hopf algebra structure in such a way that the resulting group of characters is the semidirect product group $\bar{\mathscr{G}}$, and the resulting Lie algebra of infinitesimal characters is the semidirect sum Lie algebra $\bar{\mathfrak{g}}$.

[7] In some cases, this assumption holds for most values of $\tau \in \mathbb{R}$, but fails for some particular values which gives rise to so-called numerical resonances [27].

Acknowledgements A. Murua and J.M. Sanz-Serna have been supported by projects MTM2013-46553-C3-2-P and MTM2013-46553-C3-1-P from Ministerio de Economía y Comercio, and MTM2016-77660-P(AEI/FEDER, UE) from Ministerio de Economía, Industria y Competitividad, Spain. Additionally A. Murua has been partially supported by the Basque Government (Consolidated Research Group IT649-13).

References

1. Alamo, A., Sanz-Serna, J.M.: A technique for studying strong and weak local errors of splitting stochastic integrators. SIAM J. Numer. Anal. **54**, 3239–3257 (2016)
2. Arnold, V.I.: Geometrical Methods in the Theory of Ordinary Differential Equations, 2nd edn. Springer, New York (1988)
3. Bourbaki, N.: Lie Groups and Lie Algebras. Springer, Berlin/New York (1989)
4. Butcher, J.C.: An algebraic theory of integration methods. Math. Comput. **26**, 79–106 (1972)
5. Castella, F., Chartier, Ph., Sauzeau, J.: A formal series approach to the center manifold theorem. J. Found. Comput. Math. (2017). https://doi.org/10.1007/s10208-017-9371-y
6. Chapoton, F., Livernet, M.: Pre-Lie algebras and the rooted trees operad. Int. Math. Res. Not. **8**, 395–408 (2001)
7. Chartier, P., Murua, A., Sanz-Serna, J.M.: Higher-order averaging, formal series and numerical integration I: B-series. Found. Comput. Math. **10**, 695–727 (2010)
8. Chartier, P., Murua, A., Sanz-Serna, J.M.: Higher-order averaging, formal series and numerical integration II: the quasi-periodic case. Found. Comput. Math. **12**, 471–508 (2012)
9. Chartier, P., Murua, A., Sanz-Serna, J.M.: A formal series approach to averaging: exponentially small error estimates. DCDS A **32**, 3009–3027 (2012)
10. Chartier, P., Murua, A., Sanz-Serna, J.M.: Higher-order averaging, formal series and numerical integration III: error bounds. Found. Comput. Math. **15**, 591–612 (2015)
11. Coccolo, M., Litak, G., Seoane, J.M., Sanjuán, M.A.F.: Energy harvesting enhancement by vibrational resonance. Int. J. Bifurcation Chaos **24**, 1430019 (7 pages) (2014)
12. Ebrahimi-Fard, K., Gracia-Bondia, J.M., Patras, F.: A Lie theoretic approach to renormalization. Commun. Math. Phys. **276**, 519–549 (2007)
13. Ecalle, J.: Les Fonctions résurgentes, vols. I, II, III. Publ. Math. Orsay (1981–1985)
14. Ecalle, J., Vallet, B.: Correction and linearization of resonant vector fields and diffeomorphisms. Math. Z. **229**, 249–318 (1998)
15. Fauvet, F., Menous, F.: Ecalle's arborification-coarborification transforms and Connes-Kreimer Hopf algebra. Ann. Sci. Ec. Norm. Super. **50**(1), 39–83 (2017)
16. Gan, W.L., Schedler, T.: The necklace Lie coalgebra and renormalization algebras. J. Noncommut. Geom. **2**, 195–214 (2008)
17. Grossman, R., Larson, R.G.: Hopf-algebraic structure of families of trees. J. Algebra **126**, 184–210 (1989)
18. Loday, J.-L., Ronco, M.: Combinatorial Hopf algebras. In: Blanchard, E., Ellwood, D., Khalkhali, M., Marcolli, M., Moscovici, H., Popa, S. (eds.) Quanta of Maths. Clay Mathematics Proceedings, vol. 11, pp. 347–383. American Mathematical Society, Providence (2010)
19. Manchon, D.: A short survey on pre-Lie algebras. In: Carey, A. (ed.) Noncommutative Geometry and Physics. ESI Lectures in Mathematics and Physics, pp. 89–102. European Mathematical Society, Zürich (2011)
20. Menous, F., Patras, F.: Logarithmic derivatives and generalized Dynkin operators. J. Algebraic Combin. **38**, 901–913 (2013)
21. Michelis, W.: Lie coalgebras. Adv. Math. **38**, 1–54 (1980)
22. Munthe-Kaas, H., Wright, W.: On the Hopf algebraic structure of Lie group integrators. Found. Comput. Math. **8**, 227–257 (2008)

23. Munthe-Kaas, H., Lundervold, A.: Hopf algebras of formal diffeomorphisms and numerical integration on manifolds. Contemp. Math. **539**, 295–324 (2011)
24. Murua, A.: The Hopf algebra of rooted trees, free Lie algebras and Lie series. Found. Comput. Math. **6**, 387–426 (2006)
25. Murua, A., Sanz-Serna, J.M.: Vibrational resonance: a study with high-order word-series averaging. Appl. Math. Nonlinear Sci. **1**, 239–146 (2016)
26. Murua A., Sanz-Serna, J.M.: Computing normal forms and formal invariants of dynamical systems by means of word series. Nonlinear Anal. **138**, 326–345 (2016)
27. Murua A., Sanz-Serna, J.M.: Word series for dynamical systems and their numerical integrators. Found. Comput. Math. **17**, 675–712 (2017)
28. Murua, A., Sanz-Serna, J.M.: Averaging and computing normal forms with word series algorithms. arXiv:1512.03601
29. Novelli, J.C., Paul, T., Sauzin, D., Thibon, J.Y.: Rayleigh-Schrödinger series and Birkhoff decomposition. Preprint arXiv:1608.01110 (2017)
30. Oudom, J.-M., Guin, D.: Sur l'algèbre enveloppante d'une algèbre prè-Lie. C. R. Math. Acad. Sci. Paris **340**, 331–336 (2005)
31. Patras, F., Reutenauer, C.: On Dynkin and Klyachko idempotents in graded bialgebras. Adv. Appl. Math. **28**, 560–579 (2002)
32. Paul, T., Sauzin, D.: Normalization in Lie algebras via mould calculus and applications. Preprint hal-01298047 (2016)
33. Paul, T., Sauzin, D.: Normalization in Banach scale of Lie algebras via mould calculus and applications. Preprint hal-05316595 (2016)
34. Reutenauer, C.: Free Lie Algebras. Clarendon Press, Oxford (1993)
35. Sanders, J.A., Verhulst, F., Murdock, J.: Averaging Methods in Nonlinear Dynamical Systems, 2nd edn. Springer, New York (2007)
36. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Chapman and Hall, London (1994)
37. Sanz-Serna, J.M., Murua, A.: Formal series and numerical integrators: some history and some new techniques. In: Guo, L., Zhi-Ming (eds.) Proceedings of the 8th International Congress on Industrial and Applied Mathematics (ICIAM 2015), pp. 311–331. Higher Education Press, Beijing (2015)
38. Sauzin, D.: Mould expansions for the saddle-node and resurgence monomials. In: Connes, A., Fauvet, F., Ramis, J.-P. (eds.) Renormalization and Galois Theories. IRMA Lectures in Mathematics and Theoretical Physics, vol. 15, pp. 83–163. European Mathematical Society, Zürich (2009)

Quantitative Limit Theorems for Local Functionals of Arithmetic Random Waves

Giovanni Peccati and Maurizia Rossi

Abstract We consider Gaussian Laplace eigenfunctions on the two-dimensional flat torus (arithmetic random waves), and provide explicit Berry-Esseen bounds in the 1-Wasserstein distance for the normal and non-normal high-energy approximation of the associated Leray measures and total nodal lengths, respectively. Our results provide substantial extensions (as well as alternative proofs) of findings by Oravecz et al. (Ann Inst Fourier (Grenoble) 58(1):299–335, 2008), Krishnapur et al. (Ann Math 177(2):699–737, 2013), and Marinucci et al. (Geom Funct Anal 26(3):926–960, 2016). Our techniques involve Wiener-Itô chaos expansions, integration by parts, as well as some novel estimates on residual terms arising in the chaotic decomposition of geometric quantities that can implicitly be expressed in terms of the coarea formula.

1 Introduction

The high-energy analysis of local geometric quantities associated with the *nodal set* of random Laplace eigenfunctions on compact manifolds has gained enormous momentum in recent years, in particular for its connections with challenging open problems in differential geometry (such as *Yau's conjecture* [19]), and with the striking cancellation phenomena detected by Berry in [2] – see the survey [18] for an overview of this domain of research up to the year 2012, and the Introduction of [13] for a review of recent literature. The aim of this paper is to prove quantitative limit theorems, in the high-energy limit, for *nodal lengths* and *Leray measures* (analogous to occupation densities at zero) of Gaussian Laplace eigenfunctions on

G. Peccati (✉)
Unité de Recherche en Mathématiques, Université du Luxembourg, Luxembourg, Luxembourg
e-mail: giovanni.peccati@uni.lu

M. Rossi
MAP5-UMR CNRS 8145, Université Paris Descartes, Paris, France
e-mail: maurizia.rossi@parisdescartes.fr

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_23

the two-dimensional flat torus. These random fields, first introduced by Rudnick and Wigman in [16], are called *arithmetic random waves* and are the main object discussed in the paper. The term 'arithmetic' emphasises the fact that, in the two dimensional case, the definition of toral eigenfunctions is inextricable from the problem of enumerating lattice points lying on circles with integer square radius.

Our results will allow us, in particular, to recover by an alternative (and mostly self-contained) approach the variance estimates from [11], as well as the non-central limit theorems proved in [13]. The core of our approach relies on the use of the Malliavin calculus techniques described in the monograph [14], as well as on some novel combinatorial estimates for residual terms arising in variance estimates obtained by chaotic expansions.

Although the analysis developed in the present paper focusses on a specific geometric model, we reckon that our techniques might be suitably modified in order to deal with more general geometric objects, whose definitions involve some variation of the area/coarea formulae; for instance, we believe that one could follow a route similar to the one traced below in order to deduce quantitative versions of the non-central limit theorems for phase singularities proved in [5], as well as to recover the estimates on the nodal variance of toral eigenfunctions and random spherical harmonics, respectively deduced in [16] and [17].

From now on, every random object is supposed to be defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with $\mathbb{E}$ denoting expectation with respect to $\mathbb{P}$.

1.1 Setup

As anticipated, in this paper we are interested in proving quantitative limit theorems for geometric quantities associated with Gaussian eigenfunctions of the Laplace operator $\Delta := \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$ on the flat torus $\mathbb{T} := \mathbb{R}^2/\mathbb{Z}^2$. In order to introduce our setup, we start by defining

$$S := \left\{ n \in \mathbb{Z} : n = a^2 + b^2, \ \text{ for some } a, b \in \mathbb{Z} \right\}$$

to be the set of all numbers that can be written as a sum of two integer squares [8]. It is a standard fact that the eigenvalues of $-\Delta$ are of the form $4\pi^2 n =: E_n$, where $n \in S$. The dimension $\mathscr{N}_n$ of the eigenspace $\mathscr{E}_n$ corresponding to the eigenvalue E_n coincides with the number $r_2(n)$ of ways in which n can be expressed as the sum of two integer squares (taking into account the order of summation). The quantity $\mathscr{N}_n = r_2(n)$ is a classical object in arithmetics, and is subject to large and erratic fluctuations: for instance, it grows *on average* as $\sqrt{\log n}$ but could be as small as 8 for an infinite sequence of prime numbers $p_n \equiv 1\ (4)$, or as large as a power of $\log n$ – see [10, Section 16.9 and Section 16.10] for a classical discussion, as well as [12] for recent advances. We also set

$$\Lambda_n := \left\{ \lambda = (\lambda_1, \lambda_2) \in \mathbb{Z}^2 : |\lambda|^2 := \lambda_1^2 + \lambda_2^2 = n \right\}$$

to be the class of all lattice points on the circle of radius $\sqrt{n}$ (its cardinality $|\Lambda_n|$ equals $\mathcal{N}_n$). Note that Λ_n is invariant w.r.t. rotations around the origin by $k \cdot \pi/2$, where k is any integer. An orthonormal basis $\{e_\lambda\}_{\lambda \in \Lambda_n}$ for the eigenspace $\mathcal{E}_n$ is given by the complex exponentials

$$e_\lambda(x) := \exp\left(i 2\pi \langle \lambda, x \rangle\right), \quad x = (x_1, x_2) \in \mathbb{T}.$$

We now consider a collection (indexed by the set of frequencies $\lambda \in \Lambda_n$) of identically distributed standard complex Gaussian random variables $\{a_\lambda\}_{\lambda \in \Lambda_n}$, that we assume to be independent except for the relations $\overline{a_\lambda} = a_{-\lambda}$. We recall that, by definition, every a_λ has the form $a_\lambda = b_\lambda + i c_\lambda$, where b_λ, c_λ are i.i.d. real Gaussian random variables with mean zero and variance $1/2$. We define the *arithmetic random wave* [11, 13, 15] of order $n \in S$ to be the real-valued centered Gaussian function

$$T_n(x) := \frac{1}{\sqrt{\mathcal{N}_n}} \sum_{\lambda \in \Lambda_n} a_\lambda e_\lambda(x), \quad x \in \mathbb{T}; \tag{1}$$

from (1) it is easily checked that the covariance of T_n is given by, for $x, y \in \mathbb{T}$,

$$r_n(x, y) := \mathbb{E}[T_n(x) \cdot T_n(y)] = \frac{1}{\mathcal{N}_n} \sum_{\lambda \in \Lambda_n} \cos(2\pi \langle \lambda, x - y \rangle) =: r_n(x - y). \tag{2}$$

Note that $r_n(0) = 1$, i.e. $T_n(x)$ has unit variance for every $x \in \mathbb{T}$. Moreover, as emphasised in the right-hand side (r.h.s.) of (2), the field T_n is *stationary*, in the sense that its covariance (2) depends only on the difference $x - y$. From now on, without loss of generality, we assume that T_n is stochastically independent of T_m for $n \neq m$.

For $n \in S$, we will focus on the *zero set* $T_n^{-1}(0) = \left\{x \in \mathbb{T} : T_n(x) = 0\right\}$; recall that, according e.g. to [4], with probability one $T_n^{-1}(0)$ consists of the union of a finite number of rectifiable (random) curves, called *nodal lines*, containing a finite set of isolated singular points. In this manuscript, we are more specifically interested in the following two *local* functionals associated with the nodal set $T_n^{-1}(0)$:

1. the *Leray (or microcanonical) measure* defined as [15, (1.1)]

$$\mathcal{L}_n := \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \text{meas}\left\{x \in \mathbb{T} : |T_n(x)| < \varepsilon\right\}, \tag{3}$$

   where 'meas' stands for the Lebesgue measure on $\mathbb{T}$, and the limit is in the sense of convergence in probability;
2. the (total) *nodal length* $\mathcal{L}_n$, given by (see [11])

$$\mathcal{L}_n := \text{length}\left(T_n^{-1}(0)\right); \tag{4}$$

for technical reasons, we will sometimes need to consider *restricted nodal lengths*, that are defined as follows: for every measurable $Q \subset \mathbb{T}$,

$$\mathscr{L}_n(Q) := \text{length}\left(T_n^{-1}(0) \cap Q\right). \tag{5}$$

We observe that, in the jargon of stochastic calculus, the quantity $\mathscr{Z}_n$ corresponds to the *occupation density at zero* of T_n – see [9] for a classical reference on the subject.

As already discussed, our aim is to establish quantitative limit theorems for both $\mathscr{Z}_n$ and $\mathscr{L}_n$ *in the high-energy limit*, that is, when $\mathscr{N}_n \to +\infty$.

Notation Given two positive sequences $\{a_n\}_{n\in S}$, $\{b_n\}_{n\in S}$ we will write:

1. $a_n \ll b_n$, if there exists a finite constant $C > 0$ such that $a_n \leq C b_n$, $\forall n \in S$. Similarly, $a_n \ll_\alpha b_n$ (resp. $a_n \ll_{\alpha,\beta} b_n$) will mean that C depends on α (resp. α, β);
2. "$a_n \ll b_n$, as $\mathscr{N}_n \to +\infty$" (or equivalently "$a_n = O(b_n)$, as $\mathscr{N}_n \to +\infty$") if, for every subsequence $\{n\} \subset S$ such that $\mathscr{N}_n \to \infty$, the ratio a_n/b_n is asymptotically bounded. Similarly, "$a_n \ll_\alpha b_n$, as $\mathscr{N}_n \to +\infty$", (resp. "$a_n \ll_{\alpha,\beta} b_n$, as $\mathscr{N}_n \to +\infty$") will mean that the bounding constant depends on α (resp. α, β);
3. $a_n \asymp b_n$ (resp. $a_n \asymp b_n$, $\mathscr{N}_n \to +\infty$) if both $a_n \ll b_n$ and $b_n \ll a_n$ (resp. $a_n \ll b_n$ and $b_n \ll a_n$, as $\mathscr{N}_n \to +\infty$) hold;
4. $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to +\infty$ (and analogously for subsequences);
5. $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to +\infty$ (and analogously for subsequences).

1.2 Previous Work

1.2.1 Leray Measure

The Leray measure in (3) was investigated by Oravecz, Rudnick and Wigman [15]. They found that [15, Theorem 4.1], for every $n \in S$,

$$\mathbb{E}\left[\mathscr{Z}_n\right] = \frac{1}{\sqrt{2\pi}}, \tag{6}$$

i.e. the expected Leray measure is constant, and moreover [15, Theorem 1.1],

$$\text{Var}(\mathscr{Z}_n) = \frac{1}{4\pi\mathscr{N}_n} + O\left(\frac{1}{\mathscr{N}_n^2}\right). \tag{7}$$

In particular, the asymptotic behaviour of the variance, as $\mathscr{N}_n \to +\infty$, is independent of the distribution of lattice points lying on the circle of radius $\sqrt{n}$.

1.2.2 Nodal Length

The expected nodal length was computed in [16] to be, for $n \in S$,

$$\mathbb{E}[\mathscr{L}_n] = \frac{1}{2\sqrt{2}}\sqrt{E_n}. \tag{8}$$

Computing the nodal variance is a subtler issue, and its asymptotic behaviour (in the high-energy limit) was fully characterized in [11] as follows. We start by observing that the set Λ_n induces a probability measure μ_n on the unit circle $\mathbb{S}^1$, given by $\mu_n := \frac{1}{\mathscr{N}_n}\sum_{\lambda\in\Lambda_n}\delta_{\lambda/\sqrt{n}}$, where δ_θ denotes the Dirac mass at $\theta \in \mathbb{S}^1$. One crucial fact is that, although there exists a density-1 subsequence $\{n_j\} \subset S$ such that $\mu_{n_j} \Rightarrow d\theta/2\pi$, as $j \to +\infty$,[1] there is an infinity of other weak-$*$ adherent points for the sequence $\{\mu_n\}_{n\in S}$ – see [12] for a partial classification. In particular, for every $\eta \in [-1, 1]$, there exists a subsequence $\{n_j\} \subset S$ (see [11, 12]) such that

$$\widehat{\mu_{n_j}}(4) \to \eta, \quad \text{as } j \to +\infty, \tag{9}$$

where, for a probability measure μ on the unit circle, the symbol $\widehat{\mu}(4)$ stands for the fourth Fourier coefficient $\widehat{\mu}(4) := \int_{\mathbb{S}^1}\theta^{-4}\,d\mu(\theta)$. Krishnapur, Kurlberg and Wigman in [11] found that, as $\mathscr{N}_n \to +\infty$,

$$\mathrm{Var}(\mathscr{L}_n) = c_n\frac{E_n}{\mathscr{N}_n^2}(1 + o(1)), \tag{10}$$

where $c_n := (1 + \widehat{\mu_n}(4)^2)/512$. Such a result is in stark contrast with (7): indeed, it shows that the asymptotic variance of the nodal length multiplicatively depends on the distribution of lattice points lying on the circle of radius $\sqrt{n}$, via the fluctuations of the squared Fourier coefficient $\widehat{\mu_n}(4)^2$; this also entails that the order of magnitude of the variance is $E_n/\mathscr{N}_n^2$, since the sequence $\{|\widehat{\mu_n}(4)|\}_n$ is bounded by 1. Plainly, in order to obtain an asymptotic behaviour in (10) that has no multiplicative corrections, one needs to extract a subsequence $\{n_j\} \subset S$ such that $\mathscr{N}_{n_j} \to +\infty$ and $|\widehat{\mu_{n_j}}(4)|$ converges to some $\eta \in [0, 1]$; in this case, one deduces that $\mathrm{Var}(\mathscr{L}_{n_j}) \sim c(\eta)E_{n_j}/\mathscr{N}_{n_j}^2$, where $c(\eta) := (1+\eta^2)/512$. Note that if $\mu_{n_j} \Rightarrow \mu$, then $\widehat{\mu_{n_j}}(4) \to \widehat{\mu}(4)$. By (9), the possible values of the constant $c(\eta)$ span therefore the whole interval $[1/512, 1/128]$.

The second order behavior of the nodal length was investigated in [13]. Let us define, for $\eta \in [0, 1]$, the random variable

$$\mathscr{M}_\eta := \frac{1}{2\sqrt{1+\eta^2}}\left(2 - (1+\eta)X_1^2 - (1-\eta)X_2^2\right), \tag{11}$$

[1] From now on, $\Rightarrow$ denotes weak-$*$ convergence of measures and $d\theta$ the uniform measure on $\mathbb{S}^1$.

where X_1, X_2 are i.i.d. standard Gaussians. Note that $\mathscr{M}_\eta$ is invariant in law under the transformation $\eta \mapsto -\eta$, so that if $\eta \in [-1, 0)$ we define $\mathscr{M}_\eta := \mathscr{M}_{-\eta}$.

Theorem 1.1 in [13] states that for $\{n_j\} \subset S$ such that $\mathscr{N}_{n_j} \to +\infty$ and $|\widehat{\mu_{n_j}}(4)| \to \eta$, as $j \to +\infty$, one has that

$$\widetilde{\mathscr{L}}_{n_j} \xrightarrow{d} \mathscr{M}_\eta, \tag{12}$$

where $\xrightarrow{d}$ denotes convergence in distribution and, for $n \in S$,

$$\widetilde{\mathscr{L}}_n := \frac{\mathscr{L}_n - \mathbb{E}\left[\mathscr{L}_n\right]}{\sqrt{\mathrm{Var}\left(\mathscr{L}_n\right)}} \tag{13}$$

is the normalized nodal length. Note that (12) is a non-universal and non central limit theorem: indeed, for $\eta \neq \eta'$ the (non Gaussian) laws of the random variables $\mathscr{M}_\eta$ and $\mathscr{M}_{\eta'}$ in (11) have different supports.

1.3 Main Results

The main purpose of this paper is to prove quantitative limit theorems for local functionals of nodal sets of arithmetic random waves, such as the Leray measure in (3) and the nodal length in (4). We will work with the 1-Wasserstein distance (see e.g. [14, §C] and the references therein). Given two random variables X, Y whose laws are μ_X and μ_Y, respectively, the Wasserstein distance between μ_X and μ_Y, written $d_{\mathrm{W}}(X, Y)$, is defined as

$$d_{\mathrm{W}}(X, Y) := \inf_{(A,B)} \mathbb{E}\left[|A - B|\right],$$

where the infimum runs over all pairs of random variables (A, B) with marginal laws μ_X and μ_Y, respectively. We will mainly use the dual representation

$$d_{\mathrm{W}}(X, Y) = \sup_{h \in \mathscr{H}} \left|\mathbb{E}\left[h(X) - h(Y)\right]\right|, \tag{14}$$

where $\mathscr{H}$ denotes the class of Lipschitz functions $h : \mathbb{R} \to \mathbb{R}$ whose Lipschitz constant is less or equal than 1. Relation (14) implies in particular that, if $d_{\mathrm{W}}(X_n, X) \to 0$, then $X_n \xrightarrow{d} X$ (the converse implication is false in general). Our first result is a *uniform* bound for the Wasserstein distance between the normalized Leray measure

$$\widetilde{\mathscr{Z}}_n := \frac{\mathscr{Z}_n - \mathbb{E}\left[\mathscr{Z}_n\right]}{\sqrt{\mathrm{Var}(\mathscr{Z}_n)}} \tag{15}$$

and a standard Gaussian random variable.

Theorem 1 *We have that, on S,*

$$d_W\left(\widetilde{\mathscr{Z}}_n, Z\right) \ll \mathscr{N}_n^{-1/2}, \tag{16}$$

where $\widetilde{\mathscr{Z}}_n$ *is defined in* (15)*, and* $Z \sim \mathcal{N}(0,1)$ *is a standard Gaussian random variable. In particular, if* $\{n_j\} \subset S$ *is such that* $\mathscr{N}_{n_j} \to +\infty$*, then* $\widetilde{\mathscr{Z}}_{n_j} \xrightarrow{d} Z$.

The following theorem deals with nodal lengths, providing a quantitative counterpart to the convergence result stated in (12).

Theorem 2 *As* $\mathscr{N}_n \to +\infty$*, one has that*

$$d_W\left(\widetilde{\mathscr{L}}_n, \mathscr{M}_\eta\right) \ll \mathscr{N}_n^{-1/4} \vee \left|\left|\widehat{\mu_n}(4)\right| - \eta\right|^{1/2}, \tag{17}$$

where $\widetilde{\mathscr{L}}_n$ *and* $\mathscr{M}_\eta$ *are defined, respectively, in* (12) *and* (11)*.*

Note that (17) entails the limit theorem (12): it is important to observe that, while the arguments exploited in [13] directly used the variance estimates in [11], the proof of (12) provided in the present paper is basically self-contained, except for the use of a highly non-trivial combinatorial estimate by Bombieri and Bourgain [3], appearing in our proof of Lemma 2 below – see Sect. 5. We also notice that the bound (16) for the Leray measure is uniform on S, whereas the bound (17) for the nodal length holds asymptotically, and depends on the angular distribution of lattice points lying on the circle of radius $\sqrt{n}$.

By combining the arguments used in the proofs of Theorems 1 and 2 with the content of [13, Section 4.2], one can also deduce the following multidimensional limit theorem, yielding in particular a form of *asymptotic dependence* between Leray measures and nodal lenghts.

Corollary 1 *Let* $\{n_j\} \subset S$ *be such that* $\mathscr{N}_{n_j} \to +\infty$ *and* $|\widehat{\mu_{n_j}}(4)| \to \eta \in [0,1]$*, then*

$$\left(\widetilde{\mathscr{Z}}_{n_j},\ \widetilde{\mathscr{L}}_{n_j}\right) \xrightarrow{d} \left(Z_1, \frac{q(Z)}{\sqrt{1+\eta^2}}\right),$$

where $Z = Z(\eta) = (Z_1, Z_2, Z_3, Z_4)$ *is a centered Gaussian vector with covariance matrix*

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{3+\eta}{8} & \frac{1-\eta}{8} & 0 \\ \frac{1}{2} & \frac{1-\eta}{8} & \frac{3+\eta}{8} & 0 \\ 0 & 0 & 0 & \frac{1-\eta}{8} \end{pmatrix},$$

and q *is the polynomial* $q(z_1, z_2, z_3, z_4) := 1 + z_1^2 - 2z_2^2 - 2z_3^2 - 4z_4^2$.

The details of the proof are left to the reader.

2 Outline of Our Approach

2.1 About the Proofs of the Main Results

In order to prove Theorems 1 and 2, we pervasively use *chaotic expansion* techniques (see Sect. 3). Since both $\mathscr{Z}_n$ in (3) and $\mathscr{L}_n$ in (4) are finite-variance functionals of a Gaussian field, they can be written as a series, converging in $L^2(\mathbb{P})$, whose terms can be explicitly found:

$$\mathscr{Z}_n = \sum_{q=0}^{+\infty} \mathscr{Z}_n[2q], \qquad \mathscr{L}_n = \sum_{q=0}^{+\infty} \mathscr{L}_n[2q]. \tag{18}$$

For each $q \geq 0$, the random variable $\mathscr{Z}_n[2q]$ (resp. $\mathscr{L}_n[2q]$) is the orthogonal projection of $\mathscr{Z}_n$ (resp. $\mathscr{L}_n$) onto the so-called *Wiener chaos* of order $2q$, that will be denoted by C_{2q}. Since $C_0 = \mathbb{R}$, we have $\mathscr{Z}_n[0] = \mathbb{E}[\mathscr{Z}_n]$ and $\mathscr{L}_n[0] = \mathbb{E}[\mathscr{L}_n]$; moreover, chaoses of different orders are orthogonal in $L^2(\mathbb{P})$.

2.1.1 Preliminaries to the Proof of Theorem 1

The main tool in our proof of Theorem 1 is the following result, that will be proved in Sect. 4.

Proposition 1 *For $n \in S$ (cf. (7))*

$$Var(\mathscr{Z}_n[2]) = \frac{1}{4\pi \mathscr{N}_n}. \tag{19}$$

Moreover, for every $K \geq 2$,

$$\sum_{q \geq K} Var(\mathscr{Z}_n[2q]) \ll_K \int_{\mathbb{T}} r_n(x)^{2K} dx \quad \text{on } S; \tag{20}$$

in particular, for $K = 2$,

$$\sum_{q \geq 2} Var(\mathscr{Z}_n[2q]) \ll \mathscr{N}_n^{-2}. \tag{21}$$

It should be noticed that Proposition 1 provides an alternative proof of (7) via chaotic expansions and entails also that, as $\mathscr{N}_n \to +\infty$,

$$\frac{\mathscr{Z}_n - \mathbb{E}[\mathscr{Z}_n]}{\sqrt{\mathrm{Var}(\mathscr{Z}_n)}} = \frac{\mathscr{Z}_n[2]}{\sqrt{\mathrm{Var}(\mathscr{Z}_n[2])}} + o_{\mathbb{P}}(1),$$

where $o_{\mathbb{P}}(1)$ denotes a sequence converging to 0 in probability. In particular, the Leray measure and its second chaotic component have the same asymptotic behavior, since different order Wiener chaoses are orthogonal. Let us now introduce some more notation. If $\sqrt{n}$ is an integer, we define

$$\Lambda_n^+ := \{\lambda = (\lambda_1, \lambda_2) \in \Lambda_n : \lambda_2 > 0\} \cup \{(\sqrt{n}, 0)\},$$

otherwise $\Lambda_n^+ := \{\lambda = (\lambda_1, \lambda_2) \in \Lambda_n : \lambda_2 > 0\}$. Note that $|\Lambda_n^+| = \mathcal{N}_n/2$ in both cases. The following elementary result is a further key element in the proof of Theorem 1.

Lemma 1 *For $n \in S$*

$$\mathscr{Z}_n[2] = -\frac{1}{\sqrt{2\pi}} \frac{1}{\mathcal{N}_n} \sum_{\lambda \in \Lambda_n^+} (|a_\lambda|^2 - 1).$$

Lemma 1, proven in Sect. 4 below, states that the second chaotic component is (proportional to) a sum of independent random variables. To conclude the proof of Theorem 1, note that we can write

$$d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, Z\right) \leq d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, \widetilde{\mathscr{Z}}_n[2]\right) + d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n[2], Z\right), \tag{22}$$

where $\widetilde{\mathscr{Z}}_n[2] := \mathscr{Z}_n[2]/\sqrt{\mathrm{Var}(\mathscr{Z}_n[2])}$. The first term on the right-hand side of (22) may be bounded by (21), whereas for the second term standard results apply, thanks to Lemma 1.

2.1.2 Preliminaries to the Proof of Theorem 2

The proof of Theorem 2 is similar to that one of Theorem 1, but contains a number of additional technical difficulties. In [13] it has been shown that $\mathscr{L}_n[2] = 0$ for every $n \in S$, and moreover that, as $\mathcal{N}_n \to +\infty$,

$$\mathrm{Var}(\mathscr{L}_n) \sim \mathrm{Var}(\mathscr{L}_n[4]), \tag{23}$$

by proving that the asymptotic variance of $\mathscr{L}_n[4]$ equals the r.h.s. of (10). The result stated in (23) and the orthogonality properties of Wiener chaoses entail that the fourth chaotic component and the total length have the same asymptotic behavior i.e., as $\mathcal{N}_n \to +\infty$,

$$\frac{\mathscr{L}_n - \mathbb{E}[\mathscr{L}_n]}{\sqrt{\mathrm{Var}(\mathscr{L}_n)}} = \frac{\mathscr{L}_n[4]}{\sqrt{\mathrm{Var}(\mathscr{L}_n[4])}} + o_{\mathbb{P}}(1), \tag{24}$$

where $o_{\mathbb{P}}(1)$ denotes a sequence converging to 0 in probability. Finally, in [13] it was shown that $\mathscr{L}_n[4]$ can be written as a *polynomial transform* of an asymptotically Gaussian random vector, so that the same convergence as in (12) holds when replacing the total nodal length with its fourth chaotic component.

Now let $h : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function and $\{n_j\}_j \subset S$ be such that $\mathscr{N}_{n_j} \to +\infty$ and $|\widehat{\mu_{n_j}}(4)| \to \eta$, as $j \to +\infty$. Bearing in mind (14) and (24), we write, by virtue of the triangle inequality,

$$\left|\mathbb{E}\left[h(\widetilde{\mathscr{L}}_{n_j}) - h(\mathscr{M}_\eta)\right]\right| \leq \mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_{n_j}) - h(\widetilde{\mathscr{L}}_{n_j}[4])\right|\right] + \left|\mathbb{E}\left[h(\widetilde{\mathscr{L}}_{n_j}[4]) - h(\mathscr{M}_\eta)\right]\right|, \tag{25}$$

where $\widetilde{\mathscr{L}}_{n_j}[4] := \mathscr{L}_{n_j}[4]/\sqrt{\mathrm{Var}(\mathscr{L}_{n_j}[4])}$. Let us deal with the first term on the r.h.s. of (25).

Proposition 2 *Let $h : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function and $\{n\} \subset S$ such that $\mathscr{N}_n \to +\infty$, then*

$$\mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_n) - h(\widetilde{\mathscr{L}}_n[4])\right|\right] \ll \mathscr{N}_n^{-1/4}. \tag{26}$$

In order to prove Proposition 2 in Sect. 5, we need to control the behavior of the variance tail $\sum_{q\geq 3} \mathrm{Var}(\mathscr{L}_n[2q])$, and this is done by proving the following general result

Lemma 2 *For every $K \geq 3$, on S we have*

$$\sum_{q\geq K} Var(\mathscr{L}_n[2q]) \ll_K E_n \int_{\mathbb{T}} r_n(x)^{2K}\, dx; \tag{27}$$

in particular, if $\mathscr{N}_n \to +\infty$,

$$\sum_{q\geq 3} Var(\mathscr{L}_n[2q]) \ll E_n\, \mathscr{N}_n^{-5/2}. \tag{28}$$

The proof of Lemma 2 is considerably more delicate than that of (20), see Sect. 5, and together with a precise investigation of the fourth chaotic component gives also an alternative proof of (10) via chaotic expansions.

For the second term on the r.h.s. of (25), recall from above that in [13] it was shown that $\mathscr{L}_n[4]$ can be written as a polynomial transform p of a random vector, say $W(n)$, which is asymptotically Gaussian. Let us denote by Z this limiting vector. Then, we can reformulate our problem as the estimation of the distributional

distance between $p(W(n_j))$ and $p(Z)$, the latter distributed as $\mathscr{M}_\eta$ in (11). To prove the following in Sect. 6 we can take advantage of some results in [6, 7].

Proposition 3 *Let $h : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function and let $|\widehat{\mu_{n_j}}(4)| \to \eta \in [0, 1]$, as $\mathscr{N}_{n_j} \to +\infty$, then*

$$\left| \mathbb{E}\left[h(\widetilde{\mathscr{L}}_{n_j}[4]) - h(\mathscr{M}_\eta) \right] \right| \ll \mathscr{N}_{n_j}^{-1/4} \vee ||\widehat{\mu_{n_j}}(4)| - \eta|^{1/2}. \tag{29}$$

Propositions 2 and 3 allow one to prove Theorem 2 in Sect. 6, bearing in mind (14) and (25). We now state and prove a technical result, which is an important tool for the proofs of Theorems 1 and 2.

2.2 A Technical Result

Some of the main bounds in our paper will follow from technical estimates involving pairs of cubes contained in the Cartesian product $\mathbb{T} \times \mathbb{T}$, that will be implicitly classified (for every fixed $n \in S$) according to the behaviour of the mapping $(x, y) \mapsto \mathbb{E}[T_n(x) \cdot T_n(y)] = r_n(x - y)$ appearing in (2).

Notation For every integer $M \geq 1$, we denote by $\mathscr{Q}(M)$ the partition of $\mathbb{T}$ obtained by translating in the directions k/M ($k \in \mathbb{Z}^2$) the square $Q_0 = Q_0(M) := [0, 1/M) \times [0, 1/M)$. Note that, by construction, $|\mathscr{Q}(M)| = M^2$.

Now we fix, for the rest of the paper, a small number $\epsilon \in (0, 10^{-3})$. The following statement unifies several estimates taken from [5, §6.1] (yielding Point 4), and [11, §4.1] (yielding Point 5) and [15, §6.1]. A sketch of the proof is provided for the sake of completeness.

Proposition 4 *There exists a mapping $M : S \to \mathbb{N} : n \mapsto M(n)$, as well as sets $G_0(n), G_1(n) \subset \mathscr{Q}(M(n)) \times \mathscr{Q}(M(n))$ with the following properties:*

1. *there exist constants $1 < c_1 < c_2 < \infty$ such that $c_1 E_n \leq M(n)^2 \leq c_2 E_n$ for every $n \in S$;*
2. *for every $n \in S$, $G_0(n) \cap G_1(n) = \emptyset$ and $G_0(n) \cup G_1(n) = \mathscr{Q}(M(n)) \times \mathscr{Q}(M(n))$;*
3. *$(Q, Q') \in G_0(n)$ if and only if for every $(x, y) \in Q \times Q'$, and for every choice of $i \in \{1, 2\}$ and $(i, j) \in \{1, 2\}^2$,*

$$|r_n(x - y)|,\ |\partial_i r_n(x - y)/\sqrt{n}\,|,\ |\partial_{i,j} r_n(x - y)/n| \leq \epsilon, \tag{30}$$

where $\partial_i r_n := \partial/\partial x_i\, r_n$ and $\partial_{i,j} := \partial/\partial_{x_i x_j}\, r_n$.

4. for every fixed $K \geq 2$, one has that

$$|G_1(n)| \ll_{\epsilon,K} E_n^2 \int_{\mathbb{T}} |r_n(x)|^{2K}\, dx; \tag{31}$$

5. adopting the notation (5), one has that

$$\mathrm{Var}(\mathscr{L}_n(Q_0)) \ll 1/E_n; \tag{32}$$

6. for every fixed $q \geq 2$, one has that

$$\int_{\hat{Q}_0} r_n(x)^{2q} dx \ll \frac{1}{2E_n(q+1)} \left(1 - \left(1 - \frac{E_n}{M(n)^2}\right)^{q+1}\right), \tag{33}$$

where $\hat{Q}_0 := Q_0 - Q_0$, and the constant involved in the above estimates is independent of q.

Proof (Sketch) The combination of Points 1–4 in the above statement corresponds to a slight variation of [5, Lemma 6.3]. Both estimates (32) and (33) follow from the fact that $\hat{Q}_0$ is contained in the union of four adjacent *positive singular cubes*, in the sense of [15, Definition 6.3].[2] Using such a representation of $\hat{Q}_0$, in order to prove (32) it is indeed sufficient to apply the same arguments as in [11, §4.1] for deducing that, defining the rescaled correlation 2-points function K_2 as in [11, formula (29)],

$$\mathrm{Var}(\mathscr{L}_n(Q_0)) = E_n \int_{Q_0}\int_{Q_0} K_2(x-y)dxdy \leq \frac{E_n}{M(n)^2}\int_{\hat{Q}_0} K_2(x)dx \ll \frac{1}{E_n}.$$

Finally, arguing as in [15, §6.5], we infer that $r_n(x)^2 \leq 1 - E_n\|x - x_0\|^2$, where $x_0 = (0,0)$ and the estimate holds for every $x \in \hat{Q}_0$, yielding in turn the relations

$$\int_{\hat{Q}_0} r_n(x)^{2q}\, dx \leq \int_{\|x-x_0\| \ll \frac{1}{M}} \left(1 - E_n\|x-x_0\|^2\right)^q dx \ll \int_0^{\frac{1}{M}} r(1 - E_n r^2)^q\, dr$$
$$= \frac{1}{2E_n}\frac{1}{q+1}\left(1 - \left(1 - \frac{E_n}{M(n)^2}\right)^{q+1}\right),$$

and therefore the desired conclusion. □

[2] Indeed, each one of the four cubes composing $\hat{Q}_0$ is such that its boundary contains the point $x_0 = (0,0)$, and the singularity in the sense of [15, Definition 6.3] follows by the continuity of trigonometric functions.

3 Local Functionals and Wiener Chaos

As mentioned in Sect. 2.1, for the proof of our main results we need the notion of Wiener-Itô chaotic expansions for non-linear functionals of Gaussian fields. In what follows, we will present it in a simplified form adapted to our situation; we refer the reader to [14, §2.2] for a complete discussion.

3.1 Wiener Chaos

Let ϕ denote the standard Gaussian density on $\mathbb{R}$ and $L^2(\mathbb{R}, \mathscr{B}(\mathbb{R}), \phi(t)dt) =: L^2(\phi)$ the space of square integrable functions on the real line w.r.t. the Gaussian measure $\phi(t)dt$. The sequence of normalized Hermite polynomials $\{(k!)^{-1/2}H_k\}_{k\geq 0}$ is a complete orthonormal basis of $L^2(\phi)$; recall [14, Definition 1.4.1] that they are defined recursively as follows: $H_0 \equiv 1$, and, for $k \geq 1$, $H_k(t) = tH_{k-1}(t) - H'_{k-1}(t)$, $t \in \mathbb{R}$. Recall now the definition of the arithmetic random waves (1), and observe that it involves a family of complex-valued Gaussian random variables $\{a_\lambda : \lambda \in \mathbb{Z}^2\}$ with the following properties: (i) $a_\lambda = b_\lambda + ic_\lambda$, where b_λ and c_λ are two independent real-valued centered Gaussian random variables with variance $1/2$; (ii) a_λ and $a_{\lambda'}$ are independent whenever $\lambda' \notin \{\lambda, -\lambda\}$, and (iii) $a_\lambda = \overline{a_{-\lambda}}$. Consider now the space of all real finite linear combinations of random variables ξ of the form $\xi = z\,a_\lambda + \overline{z}\,a_{-\lambda}$, where $\lambda \in \mathbb{Z}^2$ and $z \in \mathbb{C}$. Let us denote by $\mathbf{A}$ its closure in $L^2(\mathbb{P})$; it turns out that $\mathbf{A}$ is a real centered Gaussian Hilbert subspace of $L^2(\mathbb{P})$.

Definition 1 Let q be a nonnegative integer; the q-th *Wiener chaos* associated with $\mathbf{A}$, denoted by C_q, is the closure in $L^2(\mathbb{P})$ of all real finite linear combinations of random variables of the form

$$H_{p_1}(\xi_1) \cdot H_{p_2}(\xi_2) \cdots H_{p_k}(\xi_k)$$

for $k \geq 1$, where the integers $p_1, \ldots, p_k \geq 0$ satisfy $p_1 + \cdots + p_k = q$, and $(\xi_1, \ldots, \xi_k)$ is a standard real Gaussian vector extracted from $\mathbf{A}$ (note that, in particular, $C_0 = \mathbb{R}$).

It is well-known (see [14, §2.2]) that C_q and C_m are orthogonal in $L^2(\mathbb{P})$ whenever $q \neq m$, and moreover $L^2(\Omega, \sigma(\mathbf{A}), \mathbb{P}) = \bigoplus_{q\geq 0} C_q$; equivalently, every real-valued functional F of $\mathbf{A}$ can be (uniquely) represented in the form

$$F = \sum_{q=0}^{\infty} F[q], \tag{34}$$

where $F[q]$ is the orthogonal projection of F onto C_q, and the series converges in $L^2(\mathbb{P})$. Plainly, $F[0] = \mathbb{E}[F]$.

3.2 Chaotic Expansion of $\mathscr{Z}_n$

We can rewrite (3) as

$$\mathscr{Z}_n = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{\mathbb{T}} 1_{[-\varepsilon,\varepsilon]}(T_n(x))\, dx =: \lim_{\varepsilon \to 0} \mathscr{Z}_n^{\varepsilon}, \tag{35}$$

and hence formally represent the Leray measure as

$$\mathscr{Z}_n = \int_{\mathbb{T}} \delta_0(T_n(x))\, dx, \tag{36}$$

where δ_0 denotes the Dirac mass at $0 \in \mathbb{R}$. Let us now consider the sequence of coefficients $\{\beta_{2q}\}_{q \geq 0}$ defined as

$$\beta_{2q} := \frac{1}{\sqrt{2\pi}} H_{2q}(0), \tag{37}$$

where H_{2q} denotes the $2q$-th Hermite polynomial, as before. It can be seen as the sequence of coefficients corresponding to the (formal) chaotic expansion of the Dirac mass.

The following result concerns the chaotic expansion of the Leray measure in (36) and will be proved in the Appendix.

Lemma 3 *For $n \in S$, one has that $\mathscr{Z}_n \in L^2(\mathbb{P})$, and the chaotic expansion of $\mathscr{Z}_n$ is*

$$\mathscr{Z}_n = \sum_{q=0}^{+\infty} \mathscr{Z}_n[2q] = \sum_{q=0}^{+\infty} \frac{\beta_{2q}}{(2q)!} \int_{\mathbb{T}} H_{2q}(T_n(x))\, dx, \tag{38}$$

where β_{2q} is given in (37), and the convergence of the above series holds in $L^2(\mathbb{P})$.

3.3 Chaotic Expansion of $\mathscr{L}_n$

We recall now from [13] the chaotic expansion (34) for the nodal length. First, $\mathscr{L}_n$ in (4) admits the following integral representation

$$\mathscr{L}_n = \int_{\mathbb{T}} \delta_0(T_n(x)) |\nabla T_n(x)|\, dx, \tag{39}$$

where δ_0 still denotes the Dirac mass at $0 \in \mathbb{R}$ and ∇T_n the gradient of T_n; more precisely, $\nabla T_n = (\partial_1 T_n, \partial_2 T_n)$ with $\partial_i := \partial/\partial x_i$ for $i = 1, 2$. The integral in (39)

has to be interpreted in the sense that, for any sequence of bounded probability densities $\{g_k\}$ such that the associated probabilities weakly converge to δ_0, one has that $\int_{\mathbb{T}} g_k(T_n(x))|\nabla T_n(x)|\,dx \to \mathscr{L}_n$ in $L^2(\mathbb{P})$. A straightforward differentiation of the definition (1) of T_n yields, for $j = 1, 2$

$$\partial_j T_n(x) = \frac{2\pi i}{\sqrt{\mathscr{N}_n}} \sum_{(\lambda_1,\lambda_2)\in\Lambda_n} \lambda_j a_\lambda e_\lambda(x). \tag{40}$$

Hence the random fields T_n, $\partial_1 T_n$, $\partial_2 T_n$ viewed as collections of Gaussian random variables indexed by $x \in \mathbb{T}$ are all lying in $\mathbf{A}$, i.e. for every $x \in \mathbb{T}$ we have

$$T_n(x),\ \partial_1 T_n(x),\ \partial_2 T_n(x) \in \mathbf{A}.$$

It has been proved in [11] that the random variables $T_n(x)$, $\partial_1 T_n(x)$, $\partial_2 T_n(x)$ are independent for fixed $x \in \mathbb{T}$, and for $i = 1, 2$

$$\mathrm{Var}(\partial_i T_n(x)) = \frac{E_n}{2}. \tag{41}$$

We can write from (39), keeping in mind (41),

$$\mathscr{L}_n = \sqrt{\frac{E_n}{2}} \int_{\mathbb{T}} \delta_0(T_n(x))|\widetilde{\nabla} T_n(x)|\,dx, \tag{42}$$

with $\widetilde{\nabla} T_n := (\widetilde{\partial}_1 T_n, \widetilde{\partial}_2 T_n)$ and for $i = 1, 2$, $\widetilde{\partial}_i := \partial_i/\sqrt{E_n/2}$. Note that $\widetilde{\partial}_i T_n(x)$ has unit variance for every $x \in \mathbb{T}$.

Equation (39), or equivalently (42), explicitly represents the nodal length as a (finite-variance) non-linear functional of a Gaussian field. To recall its chaotic expansion, we need (37) and moreover have to introduce the collection of coefficients $\{\alpha_{2n,2m} : n, m \geq 1\}$, that is related to the Hermite expansion of the norm $|\cdot|$ in $\mathbb{R}^2$:

$$\alpha_{2n,2m} = \sqrt{\frac{\pi}{2}} \frac{(2n)!(2m)!}{n!m!} \frac{1}{2^{n+m}} p_{n+m}\left(\frac{1}{4}\right), \tag{43}$$

where for $N = 0, 1, 2, \ldots$ and $x \in \mathbb{R}$

$$p_N(x) := \sum_{j=0}^{N} (-1)^j \cdot (-1)^N \binom{N}{j} \frac{(2j+1)!}{(j!)^2} x^j,$$

$\frac{(2j+1)!}{(j!)^2}$ being the so-called "swinging factorial" restricted to odd indices. From [13, Proposition 3.2], we have for $q = 2$ or $q = 2m + 1$ odd ($m \geq 1$) $\mathscr{L}_n[q] \equiv 0$, and

for $q \geq 2$

$$\mathscr{L}_n[2q] = \sqrt{\frac{4\pi^2 n}{2}} \sum_{u=0}^{q} \sum_{k=0}^{u} \frac{\alpha_{2k,2u-2k}\beta_{2q-2u}}{(2k)!(2u-2k)!(2q-2u)!} \times$$
$$\times \int_{\mathbb{T}} H_{2q-2u}(T_n(x)) H_{2k}(\widetilde{\partial}_1 T_n(x)) H_{2u-2k}(\widetilde{\partial}_2 T_n(x))\, dx. \tag{44}$$

The Wiener-Itô chaotic expansion of $\mathscr{L}_n$ is hence

$$\mathscr{L}_n = \mathbb{E}[\mathscr{L}_n] + \sqrt{\frac{4\pi^2 n}{2}} \sum_{q=2}^{+\infty} \sum_{u=0}^{q} \sum_{k=0}^{u} \frac{\alpha_{2k,2u-2k}\beta_{2q-2u}}{(2k)!(2u-2k)!(2q-2u)!} \times$$
$$\times \int_{\mathbb{T}} H_{2q-2u}(T_n(x)) H_{2k}(\widetilde{\partial}_1 T_n(x)) H_{2u-2k}(\widetilde{\partial}_2 T_n(x))\, dx,$$

with convergence in $L^2(\mathbb{P})$.

3.3.1 Fourth Chaotic Components

In this part we investigate the fourth chaotic component $\mathscr{L}_n[4]$ (from (44) with $q = 2$), recalling also some facts from [13].

Consider, for $n \in S$, the four-dimensional random vector $W = W(n)$ given by

$$W(n) = \begin{pmatrix} W_1(n) \\ W_2(n) \\ W_3(n) \\ W_4(n) \end{pmatrix} := \frac{1}{\sqrt{\mathscr{N}_n/2}} \sum_{\lambda \in \Lambda_n^+} (|a_\lambda|^2 - 1) \begin{pmatrix} 1 \\ \lambda_1^2/n \\ \lambda_2^2/n \\ \lambda_1\lambda_2/n \end{pmatrix},$$

whose covariance matrix is

$$\Sigma_n = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{3+\widehat{\mu_n}(4)}{8} & \frac{1-\widehat{\mu_n}(4)}{8} & 0 \\ \frac{1}{2} & \frac{1-\widehat{\mu_n}(4)}{8} & \frac{3+\widehat{\mu_n}(4)}{8} & 0 \\ 0 & 0 & 0 & \frac{1-\widehat{\mu_n}(4)}{8} \end{pmatrix}, \tag{45}$$

see [13, Lemma 4.1]. Note that for every $n \in S$

$$W_2(n) + W_3(n) = W_1(n). \tag{46}$$

The following will be proved in the Appendix and is a finer version of [13, Lemma 4.2].

Lemma 4 *For every $n \in S$,*

$$\mathscr{L}_n[4] = \sqrt{\frac{E_n}{\mathscr{N}_n^2}} \frac{1}{\sqrt{512}} \left(W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2 + \frac{1}{2}\frac{1}{\mathscr{N}_n} \sum_{\lambda \in \Lambda_n} |a_\lambda|^4 \right), \tag{47}$$

and moreover,

$$Var(\mathscr{L}_n[4]) = \frac{E_n}{512\mathscr{N}_n^2} \left(1 + \widehat{\mu_n}(4)^2 + \frac{34}{\mathscr{N}_n} \right). \tag{48}$$

It is worth noticing that Lemma 2 and (48) immediately give an alternative proof of (10) via chaotic expansion.

We recall here from [13, Lemma 4.3] that, for $\{n_j\} \subseteq S$ such that $\mathscr{N}_{n_j} \to +\infty$ and $\widehat{\mu}_{n_j}(4) \to \eta \in [-1, 1]$, as $j \to \infty$, the following CLT holds:

$$W(n_j) \xrightarrow{d} Z = Z(\eta) = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix}, \tag{49}$$

where $Z(\eta)$ is a centered Gaussian vector with covariance

$$\Sigma = \Sigma(\eta) = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{3+\eta}{8} & \frac{1-\eta}{8} & 0 \\ \frac{1}{2} & \frac{1-\eta}{8} & \frac{3+\eta}{8} & 0 \\ 0 & 0 & 0 & \frac{1-\eta}{8} \end{pmatrix}. \tag{50}$$

The eigenvalues of Σ are $0, \frac{3}{2}, \frac{1-\eta}{8}, \frac{1+\eta}{4}$ and hence, in particular, Σ is singular. Moreover,

$$\frac{\mathscr{L}_{n_j}[4]}{\sqrt{\mathrm{Var}(\mathscr{L}_{n_j}[4])}} \xrightarrow{d} \mathscr{M}_{|\eta|},$$

where $\mathscr{M}_{|\eta|}$ is defined as in (11), see [13, Proposition 2.2].

4 Proof of Theorem 1

Note first that, from (37) and (38) for $q = 0$

$$\mathscr{Z}_n[0] = \beta_0 = \frac{1}{\sqrt{2\pi}},$$

cf. (6). Let us now focus on the second chaotic component of the Leray measure in (38), by proving Lemma 1.

Proof (Lemma 1) By (37) and (38) for $q = 1$, recalling that $H_2(t) = t^2 - 1$,

$$\mathscr{Z}_n[2] = -\frac{1}{2\sqrt{2\pi}} \int_{\mathbb{T}} (T_n(x)^2 - 1)\, dx.$$

Finally, (1) allows us to conclude the proof. □

We can now prove Proposition 1.

Proof (Proposition 1) From Lemma 1, straightforward computations based on independence yield that

$$\mathrm{Var}(\mathscr{Z}_n[2]) = \frac{1}{4\pi \mathscr{N}_n},$$

that is (19). We can rewrite (20) as

$$\sum_{q \geq K} \mathrm{Var}(\mathscr{Z}_n[2q]) = \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \int_{\mathbb{T}} r_n(x)^{2q}\, dx \ll \int_{\mathbb{T}} r_n(x)^{2K}\, dx \tag{51}$$

(note that the first equality in (51) is a direct consequence of (38), [14, Proposition 1.4.2] and stationarity of T_n). Our proof of the second equality in (51), which is (20), uses the content of Proposition 4. We can rewrite the middle term in (51), by stationarity of T_n, as

$$\begin{aligned}
\sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \int_{\mathbb{T}} r_n(x)^{2q}\, dx &= \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \int_{\mathbb{T}} \int_{\mathbb{T}} r_n(x-y)^{2q}\, dxdy \\
&= \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \sum_{(Q,Q') \in G_0(n)} \int_Q \int_{Q'} r_n(x-y)^{2q}\, dxdy \\
&\quad + \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \sum_{(Q,Q') \in G_1(n)} \int_Q \int_{Q'} r_n(x-y)^{2q}\, dxdy \\
&=: A(n) + B(n).
\end{aligned} \tag{52}$$

Using Point 3 in Proposition 4 one infers that

$$\begin{aligned}
A(n) &\leq \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \epsilon^{2q-2K} \sum_{(Q,Q') \in G_0(n)} \int_Q \int_{Q'} r_n(x-y)^{2K}\, dxdy \\
&\leq \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \epsilon^{2q-2K} \int_{\mathbb{T}} r_n(x)^{2K}\, dx.
\end{aligned} \tag{53}$$

It is easy to check that, since $\epsilon \in (0, 1)$, then

$$\sum_{q=1}^{+\infty} \frac{\beta_{2q}^2}{(2q)!} \epsilon^{2q} < \infty$$

(indeed, $\beta_{2q}^2/(2q)! \asymp 1/\sqrt{q}$, as $q \to \infty$), finally yielding

$$A(n) \ll_{\epsilon,K} \int_{\mathbb{T}} r_n(x)^{2K}\, dx. \tag{54}$$

Let us now focus on $B(n)$. For every pair $(Q, Q') \in G_1(n)$ and every $q \geq 1$, we can use Cauchy-Schwartz inequality and then exploit the stationarity of T_n to write

$$\begin{aligned}\int_Q \int_{Q'} r_n(x-y)^{2q}\, dxdy &= (2q)!^{-1}\mathbb{E}\left[\int_Q H_{2q}(T_n(x))dx \int_{Q'} H_{2q}(T_n(y))dy\right] \\ &\leq (2q)!^{-1}\mathrm{Var}\left(\int_{Q_0} H_{2q}(T_n(x))dx\right) = \int_{Q_0}\int_{Q_0} r_n(x-y)^{2q}\, dxdy \\ &\leq \int_{Q_0} dy \int_{\hat{Q}_0} r_n(x)^{2q}\, dx \ll \frac{1}{E_n}\int_{\hat{Q}_0} r_n(x)^{2q}\, dx,\end{aligned}$$

where the constant involved in the last estimate is independent of q. Using (31) and (33), one therefore deduces that

$$\begin{aligned}B(n) &\ll \int_{\mathbb{T}} r_n(x)^{2K}\, dx \times \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!}\frac{1}{q+1}\left(1-\left(1-\frac{E_n}{M^2}\right)^{q+1}\right) \\ &= \int_{\mathbb{T}} r_n(x)^{2K}\, dx \times \left(\sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!}\frac{1}{q+1} - \sum_{q=K}^{+\infty} \frac{\beta_{2q}^2}{(2q)!}\frac{1}{q+1}\left(1-\frac{E_n}{M^2}\right)^{q+1}\right).\end{aligned} \tag{55}$$

Since the series appearing in the above expression are both convergent, substituting (53) and (55) in (52), bearing in mind (51), we immediately have (20). To prove (21), it suffices to recall (from (2)) that for every integer $K \geq 1$

$$\int_{\mathbb{T}} r_n(x)^{2K}\, dx = \frac{|S_{2K}(n)|}{\mathcal{N}_n^{2K}}, \tag{56}$$

where

$$S_{2K}(n) = \{(\lambda_1, \lambda_2, \ldots, \lambda_{2K}) \in \Lambda_n^{2K} : \lambda_1 + \lambda_2 + \cdots + \lambda_{2K} = 0\}. \tag{57}$$

For $K = 2$, from [11] we have

$$|S_4(n)| = 3\mathscr{N}_n(\mathscr{N}_n - 1), \tag{58}$$

so that substituting (58) into (20) for $K = 2$, bearing in mind (56), we obtain (21). □

This section ends with the proof of Theorem 1.

Proof (Theorem 1) We write for (22)

$$\begin{aligned} d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, Z\right) &\le d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, \widetilde{\mathscr{Z}}_n[2]\right) + d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n[2], Z\right) \\ &\le d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, \mathscr{Z}_n[2]/\sqrt{\mathrm{Var}(\mathscr{Z}_n)}\right) + d_{\mathrm{W}}\left(\mathscr{Z}_n[2]/\sqrt{\mathrm{Var}(\mathscr{Z}_n)}, \widetilde{\mathscr{Z}}_n[2]\right) \\ &\quad + d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n[2], Z\right). \end{aligned} \tag{59}$$

Bearing in mind (14), the first term on the r.h.s. of (59) can be dealt with as follows

$$d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n, \mathscr{Z}_n[2]/\sqrt{\mathrm{Var}(\mathscr{Z}_n)}\right) \le \sqrt{\frac{\sum_{q\ge 2}\mathrm{Var}(\mathscr{Z}_n[2q])}{\mathrm{Var}(\mathscr{Z}_n)}} \ll \mathscr{N}_n^{-1/2}, \tag{60}$$

where the last estimate comes from (21), and the trivial lower bound for the total variance $\mathrm{Var}(\mathscr{Z}_n) \ge \mathrm{Var}(\mathscr{Z}_n[2])$. For the second term on the r.h.s. of (59) we have

$$d_{\mathrm{W}}\left(\mathscr{Z}_n[2]/\sqrt{\mathrm{Var}(\mathscr{Z}_n)}, \widetilde{\mathscr{Z}}_n[2]\right) \le \left| \frac{1}{\sqrt{1 + \frac{\sum_{q\ge 2}\mathrm{Var}(\mathscr{Z}_n[2q])}{\mathrm{Var}(\mathscr{Z}_n[2])}}} - 1 \right| \ll \mathscr{N}_n^{-1}, \tag{61}$$

where we used (19) and (21). We have finally to deal with the term $d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n[2], Z\right)$ on the right-hand side of (59). Since $\widetilde{\mathscr{Z}}_n[2]$ has unit variance and is an element of the second Wiener chaos associated with $\mathbf{A}$ (see Definition 1), we can apply [14, formula (5.2.15)], in order to deduce that

$$d_{\mathrm{W}}\left(\widetilde{\mathscr{Z}}_n[2], Z\right) \le \sqrt{\kappa_4\left(\widetilde{\mathscr{Z}}_n[2]\right)},$$

where $\kappa_4(Y)$ indicates the fourth cumulant of a random variable Y (see e.g. [14, Section A.2] and the references therein). Using Lemma 1, one also has that, for $n \in S$,

$$\widetilde{\mathscr{Z}}_n[2] = -\frac{\sqrt{2}}{\sqrt{\mathscr{N}_n}} \sum_{\lambda\in\Lambda_n^+} \left\{ \frac{1}{2}u_\lambda^2 + \frac{1}{2}v_\lambda^2 - 1 \right\},$$

where $\{u_\lambda, v_\lambda : \lambda \in \Lambda_n^+\}$ is a collection of i.i.d. centered standard Gaussian random variables. Exploiting independence together with the homogeneity of cumulants (see again [14, Section A.2]), one infers that

$$\kappa_4\left(\widetilde{\mathscr{Z}}_n[2]\right) = \frac{4}{\mathscr{N}_n^2} \cdot \frac{\mathscr{N}_n}{16} \cdot \kappa_4(Z^2) = \frac{12}{\mathscr{N}_n},$$

(recall that Z is a centred standard Gaussian random variable) and the desired conclusion follows at once. □

5 Proof of Proposition 2

In this section we will prove Proposition 2. Let us first give the proof of Lemma 2.

Proof (Lemma 2) Fix $K \geq 3$, and recall the notation (5). In order to simplify the discussion, for every $n \in S$ and given $Q \in \mathscr{Q}(M(n))$, we shall denote by $\mathscr{L}_n(Q\,;\geq 2K)$, the projection of the random variable $\mathscr{L}_n(Q)$ onto the direct sum of chaoses $\bigoplus_{q\geq K} C_{2q}$. For the l.h.s. of (27) we write

$$\sum_{q\geq K} \mathrm{Var}(\mathscr{L}_n[2q]) = \sum_{(Q,Q')} \mathrm{Cov}\left(\mathscr{L}_n(Q\,;\geq 2K), \mathscr{L}_n(Q'\,;\geq 2K)\right),$$

where the sum runs over the cartesian product $\mathscr{Q}(M(n)) \times \mathscr{Q}(M(n))$. We now write $\sum_{(Q,Q')} = \sum_{(Q,Q')\in G_0(n)} + \sum_{(Q,Q')\in G_1(n)}$, and study separately the two terms. By virtue of Cauchy-Schwarz and stationarity of T_n, one has that

$$\begin{aligned}\sum_{(Q,Q')\in G_1(n)} \mathrm{Cov}\left(\mathscr{L}_n(Q\,;\geq 2K), \mathscr{L}_n(Q'\,;\geq 2K)\right) &\leq |G_1(n)|\mathrm{Var}(\mathscr{L}_n(Q_0)) \\ &\ll E_n \int_{\mathbb{T}} r_n(x)^{2K}\, dx,\end{aligned}$$

where we have used (31) and (32), together with the fact that, by orthogonality, $\mathrm{Var}(\mathscr{L}_n(Q\,;\geq 2K)) \leq \mathrm{Var}(\mathscr{L}_n(Q)) = \mathrm{Var}(\mathscr{L}_n(Q_0))$. The rest of the proof follows closely the arguments rehearsed in [5, §6.2.2]. For all $Q \in \mathscr{Q}(M(n))$, we write

$$\begin{aligned}\mathscr{L}_n(Q\,;\geq 2K) = \sqrt{\frac{E_n}{2}} \sum_{q\geq K} \sum_{i_1+i_2+i_3=2q} \frac{\beta_{i_1}\alpha_{i_2,i_3}}{i_1!i_2!i_3!} \times \\ \times \int_Q H_{i_1}(T_n(x)) H_{i_2}(\widetilde{\partial}_1 T_n(x)) H_{i_3}(\widetilde{\partial}_2 T_n(x))\, dx,\end{aligned}$$

where the sum runs over all even integers $i_1, i_2, i_3 \geq 0$. We have

$$\begin{aligned}
&\left| \sum_{(Q,Q')\in G_0(n)} \mathrm{Cov}\left(\mathscr{L}_n(Q\,;\geq 2K),\ \mathscr{L}_n(Q'\,;\geq 2K)\right)\right| \\
&\leq E_n \sum_{q\geq 2K} \sum_{i_1+i_2+i_3=2q} \sum_{a_1+a_2+a_3+=2q} \left|\frac{\beta_{i_1}\alpha_{i_2,i_3}}{i_1!i_2!i_3!}\right| \cdot \left|\frac{\beta_{a_1}\alpha_{a_2,a_3}}{a_1!a_2!a_3!}\right| \\
&\quad \times \left| \sum_{(Q,Q')\in G_0(n)} \int_Q \int_{Q'} \mathbb{E}\Big[H_{i_1}(T_n(x)) H_{i_2}(\widetilde{\partial}_1 T_n(x)) H_{i_3}(\widetilde{\partial}_2 T_n(x)) \right. \\
&\qquad \left. \times\, H_{a_1}(T_n(y)) H_{a_2}(\widetilde{\partial}_1 T_n(y)) H_{a_3}(\widetilde{\partial}_2 T_n(y))\Big]\, dxdy \right|.
\end{aligned} \tag{62}$$

For $n \in S$, we now introduce the notation

$$(X_0(x), X_1(x), X_2(x)) := (T_n(x),\ \widetilde{\partial}_1 T_n(x),\ \widetilde{\partial}_2 T_n(x)), \quad x \in \mathbb{T}.$$

Applying the Leonov-Shyraev formulae for cumulants, in a form analogous to [5, Proposition 2.2], we infer that

$$\begin{aligned}
&\left| \sum_{(Q,Q')\in G_0(n)} \mathrm{Cov}\left(\mathscr{L}_n(Q\,;\geq 2K),\ \mathscr{L}_n(Q'\,;\geq 2K)\right)\right| \qquad (63)\\
&\leq E_n \sum_{q\geq 2K} \sum_{i_1+i_2+i_3=2q} \sum_{a_1+a_2+a_3=2q} \left|\frac{\beta_{i_1}\alpha_{i_2,i_3}}{i_1!i_2!i_3!}\right| \cdot \left|\frac{\beta_{a_1}\alpha_{a_2,a_3}}{a_1!a_2!a_3!}\right| \\
&\quad \times \mathbf{1}_{\{i_1+i_2+i_3=a_1+a_2+a_3\}} \Big| U(i_1,i_2,i_3;a_1,a_2,a_3)\Big|, \\
&=: E_n \times Z, \qquad (64)
\end{aligned}$$

where each summand $U = U(i_1, i_2, i_3; a_1, a_2, a_3)$ is the sum of at most $(2q)!$ terms of the type

$$u = \sum_{(Q,Q')\in G_0(n)} \int_Q \int_{Q'} \prod_{u=1}^{2q} R_{l_u,k_u}(x,y)\, dxdy, \tag{65}$$

with $k_u, l_u \in \{0, 1, 2\}$ and, for $l, k = 0, 1, 2$ and $x, y \in \mathbb{T}$, and we set

$$R_{l,k}(x,y) := \mathbb{E}\left[X_l(x) X_k(y)\right] = R_{l,k}(x-y),$$

where the last equality (with obvious notation) emphasises the fact that $R_{l,k}(x, y)$ only depends on the difference $x - y$. We will also exploit the following relation,

valid for every even integer p:

$$\int_{\mathbb{T}} R_{l,k}(x)^p \, dx \leq \int_{\mathbb{T}} r(x)^p \, dx; \tag{66}$$

also, for $x, y \in \mathbb{T}$, one has $|R_{l,k}(x-y)| \leq 1$, and, for $(x, y) \in Q \times Q'$,

$$|R_{l,k}(x-y)| \leq \epsilon. \tag{67}$$

Using the properties of $G_0(n)$ put forward in Proposition 4, as well as the fact that the sum defining Z in (64) involves indices $q \geq 2K$, one infers that, for u as in (65),

$$\begin{aligned} |u| &\leq \epsilon_1^{2q-2K} \sum_{(Q,Q') \in G_0(n)} \int_Q \int_{Q'} \prod_{u=1}^{2K} \left| R_{l_u,k_u}(x,y) \right| dxdy \\ &\leq \epsilon_1^{2q-2K} \int_{\mathbb{T}} \prod_{u=1}^{2K} \left| R_{l_u,k_u}(x) \right| dx \leq \epsilon_1^{2q-2K} R_n(2K), \end{aligned}$$

where $R_n(2K) = \int_{\mathbb{T}} r_n(x)^{2K} \, dx$, and we have applied a generalised Hölder inequality together with (66) in order to obtain the last estimate. This relation yields that each of the terms U contributing to Z can be bounded as follows:

$$\begin{aligned} &\left| U(i_1, i_2, i_3; a_1, a_2, a_3) \right| \\ &\quad \leq (2q)! \frac{R_n(2K)}{\epsilon^{2K}} \epsilon^{2q} = (2q)! \frac{R_n(2K)}{\epsilon^{2K}} (\sqrt{\epsilon})^{i_1+i_2+i_3} (\sqrt{\epsilon})^{a_1+a_2+a_3}. \end{aligned}$$

This yields that

$$\begin{aligned} Z \leq \frac{R_n(2K)}{\epsilon^{2K}} \sum_{q \geq 2K} (2q)! \sum_{i_1+i_2+i_3=2q} \sum_{a_1+a_2+a_3=2q} \left| \frac{\beta_{i_1} \alpha_{i_2,i_3}}{i_1! i_2! i_3!} \right| \times \\ \left| \frac{\beta_{a_1} \alpha_{a_2,a_3}}{a_1! a_2! a_3!} \right| \times (\sqrt{\epsilon})^{i_1+i_2+i_3} (\sqrt{\epsilon})^{a_1+a_2+a_3} =: \frac{R_n(2K)}{\epsilon^{2K}} \times S. \end{aligned}$$

The fact that $S < \infty$ now follows from standard estimates, such as the ones appearing in [5, end of §6.2.2]. This concludes the proof of (27). To prove (28), it suffices to recall (56) for $K = 3$, and use an estimate by Bombieri-Bourgain (see [3, Theorem 1]), stating that $|S_6(n)| = O(\mathscr{N}_n^{7/2})$, as $\mathscr{N}_n \to +\infty$. □

We are now ready to prove Proposition 2.

Proof (Proposition 2) By the triangle inequality, for the l.h.s. of (26) we write

$$\mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_n) - h(\widetilde{\mathscr{L}}_n[4])\right|\right] \leq \mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_n) - h(\mathscr{L}_n[4]/\sqrt{\mathrm{Var}(\mathscr{L}_n)})\right|\right] + \mathbb{E}\left[\left|h(\mathscr{L}_n[4]/\sqrt{\mathrm{Var}(\mathscr{L}_n)}) - h(\widetilde{\mathscr{L}}_n[4])\right|\right]. \tag{68}$$

For the first term on the r.h.s. of (68), since h is Lipschitz, from (18) and Cauchy-Schwartz

$$\mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_n) - h(\mathscr{L}_n[4]/\sqrt{\mathrm{Var}(\mathscr{L}_n)})\right|\right] \leq \frac{1}{\sqrt{\mathrm{Var}(\mathscr{L}_n)}}\mathbb{E}\left[\left|\sum_{q\geq 3}\mathscr{L}_n[2q]\right|\right] \leq \sqrt{\frac{\sum_{q\geq 3}\mathrm{Var}(\mathscr{L}_n[2q])}{\mathrm{Var}(\mathscr{L}_n)}} \ll \mathscr{N}_n^{-1/4},$$

where the last upper bound follows from (10) and Lemma 2. For the second term on the r.h.s. of (68), we have again by the Lipschitz property and some standard steps

$$\begin{aligned}
&\mathbb{E}\left[\left|h(\mathscr{L}_n[4]/\sqrt{\mathrm{Var}(\mathscr{L}_n)}) - h(\widetilde{\mathscr{L}}_n[4])\right|\right] \\
&\leq \left|\frac{1}{\sqrt{\mathrm{Var}(\mathscr{L}_n)}} - \frac{1}{\sqrt{\mathrm{Var}(\mathscr{L}_n[4])}}\right|\mathbb{E}\left[\left|\mathscr{L}_n[4]\right|\right] \\
&= \frac{1}{\sqrt{\mathrm{Var}(\mathscr{L}_n[4])}}\left|\frac{1}{\sqrt{1+\frac{\sum_{q\geq 3}\mathrm{Var}(\mathscr{L}_n[2q])}{\mathrm{Var}(\mathscr{L}_n[4])}}} - 1\right|\mathbb{E}\left[\left|\mathscr{L}_n[4]\right|\right] \\
&\leq \left|\frac{1}{\sqrt{1+\frac{\sum_{q\geq 3}\mathrm{Var}(\mathscr{L}_n[2q])}{\mathrm{Var}(\mathscr{L}_n[4])}}} - 1\right| \ll \mathscr{N}_n^{-1/4},
\end{aligned}$$

where the last bound comes from (48) and Lemma 2. □

6 Proofs of Proposition 3 and Theorem 2

Recall (46), then we can rewrite (47) as

$$\mathscr{L}_n[4] = \sqrt{\frac{E_n}{\mathscr{N}_n^2}} \frac{1}{\sqrt{512}} \left(p(\widehat{W}) + \psi_n \right), \tag{69}$$

where

$$\psi_n := \frac{1}{2} \frac{1}{\mathscr{N}_n} \sum_{\lambda \in \Lambda_n} (|a_\lambda|^4 - 2), \tag{70}$$

$$\widehat{W} := (W_1, W_2, W_4), \tag{71}$$

and p is the polynomial

$$p(x, y, z) := 1 - x^2 - 4y^2 + 4xy - 4z^2. \tag{72}$$

The following statement is a key step in order to prove Proposition 3.

Lemma 5 *Let $h : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function, define $\widehat{W}$ as in (71) for a fixed $n \in S$, and select $\eta \in [-1, 1]$. Then, on S,*

$$\left| \mathbb{E}\left[h\left(p(\widehat{W}) \right) \right] - \mathbb{E}\left[h\left(p(\widehat{Z}) \right) \right] \right| \ll |\widehat{\mu_n}(4) - \eta|^{1/2} \vee \mathscr{N}_n^{-1/4}, \tag{73}$$

where the constant involving in the previous estimation is independent of η and h, p is the second degree polynomial defined in (72) and $\widehat{Z} = \widehat{Z}(\eta) := (Z_1, Z_2, Z_4)$ is defined according to (49).

Proof We will apply an approximation argument from Ch. Döbler's dissertation [6]. Indeed, according to [6, Proposition 2.7.5, Corollary 2.7.6 and Lemma 2.7.7], to every Lipschitz mapping h as in the statement one can associate a collection of real-valued functions $\{h_\rho : \rho \geq 1\}$, such that the following properties are verified for every ρ: (i) h_ρ equals the convolution of h with a centered Gaussian density with variance $1/\rho^2$, (ii) h_ρ is continuously infinitely differentiable, and $\|h_\rho^{(m)}\|_\infty \leq \rho^{m-1}$ (with $h_\rho^{(m)}$ denoting the mth derivative of h_ρ), and (iii) for every integrable random variable X, one has that $|\mathbb{E}[h(X) - h_\rho(X)]| \leq \rho^{-1}$. From Point (iii) it follows in particular that

$$\left| \mathbb{E}\left[h\left(p(\widehat{W}) \right) \right] - \mathbb{E}\left[h\left(p(\widehat{Z}) \right) \right] \right| \leq \frac{2}{\rho} + \left| \mathbb{E}\left[F_\rho(\widehat{W}) \right] - \mathbb{E}\left[F_\rho(\widehat{Z}) \right] \right| =: \frac{2}{\rho} + B(\rho),$$

with $F_\rho := h_\rho \circ p$. Note that F_ρ is an infinitely differentiable mapping, whose partial derivatives have at most polynomial growth. This implies that we can directly apply the same interpolation and integration by parts argument one can find in [14, Proof of Theorem 6.1.2], to deduce that

$$B(\rho) \leq \underbrace{\sum_{i,j=1}^{3} |\widehat{\Sigma}(i,j) - \widehat{\Sigma}_n(i,j)| \mathbb{E}[|\partial^2_{i,j} F_\rho(\widehat{W}(n))|]}_{:=I_1}$$

$$+ \underbrace{\sum_{i,j=1}^{3} \sqrt{\mathbb{E}[|\partial^2_{i,j} F_\rho(\widehat{W}(n))|^2]\, \mathbb{E}[|\widehat{\Sigma}_n(i,j) - \langle D\widehat{W}_j(n), -DL^{-1}\widehat{W}_i(n)\rangle|^2]}}_{:=I_2},$$

where $\partial^2_{i,j} := \partial^2/\partial x_i \partial x_j$, D denotes the Malliavin derivative (see [14, Definition 1.1.8]), L^{-1} the inverse of the infinitesimal generator of the Ornstein-Uhlenbeck semigroup (see [14, §1.3]) and $\langle \cdot, \cdot \rangle$ stands for the inner product of an appropriate real separable Hilbert space $\mathscr{H}$ (whose exact definition is immaterial for the present proof). Standard arguments based on hypercontractivity and Point (ii) discussed above (together with the fact that $\rho \geq 1$) yield that $E[|\partial^2_{i,j} F_\rho(\widehat{W}(n))|^2]^{1/2} \leq C\rho$, for some absolute constant C. In view of these facts, relations (45) and (50) imply therefore that

$$I_1 \ll |\widehat{\mu}_n(4) - \eta|. \tag{74}$$

To deal with I_2, we can use the upper bound in [14, formula (6.2.6)], together with the fact that each $\widehat{W}_i(n)$ belongs to the second Wiener chaos; it hence remains to compute the fourth cumulant $k_4(\widehat{W}_i(n)) = \mathbb{E}[\widehat{W}_i(n)^4] - 3\mathbb{E}[\widehat{W}_i(n)^2]^2$ for every i (note that these cumulants are necessarily positive). Standard computations yield that,

$$\kappa_4(W_1(n)) \ll \frac{1}{\mathscr{N}_n}, \quad \kappa_4(W_2(n)) \ll \frac{1}{\mathscr{N}_n}\frac{1}{\mathscr{N}_n}\sum_{\lambda} \frac{\lambda_1^8}{n^4},$$

$$\kappa_4(W_4(n)) \ll \frac{1}{\mathscr{N}_n}\frac{1}{\mathscr{N}_n}\sum_{\lambda} \frac{\lambda_1^4\lambda_2^4}{n^4},$$

from which we deduce

$$I_2 \ll \sqrt{\frac{1}{\mathscr{N}_n}}. \tag{75}$$

We have therefore proved the existence of an absolute constant C such that

$$\left|\mathbb{E}\left[h\left(p(\widehat{W})\right)\right]-\mathbb{E}\left[h\left(p(\widehat{Z})\right)\right]\right|\leq C\left\{\frac{1}{\rho}+\rho\gamma_n\right\},$$

with $\gamma_n := (2\mathscr{N}_n^{1/2})^{-1}|\widehat{\mu}_n(4)-\eta| \leq 1$. Since the right-hand side of the previous inequality is maximised at the point $\rho=\gamma_n^{-1/2}$, we immediately obtain the desired conclusion. □

Let us now prove Proposition 3.

Proof (Proposition 3) We can rewrite the l.h.s. of (29) as

$$\left|\mathbb{E}\left[h\left(\frac{p(\widehat{W})+\psi_{n_j}}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}\right)-h\left(\frac{p(Z)}{\sqrt{1+\eta^2}}\right)\right]\right|,$$

where for $n\in S$, ψ_n is given in (70). By the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|h\left(\frac{p(\widehat{W})+\psi_{n_j}}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}\right)-h\left(\frac{p(Z)}{\sqrt{1+\eta^2}}\right)\right|\right]\\
\leq\ &\mathbb{E}\left[\left|h\left(\frac{p(\widehat{W})+\psi_{n_j}}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}\right)-h\left(\frac{p(\widehat{W})}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}\right)\right|\right]\\
&+\mathbb{E}\left[\left|h\left(\frac{p(\widehat{W})}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}\right)-h\left(\frac{p(\widehat{W})}{\sqrt{1+\eta^2}}\right)\right|\right]\\
&+\left|\mathbb{E}\left[h\left(\frac{p(\widehat{W})}{\sqrt{1+\eta^2}}\right)-h\left(\frac{p(Z)}{\sqrt{1+\eta^2}}\right)\right]\right|\\
=:\ &I_{n_j}+J_{n_j}+K_{n_j}.
\end{aligned}\tag{76}$$

For the first term we simply have, since h is Lipschitz,

$$I_{n_j}\ll \mathrm{Var}(\psi_{n_j})=\frac{10}{\mathscr{N}_{n_j}},\tag{77}$$

where the last equality is (85). Let us now deal with J_{n_j}. By the Lipschitz property,

$$J_{n_j} \leq \sqrt{1+\widehat{\mu_{n_j}}(4)^2}\left|\frac{1}{\sqrt{1+\widehat{\mu_{n_j}}(4)^2+34/\mathscr{N}_{n_j}}}-\frac{1}{\sqrt{1+\eta^2}}\right|$$

$$=\sqrt{\frac{1+\widehat{\mu_{n_j}}(4)^2}{1+\eta^2}}\left|\frac{1}{\sqrt{1+\frac{\widehat{\mu_{n_j}}(4)^2-\eta^2+34/\mathscr{N}_{n_j}}{1+\eta^2}}}-1\right|$$

$$\ll |\widehat{\mu_{n_j}}(4)^2-\eta^2|+34\mathscr{N}_{n_j}^{-1} \ll ||\widehat{\mu_{n_j}}(4)|-\eta| \vee \mathscr{N}_{n_j}^{-1}. \tag{78}$$

Finally, note that Lemma 5 and the equality in law $\mathscr{M}_\eta = \mathscr{M}_{-\eta}$ give

$$K_n \ll ||\widehat{\mu_{n_j}}(4)|-\eta|^{1/2} \vee \mathscr{N}_{n_j}^{-1/4}.$$

Plugging the latter bound, (77) and (78) into (76) we conclude the proof of Proposition 3. □

6.1 Proof of Theorem 2

Proof (Theorem 2) For every $j \geq 1$, reasoning as in (25),

$$\mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_{n_j})-h(\mathscr{M}_\eta)\right|\right] \leq \mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_{n_j})-h(\widetilde{\mathscr{L}}_{n_j}[4])\right|\right]$$

$$+\mathbb{E}\left[\left|h(\widetilde{\mathscr{L}}_{n_j}[4])-h(\mathscr{M}_\eta)\right|\right]$$

$$\ll \mathscr{N}_{n_j}^{-1/4} \vee ||\widehat{\mu_{n_j}}(4)|-\eta|^{1/2},$$

where the last step directly follows from Propositions 2 and 3. □

Appendix

Proof (Lemma 3) From [13, Lemma 3.4], we have that the chaotic expansion of $\mathscr{Z}_n^\varepsilon$ is

$$\mathscr{Z}_n^\varepsilon = \sum_{q=0}^{+\infty} \mathscr{Z}_n^\varepsilon[2q] = \sum_{q=0}^{+\infty} \frac{\beta_{2q}^\varepsilon}{(2q)!}\int_{\mathbb{T}} H_{2q}(T_n(x))\,dx, \tag{79}$$

where H_{2q} denotes the $2q$-th Hermite polynomial, and

$$\beta_0^\varepsilon = \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}\phi(t)\,dt, \qquad \beta_{2q}^\varepsilon = -\frac{1}{\varepsilon}\phi(\varepsilon)H_{2q-1}(\varepsilon),\ q \geq 1, \tag{80}$$

ϕ still denoting the Gaussian density. Taking the limit for ε going to 0 in (80) we obtain the collection of coefficients (37), related to the (formal) Hermite expansion of the Dirac mass δ_0. Note that

$$\sum_{q=1}^{+\infty}\frac{(\beta_{2q})^2}{(2q)!}\int_{\mathbb{T}} r_n(x)^{2q}\,dx = \frac{1}{2\pi}\int_{\mathbb{T}}\left(\frac{1}{\sqrt{1-r_n(x)^2}}-1\right)dx < +\infty, \tag{81}$$

since the collection $\{(\beta_{2q})^2/(2q)!\}_q$ coincides with the sequence of Taylor coefficients of the function $x \mapsto 1/(2\pi\sqrt{1-x^2})$ around zero; thanks to Lemma 5.3 in [15] we have the finiteness of the integral. Therefore the series

$$\sum_{q=0}^{+\infty}\frac{\beta_{2q}}{(2q)!}\int_{\mathbb{T}} H_{2q}(T_n(x))\,dx,$$

is a well-defined random variable in $L^2(\mathbb{P})$, its variance being the series on the l.h.s. of (81). Moreover, from [1, 22.14.16] and (81)

$$\sum_{q=1}^{+\infty}\frac{(\beta_{2q}^\varepsilon-\beta_{2q})^2}{(2q)!}\int_{\mathbb{T}} r_n(x)^{2q}\,dx \leq 2\sum_{q=1}^{+\infty}\frac{(\beta_{2q})^2}{(2q)!}\int_{\mathbb{T}} r_n(x)^{2q}\,dx < +\infty,$$

that implies, by the dominated convergence theorem, $\mathscr{Z}_n^\varepsilon \to \mathscr{Z}_n$, $\varepsilon \to 0$, in $L^2(\mathbb{P})$. □

Proof (Lemma 4) From (44) with $q = 2$

$$\begin{aligned}\mathscr{L}_n[4] = \frac{\sqrt{E_n}}{128\sqrt{2}}\Big(&8\int_{\mathbf{T}} H_4(T_n(x))\,dx - \int_{\mathbf{T}} H_4(\widetilde{\partial}_1 T_n(x))\,dx - \int_{\mathbf{T}} H_4(\widetilde{\partial}_2 T_n(x))\,dx\\ &-8\int_{\mathbf{T}} H_2(T_n(x))H_2(\widetilde{\partial}_1 T_n(x))\,dx - 8\int_{\mathbf{T}} H_2(T_n(x))H_2(\widetilde{\partial}_2 T_n(x))\,dx\\ &-2\int_{\mathbf{T}} H_2(\widetilde{\partial}_1 T_n(x))H_2(\widetilde{\partial}_2 T_n(x))\,dx.\end{aligned}$$

Lemmas 5.2 and 5.5 in [13] together with some straightforward computations allow one to write, from (82),

$$\mathscr{L}_n[4] = \sqrt{\frac{E_n}{\mathscr{N}_n^2}}\frac{1}{128\sqrt{2}}\Big(8W_1^2 - 16W_2^2 - 16W_3^2 - 32W_4^2$$
$$+\frac{1}{\mathscr{N}_n}\sum_{\lambda\in\Lambda_n}|a_\lambda|^4\left(-8+12\left(\left(\frac{\lambda_1}{\sqrt{n}}\right)^2+\left(\frac{\lambda_2}{\sqrt{n}}\right)^2\right)^2\right)\Big).$$

Recalling that $\lambda_1^2+\lambda_2^2=n$, we obtain (47). Let us now note that we can write

$$W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2$$
$$= \frac{1}{\mathscr{N}_n/2}\sum_{\lambda,\lambda'\in\Lambda_n^+}\left(1-\frac{2}{n^2}(\lambda_1\lambda_1'+\lambda_2\lambda_2')^2\right)(|a_\lambda|^2-1)(|a_{\lambda'}|^2-1). \tag{82}$$

Then it is immediate to compute from (82)

$$\mathbb{E}\left[W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2\right] = -1. \tag{83}$$

Bearing in mind Lemma 4.1 in [13], still from (82) some straightforward computations lead to

$$\mathbb{E}\left[(W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2)^2\right] = 2 + \widehat{\mu_n}(4)^2 + \frac{48}{\mathscr{N}_n}. \tag{84}$$

From (83) and (84) hence we find

$$\mathrm{Var}(W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2) = 1 + \widehat{\mu_n}(4)^2 + \frac{48}{\mathscr{N}_n}.$$

Recalling that $(\sqrt{2}|a_\lambda|)^2$ is distributed as a chi-square random variable with two degrees of freedom,

$$\mathrm{Var}\left(\frac{1}{2}\frac{1}{\mathscr{N}_n}\sum_{\lambda\in\Lambda_n}|a_\lambda|^4\right) = \frac{10}{\mathscr{N}_n}, \tag{85}$$

and moreover

$$\mathrm{Cov}\left(W_1^2 - 2W_2^2 - 2W_3^2 - 4W_4^2, \frac{1}{2}\frac{1}{\mathscr{N}_n}\sum_{\lambda\in\Lambda_n}|a_\lambda|^4\right) = -\frac{12}{\mathscr{N}_n}.$$

This concludes the proof of Lemma 4. □

Acknowledgements We thank Ch. Döbler for useful discussions (in particular, for pointing out the relevance of [6]), as well as two Referees for several useful remarks. The research leading to this work has been supported by the grant F1R-MTH-PUL-15STAR (STARS) at the University of Luxembourg.

References

1. Abramovitz, M., Stegun, I.-A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series, vol. 55. United States Department of Commerce, National Bureau of Standards (NBS), Washington, DC (1964)
2. Berry, M.V.: Statistics of nodal lines and points in chaotic quantum billiards: perimeter corrections, fluctuations, curvature. J. Phys. A **35**, 3025–3038 (2002)
3. Bombieri E., Bourgain, J.: A problem on sums of two squares. Int. Math. Res. Not. **11**, 3343–3407 (2015)
4. Cheng, S.Y.: Eigenfunctions and nodal sets. Comment. Math. Helv. **51**(1), 43–55 (1976)
5. Dalmao, F., Nourdin, I., Peccati, G., Rossi, M.: Phase singularities in complex arithmetic random waves. Preprint (2016). ArXiv 1608.05631
6. Döbler, C.: New developments in Stein's method with applications. Ph.D. Thesis, Ruhr-Universität Bochum (2012)
7. Döbler, C.: The Stein equation beyond the support with applications. In preparation (2018+)
8. Erdös, P., Hall, R.-R.: On the angular distribution of Gaussian integers with fixed norm. Discrete Math. **200**(1–3), 87–94 (1999)
9. Geman, D., Horowitz, J.: Local times for real and random functions. Duke Math. J. **43**(4), 809–828 (1976)
10. Hardy, G.H., Wright, E.-M.: An Introduction to the Theory of Numbers, 5th edn. The Clarendon Press/Oxford University Press, New York (1979)
11. Krishnapur, M., Kurlberg, P., Wigman, I.: Nodal length fluctuations for arithmetic random waves. Ann. Math. **177**(2), 699–737 (2013)
12. Kurlberg, P., Wigman, I.: On probability measures arising from lattice points on circles. Math. Ann. (2015, in press). ArXiv: 1501.01995
13. Marinucci, D., Peccati, G., Rossi, M., Wigman, I.: Non-universality of nodal length distribution for arithmetic random waves. Geom. Funct. Anal. **26**(3), 926–960 (2016)
14. Nourdin, I., Peccati, G.: Normal approximations with Malliavin calculus. From Stein's method to universality. Cambridge Tracts in Mathematics, vol. 192. Cambridge University Press, Cambridge (2012)
15. Oravecz, F., Rudnick, Z., Wigman, I.: The Leray measure of nodal sets for random eigenfunctions on the torus. Ann. Inst. Fourier (Grenoble) **58**(1), 299–335 (2008)
16. Rudnick, Z., Wigman, I.: On the volume of nodal sets for eigenfunctions of the Laplacian on the torus. Ann. Henri Poincaré. **9**(1), 109–130 (2008)
17. Wigman, I.: Fluctuations of the nodal length of random spherical harmonics. Commun. Math. Phys. **298**(3), 787–831 (2010)
18. Wigman, I.: On the nodal lines of random and deterministic Laplace eigenfunctions. In: Spectral Geometry. Proceedings of Symposia in Pure Mathematics, vol. 84, pp. 285–297. American Mathematical Society, Providence (2012)
19. Yau, S.T.: Survey on partial differential equations in differential geometry. Seminar on Differential Geometry. Annals of Mathematics Studies, vol. 102, pp. 3–71. Princeton University Press, Princeton (1982)

Combinatorics on Words and the Theory of Markoff

Christophe Reutenauer

Abstract This is a survey on the theory of Markoff, in its two aspects: quadratic forms (the original point of view of Markoff), approximation of reals. A link wih combinatorics on words is shown, through the notion of Christoffel words and special palindromes, called central words. Markoff triples may be characterized, by using some linear representation of the free monoid, restricted to these words, and Fricke relations. A double iterated palindromization allows to construct all Markoff numbers and to reformulate the Markoff numbers injectivity conjecture (Frobenius, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin 26:458–487, 1913).

1 Introduction

In a short article written in Latin in 1875, Christoffel [11] introduced a family of words on a two letter alphabet, that we call *Christoffel words*. Shortly after, they were also considered by Smith [36], who did not know, as he says and regrets, Christoffel's work. These words were followed in the twentieth century by the theory of Sturmian sequences, introduced in 1940 by Morse and Hedlund [28] in Symbolic Dynamics. More recently, there has been a lot of work, beginning by Jean Berstel and Aldo de Luca, on these words, from the point of view of Combinatorics on Words and also in Discrete Geometry, see among others [2, 7, 23].

Independently from Christoffel, Markoff (= Markov, famous for the Markov processes, but writing his name in the French way) wrote as young student two brilliant articles [26, 27] in 1879 and 1880, on the theory called now *Theory of Markoff*. This theory has two sides: it characterizes on one hand certain quadratic forms, and on the other, certain real numbers, defined by some extremal conditions (by their minima for quadratic forms, and by their rational approximations for

C. Reutenauer (✉)
Département de mathématiques, Université du Québec à Montréal, Montréal, QC, Canada
e-mail: reutenauer.christophe@uqam.ca

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_24

real numbers). They are constructed using some special integers called *Markoff numbers*, which are among others characterized by a special Diophantine equation, the *Markoff equation*.

Looking at the subsequent literature, it is seen that Markoff's theory has visibly fascinated many mathematicians, which have developed, and often reproved one or the other side of the theory: Hurwitz [21], Frobenius [19], Perron [30], Remak [31], Dickson [17], Cassels [9, 10], Cohn [12], Bombieri [5], and the list is much longer. Three books must be cited here: the unavoidable book by Cusick and Flahive [14], the book by Perrine [29] who gives among others a lot of matrix constructions, and the recent book by Aigner [1], celebrating the 100th anniversary of Frobenius' *injectivity conjecture for Markoff numbers*.

Markoff constructs the special quadratic forms that appear in his theory by using special patterns of 1s and 2s (in the continued fraction expansion); these patterns happen to be Christoffel words. The link between Christoffel words and the theory of Markoff was explicitly noted by Frobenius in 1913 [19], and somewhat forgotten until recently (but it was known to Caroline Series [35]). The scope of the present survey is to present Markoff's theory, from the point of view of Combinatorics on words, especially Christoffel words. In the two final sections, we review some tree constructions, and relate calculations made by Frobenius to the standard factorization of Christoffel words (this seems to be new).

The interested reader will find many results and proofs in the forthcoming book [33]. For the theory of Christoffel and related words, see also Chapter 2 of [25] and the first part of the book [3].

The author thanks the two referees for their useful and kind comments.

2 Christoffel Words

A *lattice path* is a sequence of consecutive elementary steps in the plane; each *elementary step* is a segment $[(x, y), (x+1, y)]$ or $[(x, y), (x, y+1)]$, with $x, y \in \mathbf{Z}$.

Let p, q be relatively prime natural integers. Consider the segment from some integral point A to $B = A + (p, q)$ and the lattice path from A to B located below this segment and such that the polygon delimited by the segment and the path has no interior integer point.

Given a totally ordered alphabet $\{a < b\}$, the *lower Christoffel word of slope* q/p is the word in the free monoid[1] $\{a, b\}^*$ coding the above path, where a (resp. b) codes an horizontal (resp. vertical) elementary step. See Fig. 1, where is represented the path with $(p, q) = (7, 4)$ corresponding to the Christoffel word $aabaabaabab$ of slope 4/7. Note that the slope of a Christoffel word is equal to the slope of the

[1]The free monoid A^* is the set of words (= strings = finite sequences) on the set A, including the empty one; this is a monoid, the product of two words being the concatenation.

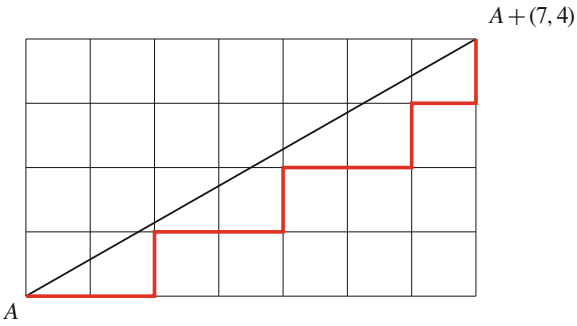


Fig. 1 The lower Christoffel words *aabaabaabab* of slope 4/7

segment in the plane delimited by the extreme points of the corresponding discrete path.

We say that the above path, and the lower Christoffel word, *discretizes from below* the segment AB.

The upper Christoffel word of slope q/p is defined similarly, by considering the lattice path located above the segment. Since the rectangle with opposite vertices A and B and sides parallel to the coordinate lines has a symmetry around its center, it follows that *the upper Christoffel word of a given slope is the reversal* $\tilde{w}$ *of the lower Christoffel word* w *of the same slope*. It is known also that *a lower Christoffel word* w *and the corresponding upper Christoffel word* $\tilde{w}$ *are conjugate*[2] *in the free monoid* $\{a < b\}^*$. See [3] Lemma 2.7.

Clearly, the number of a's in the lower and upper Christoffel word of slope q/p is p, while the number of b's is q. In particular, $|w|_a$, $|w|_b$ are relatively prime when w is a lower or upper Christoffel word and w cannot be a nontrivial power of another word.

The letters a and b are Christoffel words. The other Christoffel words are called *proper*. The words $a^n b$ and ab^n, for $n \geq 0$, are lower Christoffel words.

On the path defining the Christoffel word, consider the integral point, not equal to the first nor to the last, which is the closest to the diagonal AB of the rectangle. This defines a factorization of the Christoffel word, called its *standard factorization*. In the example of Fig. 1, the point is $A + (2, 1)$, and the factorization is $aab.aabaabab$. It follows from a theorem of Borel and Laubie that the two factors are themselves Christoffel words, and that this factorization is unique [3, 6] Theorem 3.3.

[2]This means that $w = uv$ and $\tilde{w} = vu$ for some words u, v.

3 Markoff Triples and Numbers

A *Markoff triple* is a multiset $\{x, y, z\}$ of positive integers satisfying the *Markoff equation*

$$x^2 + y^2 + z^2 = 3xyz.$$

Examples are $\{1, 1, 1\}$, $\{1, 1, 2\}$ and $\{1, 2, 5\}$. We sometimes use ordered triples for representing Markoff triples. A Markoff triple is called *proper* if the three numbers are distinct. Otherwise we call it *improper*. The improper Markoff triples are $\{1, 1, 1\}$ and $\{1, 1, 2\}$. A *Markoff number* is an element of a Markoff triple.

Consider the monoid homomorphism μ from the free monoid $\{a, b\}^*$ into $SL_2(\mathbb{Z})$ defined by

$$\mu(a) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mu(b) = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}.$$

The matrix construction of the Markoff triples as shown in the following theorem was obtained by Cohn [12, 13]. He used the third Fricke relation (see the lemma below), having noted the striking analogy between Markoff's equation and this relation. Uniqueness was noticed by Bombieri [5] Theorem 26, and independently by the author [32] Theorem 1.

Theorem 3.1 *The mapping sending each Christoffel word w with standard factorization uv onto the multiset $\{\frac{1}{3}Tr(\mu(u)), \frac{1}{3}Tr(\mu(v)), \frac{1}{3}Tr(\mu(w))\}$ is a bijection from the set of proper lower Christoffel words onto the set of proper Markoff triples.*

It is useful to know that for w a lower Christoffel word, one has $\frac{1}{3}Tr(\mu(w)) = \mu(w)_{12}$ (see [3] Lemma 8.7). The Fricke relations are given in the following lemma.

Lemma 3.1 (Fricke relations [18], (6) p. 91) *Let A, B be matrices in $SL_2(\mathbb{Z})$. Then:* $\mathrm{Tr}(A^2B) + \mathrm{Tr}(B) = \mathrm{Tr}(A)\,\mathrm{Tr}(AB)$, $\mathrm{Tr}(AB^2) + \mathrm{Tr}(A) = \mathrm{Tr}(AB)\,\mathrm{Tr}(B)$ *and* $\mathrm{Tr}(A)^2 + \mathrm{Tr}(B)^2 + \mathrm{Tr}(AB)^2 = \mathrm{Tr}(A)\,\mathrm{Tr}(B)\,\mathrm{Tr}(AB) + \mathrm{Tr}(ABA^{-1}B^{-1}) + 2$.

It follows from Theorem 3.1 that for each Markoff number m, there exists a lower Christoffel word w such that $m = \frac{1}{3}Tr(\mu(w)) = \mu(w)_{12}$. We say that m is the Markoff number *associated to the Christoffel word w*. The so-called *conjecture of Frobenius* [19], also called *Markoff numbers injectivity conjecture*, is the following open question[3]: is the mapping $w \mapsto m$ injective? Of course, the theorem has a striking analogy with this conjecture, since the former asserts that the mapping which to w associates its Markoff triple is bijective. Some partial answers to the

[3]Frobenius states it, in two different forms, as an open problem, not a conjecture [19] p. 601 and 614.

conjecture have been given (see the book [1] by Martin Aigner), but the general case seems to be very difficult.

The Frobenius conjecture is equivalent to the conjecture that for each Markoff number m, there is a unique Markoff triple of which m is the maximum, see [1] p. 39 (this works also for the improper triples). For example, 1, 2, 5 are respectively the unique maxima of (1, 1, 1), 1, 1, 2), (1, 2, 5).

4 Lagrange Number of a Real Number

Let x be an irrational real number. Consider the set of real numbers L such that the inequality $|x - p/q| < 1/Lq^2$ holds for infinitely many rational numbers p/q.

Define $L(x)$ to be the supremum of all these L. It is called the *Lagrange number* of x.

Let x be represented by its infinite continued fraction $[a_0, a_1, a_2, \ldots]$, and define $x_n = [a_n, a_{n+1}, a_{n+2}, \ldots]$; moreover, for $n \geq 1$, define $y_n = [a_n, \ldots, a_1]$. Finally, for $n \geq 2$, let $\lambda_n(x) = x_n + y_{n-1}^{-1}$. We have

$$\begin{aligned} \lambda_n(x) &= [a_n, a_{n+1}, \ldots] + [a_{n-1}, \ldots, a_1]^{-1} \\ &= a_n + [a_{n+1}, a_{n+2}, \ldots]^{-1} + [a_{n-1}, \ldots, a_1]^{-1} \\ &= [a_{n+1}, a_{n+2}, \ldots]^{-1} + [a_n, a_{n-1}, \ldots, a_1]. \end{aligned} \tag{1}$$

For $n = 1$, we define by the last equation: $\lambda_1(x) = [a_2, a_3, \ldots]^{-1} + a_1 = x_1$. The next result is essential for the determination of the Lagrange number of a sequence; it is stated without proof by Hurwitz [21] p. 283, see [1] p. 23 for a proof.

Theorem 4.1

$$L(x) = \limsup_{n \to \infty} \lambda_n(x).$$

The main tool in the proof is the following classical identity (where p_n/q_n is the n-th convergent of x)

$$|x - \frac{p_n}{q_n}| = \frac{1}{\lambda_{n+1}(x) q_n^2}. \tag{2}$$

Recall that two irrational real numbers are called *equivalent* if their expansions into continued fractions coincide after some rank (which may be not the same rank for both numbers). It follows from the previous theorem that in this case they have the same Lagrange number.

5 Main Technical Result

Motivated by the previous result, we define for any infinite word $s = a_0a_1a_2\cdots$ over the $\mathbb{P}$ set of positive natural integers

$$\lambda_i(s) = [a_{n-1}, a_{n-2}, \ldots, a_1]^{-1} + [a_n, a_{n+1}, a_{n+2}, \ldots].$$

The next result has a strong similarity with previous results involving doubly infinite words, as used by Markoff and subsequent authors (for example Dickson and Bombieri). It was however tempting to find a version for infinite words, since continued fractions are such words. This result will be used to prove Markoff's theorems for continued fractions and Markoff's theorem on quadratic forms. Denote by χ the monoid homomorphism from the free monoid $\{a, b\}^*$ into the free monoid $\{1, 2\}^*$ sending a onto 11 and b onto 22. For a nonempty word $w = a_0 \ldots a_{n-1}$, we denote by w^∞ the infinite word $b_0b_1b_2b_3\cdots$ whose letter b_i in position i satisfies $b_i = a_{i \mod n}$.

Theorem 5.1 *Let s be an infinite word over $\mathbb{P}$ such that for some I and some $\theta < 3$ one has $\lambda_i(s) < \theta$ for any $i \geq I$. Then $s = u\chi(w)^\infty$ for some lower Christoffel word w, and some word u whose length is bounded by a function depending only on I and θ.*

This result (except the bounds) may be deduced from similar results for bi-infinite words (see for example [1, 5]). A direct proof will be found in the forthcoming book [33].

6 Markoff's Theorem for Approximations

For the results in these sections, see among others [10] Theorem III p.41, [5] Theorem 1 and [1] Theorem p.185 (and also [33]).

If $s = b_0b_1b_2b_3\cdots$ is an infinite word on $\mathbb{P}$, we denote by $[s]$ the real number whose expansion into continued fractions is $[b_0, b_1, b_2, b_3, \ldots]$.

Theorem 6.1 *Let x be an irrational real number. Then its Lagrange number $L(x)$ is < 3 if and only if x is equivalent to some number $x_w = [\chi(\tilde{w})^\infty]$ (or equivalently to $[\chi(w)^\infty]$) for some lower Christoffel word w. In this case, let m be the Markoff number $\mu(w)_{12} = \frac{1}{3}Tr(\mu(w))$ associated to w. Then $L(x) = L(x_w) = \sqrt{9 - \frac{4}{m^2}}$ and $x_w = \frac{p-s}{2m} + \frac{1}{2}\sqrt{9 - \frac{4}{m^2}}$ with $\mu(w) = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$ and therefore $m = q = \frac{1}{3}(p+s)$.*

Note that for some technical reasons, we have chosen x_w as in the statement. One could choose $[\chi(w)^\infty]$ instead, at the cost of transposing $\mu(w)$, or taking upper Christoffel words instead of lower ones. This is not fundamental.

Markoff's theorem is often stated as a series of progressively better approximations, with exceptions, the first exception being the golden ratio and the numbers equivalent to it. Formally, this is the following result. Informally, see the examples following it.

Corollary 6.1 *Let M be a finite set of Markoff numbers, such that for any Markoff numbers n, m, if $n < m$ and $m \in M$, then $n \in M$. Let W be the finite set of all lower Christoffel words corresponding to the Markoff numbers in M (in other words, $W = \{w \in W | \mu(w)_{12} \in M\}$). Let m be the smallest Markoff number not in M. Then for each irrational real number not equivalent to any x_w, $w \in W$, there are infinitely many rational approximations p/q of x such that $|x - p/q| < 1/\sqrt{9 - \frac{4}{m^2}}q^2$.*

We give three examples. First, let $M = \emptyset$. Then $m = 1$, $W = \emptyset$ and $\sqrt{9 - \frac{4}{m^2}} = \sqrt{5}$. We obtain that each real irrational number has infinitely many rational approximations satisfying $|x - p/q| < 1/\sqrt{5}q^2$. This is a theorem of Hurwitz [21], Satz 1 p. 279, that we state as corollary.

Corollary 6.2 *For each real number x there are infinitely many rational fractions $\frac{p}{q}$ such that $|x - p/q| < 1/\sqrt{5}q^2$.*

Now let $M = \{1\}$. Then $m = 2$, $W = \{a\}$, $\sqrt{9 - \frac{4}{m^2}} = \sqrt{8}$. Moreover $x_a = \frac{\sqrt{5}+1}{2}$, the *golden ratio*. We obtain that each real irrational number x not equivalent to x_a has infinitely many rational approximations p/q satisfying $|x - p/q| < 1/\sqrt{8}q^2$.

Finally, let $M = \{1, 2\}$. Then $W = \{a, b\}$, $m = 5$, $\sqrt{9 - \frac{4}{m^2}} = \frac{\sqrt{221}}{5}$, $x_b = 1 + \sqrt{2}$. We obtain that each irrational real number not equivalent to x_a nor to x_b has infinitely many rational approximations satisfying $|x - p/q| < 1/\frac{\sqrt{221}}{5}q^2$.

7 Markoff's Theorem for Quadratic Forms

For the results in these sections, see among others [10] Theorem II p.39, [17] Theorem 62 p.79 and seq., [14] theorem 6 p.10 (and also [33]).

A *real binary quadratic form* is a polynomial $f(x, y) = \alpha x^2 + \beta xy + \gamma y^2$ in the variables x, y and real coefficients α, β, γ not all zero. Its *discriminant* is $d(f) = \beta^2 - 4\alpha\gamma$. If the latter number is positive, the form is called *indefinite*.

We are interested here in the *greatest lower bound* of such a form, defined by $L(f) = \inf\{|f(x, y|, x, y \in \mathbb{Z}, (x, y) \neq (0, 0)\}$. We say that the lower bound is *attained* if there exist $(x, y) \in \mathbb{Z}^2 \setminus (0, 0)$ such that $L(f) = f(x, y)$. If the coefficients of f are integers (the interesting case), then the bound is clearly attained.

Two quadratic forms are *equivalent* if each of them is obtained from the other by a change of variables over $\mathbb{Z}$.

Let w be a lower Christoffel word and let $\mu(w) = \begin{pmatrix} p & m \\ r & s \end{pmatrix}$, where m is the Markoff number associated to w and therefore $m = \frac{1}{3}(p+s)$. Define the associated *Markoff quadratic form* by $f_w(x, y) = mx^2 + (s-p)xy - ry^2$.

Theorem 7.1 *Let $f(x, y)$ be a indefinite binary quadratic form. Assume that its greatest lower bound $L(f)$ and its discriminant $d(f)$ satisfy the inequality $\sqrt{d(f)} < 3L(f)$. Then f is equivalent to a multiple of some Markoff form f_w. Let m be the Markoff number associated to the lower Christoffel word w. One has $d(f_w) = 9 - 4m^2$, $L(f_w) = f_w(1, 0) = m$, so that the lower bounds of f_w and f are attained, and $\frac{\sqrt{d(f)}}{L(f)} = \frac{\sqrt{d(f_w)}}{L(f_w)} = \sqrt{9 - \frac{4}{m^2}}$.*

Remark 7.1 The first inequality in the theorem implies that $L(f)$ does not vanish (that is, there is no pair $(x, y) \neq (0, 0)$ of integers such that $f(x, y) = 0$). If we consider only quadratic forms satisfying this, then we see below (Corollary 7.2) that $\sqrt{d(f)}/L(f)$ is always at least equal to $\sqrt{5}$ (which is smaller that 3). Markoff's theorem is about such quadratic forms satisfying $\sqrt{d(f)}/L(f) < 3$. Note that there exist quadratic forms with $\sqrt{d(f)}/L(f)$ arbitrarily large: this set of real numbers, for all possible f, is called the *Markoff spectrum*, see [14].

This theorem is also stated as a series of progressively better inequalities, with exceptions, as follows.

Corollary 7.1 *Let M be a finite set of Markoff numbers, such that for any Markoff numbers n, m, if $n < m$ and $m \in M$, then $n \in M$. Let W be the finite set of all lower Christoffel words corresponding to the Markoff numbers in M (in other words, $W = \{w \in W | \mu(w)_{12} \in M\}$). Let m be the smallest Markoff number not in M. The for each indefinite binary quadratic form $f(x, y)$, not equivalent to a multiple of any form f_w, $w \in W$, one has $\sqrt{d(f)} \geq \sqrt{9 - \frac{4}{m^2}} L(f)$.*

We give several examples. First, let $M = \emptyset$. Then $m = 1$, $W = \emptyset$ and $\sqrt{9 - \frac{4}{m^2}} = \sqrt{5}$. We obtain a result due to Korkine and Zolotareff [24].

Corollary 7.2 *Let $f(x, y)$ be a indefinite binary quadratic form. Then $\sqrt{d(f)} \geq \sqrt{5}L(f)$.*

Now let $M = \{1\}$. Then $m = 2$, $W = \{a\}$, $\sqrt{9 - \frac{4}{m^2}} = \sqrt{8}$. We have $\mu(a) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ so that $f_a = x^2 - xy - y^2$. Thus we obtain the following result, also due to Korkine and Zolotareff [24].

Corollary 7.3 *Let $f(x, y)$ be a indefinite binary quadratic form. If f is not equivalent to a multiple of the form $f_a = x^2 - xy - y^2$, then $\sqrt{d(f)} \geq \sqrt{8}L(f)$.*

These two results are mentioned by Markoff [26], who gives them as the motivation of his own work. The next example is $M = \{1, 2\}$. Then $W = \{a, b\}$, $m = 5$, $\sqrt{9 - \frac{4}{m^2}} = \frac{\sqrt{221}}{5}$, $x_b = 1 + \sqrt{2}$. Since $\mu(b) = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$, we have $f_b = 2x^2 - 4xy - 2y^2$. Thus, for each form not equivalent to a multiple of $f_a = x^2 - xy - y^2$ nor of $f_b = 2x^2 - 4xy - 2y^2$, one has $\sqrt{d(f)} \geq \frac{\sqrt{221}}{5} L(f)$.

8 Several Binary Complete Infinite Trees

The set of Markoff triples may be organized as the set of nodes of an infinite binary tree; this follows from an operation on triples, already present in Markoff's work, that transforms each triple in three other ones: in the tree this corresponds to going to its parent, or to one of its two children. Other trees appear in the literature. All these trees are specialization of the first one, which we define now.

The nodes of the first tree are the pairs (u, v) where uv is a lower Christoffel word with its standard factorization. It was introduced by Jean Berstel and Aldo de Luca in [2]. Its root is (a, b) and the tree is constructed using the rule given in Fig. 2.

We call this tree the *tree of Christoffel pairs*, see Fig. 3.

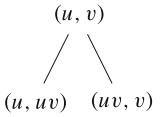


Fig. 2 The rule for constructing the tree of Christoffel pairs

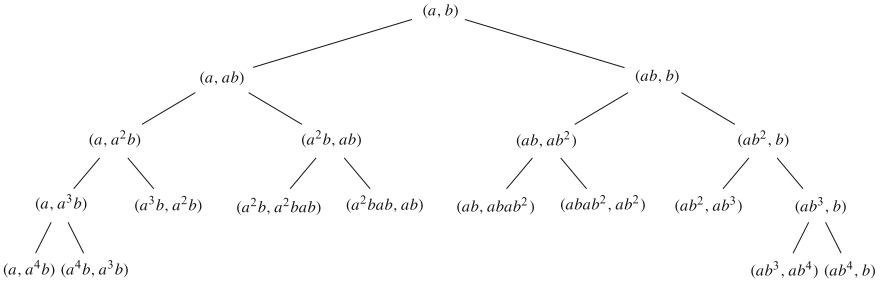


Fig. 3 The tree of Christoffel pairs

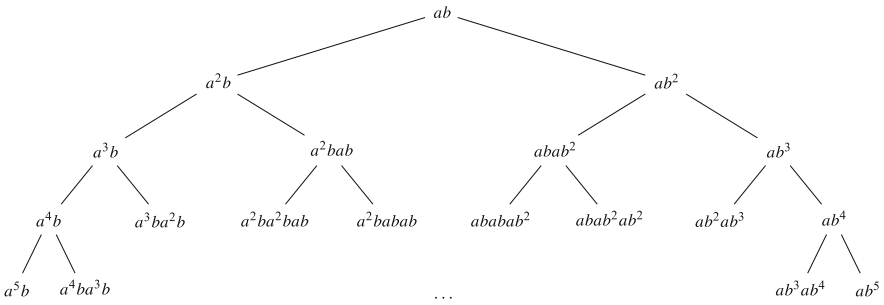


Fig. 4 The tree of Christoffel words

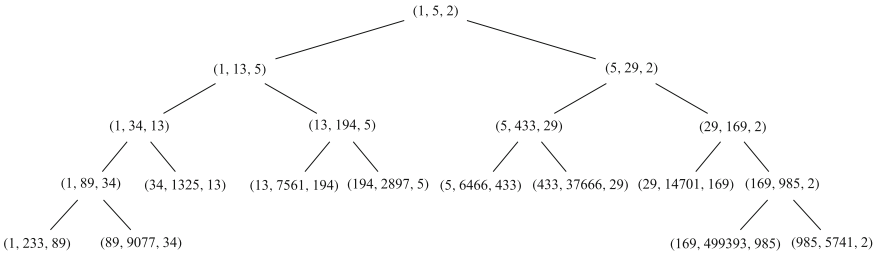


Fig. 5 The tree of Markoff triples

The second tree, the *tree of (lower) Christoffel words* is obtained from the previous one by replacing each node (u, v) by uv. Its nodes are exactly all lower Christoffel words. See Fig. 4.

This tree may constructed directly by taking as root ab, and any other node w is obtained as follows: consider the path from w towards the root.

(i) Suppose that w is not on the two extreme branches of the tree; then this path has north-west steps and north-east steps; let u be the node after the first north-west step and v be the node after the first north-east step. Then $w = uv$.
(ii) If w is on the left (respectively right) extreme branch, then no north-west (respectively north-east) step exists. Choose $u = a$ and v as in (i) (respectively u as in (i) and $v = b$).

For example, the node $w = a^2babab$ is the product of the nodes $u = a^2bab$ and $v = ab$.

The third tree is called the *tree of Markoff triples*. It is obtained by replacing each node (u, v) in the tree of Christoffel pairs by the triple $(\mu(u)_{12}, \mu(uv)_{12}, \mu(v)_{12}) = (\frac{1}{3}Tr(\mu(u)), \frac{1}{3}Tr(\mu(uv)), \frac{1}{3}Tr(\mu(v)))$. By Theorem 3.1, the nodes of this tree are the proper Markoff triples, each of which is represented by some ordered triple. See Fig. 5.

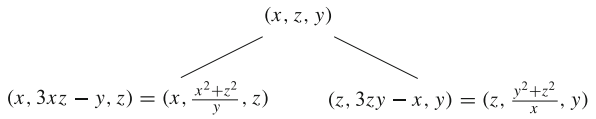


Fig. 6 The rule for building the tree of Markoff triples

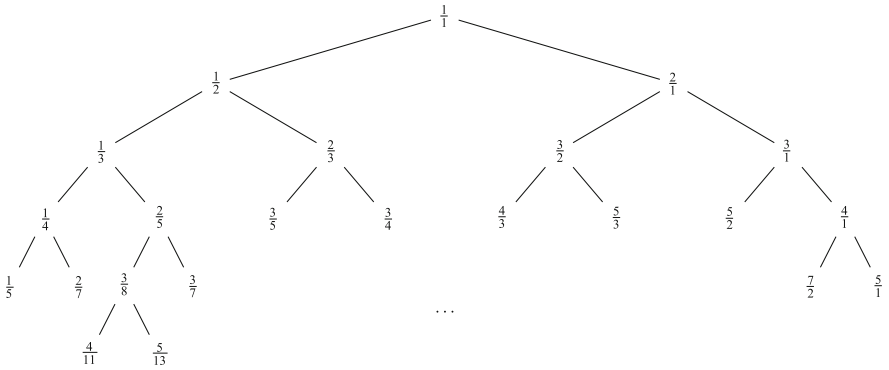


Fig. 7 The Stern-Brocot tree

This tree may be constructed directly by the rule given in Fig. 6, as follows from the Fricke relations, see Lemma 3.1.

The fourth tree, the *Stern-Brocot tree*, is obtained, following [2], from the tree of Christoffel pairs by replacing each node (u, v) by the *slope* of the word uv, that is the quotient of its number of b's divided by its number of a's. The nodes of the Stern-Brocot tree are the positive rational numbers. See Fig. 7.

The Stern-Brocot tree may be constructed directly, by mimicking the direct construction of the tree of Christoffel pairs: its root is $\frac{1}{1}$; consider the path from some other node $\frac{q}{p}$ to the root, and let $\frac{q'}{p'}$ (resp. $\frac{q''}{p''}$) be the node immediately after the first north-west (resp. north-east) step in this path (if no such step exists, then take $\frac{0}{1}$ resp. $\frac{1}{0}$). Then $\frac{q}{p}$ is the *mediant* of $\frac{q'}{p'}$ and $\frac{q''}{p''}$, that is $p = p' + p''$ and $q = q' + q''$. For example, $\frac{3}{4}$ is the mediant of $\frac{2}{3}$ and $\frac{1}{1}$, and $\frac{1}{3}$ is the mediant of $\frac{0}{1}$ and $\frac{1}{2}$.

The Stern-Brocot tree is a variant of the notion of continued fractions. It contains all *semi-convergents* of any real number. See [20].

The fifth tree is the *Raney tree* of [2] (see also [8]). It is obtained from the tree of Christoffel pairs by replacing each node (u, v) by the rational number $\frac{|u|}{|v|}$. The nodes of the Raney tree are the positive rational numbers (Fig. 8).

This tree may be constructed directly by applying the following rule: the root is $\frac{1}{1}$; if $\frac{q}{p}$ is a node, then its left child is $\frac{q}{p+q}$ and its right child is $\frac{p+q}{p}$. This follows directly from the rule for constructing the tree of Christoffel pairs, see Fig. 2.

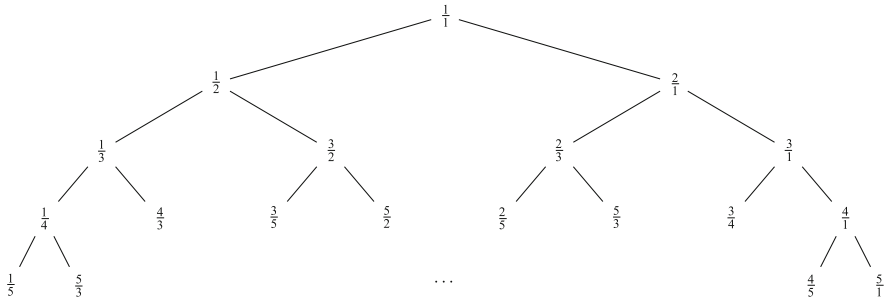


Fig. 8 The Raney tree

9 Frobenius Congruences Through Christoffel Words

In [19] p.602, Frobenius establishes several congruences related to a given Markoff triple, see also [1] p.33. We show that these congruences may be obtained through the standard factorization of Christoffel words.

Consider the monoid homomorphism ω from the free monoid $\{a, b\}^*$ into $SL_2(\mathbb{Z})$ defined by

$$\omega(a) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \omega(b) = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

This homomorphism is related to the homomorphism μ of Sect. 3 as follows.

Lemma 9.1 *Let* $P = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. *Then for any word* $x \in \{a, b\}^*$, *one has* $\mu(x) = P^{-1}(\omega \circ G(x))P$ *where* G *is the endomorphism of the free monoid* $\{a, b\}^*$ *sending* a *onto* a *and* b *onto* ab. *Moreover* $P^{-1}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ *and* $\begin{pmatrix} 1 & 1 \end{pmatrix} P = \begin{pmatrix} 2 & 1 \end{pmatrix}$.

We borrowed the notation G (meaning "gauche") from the notations for Sturmian morphisms as presented by Berstel and Séébold in [25].

Proof Clearly, $\begin{pmatrix} 1 & 1 \end{pmatrix} P = \begin{pmatrix} 2 & 1 \end{pmatrix}$. One has $P^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$. Hence $P^{-1}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Since $x \mapsto P^{-1}(\omega \circ G(x))P$ and μ are homomorphisms, it suffices to show

that they coincide for $x = a$ and $x = b$. One has

$$P^{-1}\omega(G(a))P = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \mu(a)$$

and

$$P^{-1}\omega(G(b))P = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix} = \mu(b).$$

□

It is easy to see that if $w = amb$ is a lower Christoffel word, then m is a palindrome. These palindromes are called *central words*. We use below the *iterated palindromization mapping*, denoted Pal: it is a bijection from $\{a, b\}^*$ onto the set of central words. One defines first the *palindromic closure* $u^{(+)}$ of a word: it is the shortest palindrome having u as prefix; it is determined by the equalities $u = ps$, $u^{(+)} = ps\tilde{p}$, where s is the longest palindromic suffix of u and where $\tilde{p}$ is the *reversal* of p. Then Pal is defined recursively by $Pal(1) = 1$ (the empty word) and $Pal(ux) = (Pal(u)x)^{(+)}$ for any word u and any letter x. All these notions and results are due to Aldo de Luca [15].

For example, for the lower Christoffel word $aabaabaabab$, the associated central word is the palindrome $abaabaaba$, which is equal to $Pal(abaa)$: indeed, $Pal(a) = a^{(+)} = a$, $Pal(ab) = (a\underline{b})^{+} = aba$, $Pal(aba) = (ab\underline{aa})^{+} = abaaba$ and $Pal(abaa) = (ab\underline{aabaa})^{+} = abaabaaba$, where the longest palindromic suffixes have been underlined.

The next result, which characterizes Markoff numbers, is due to Laurent Vuillon and the author [34].

Corollary 9.1 *Let v be any word in $\{a, b\}^*$ and w be the Christoffel word $w = aPal(v)b$. Then the Markoff number $\mu(w)_{12}$ is equal to the length of the lower Christoffel word $a(Pal \circ \theta \circ Pal(av))b$.*

Here θ denote the endomorphism of the free monoid that sends a onto ab and b onto ba, called the *Thue-Morse substitution*. Note that the mapping $\theta \circ Pal$ is the *iterated anti-palindromization mapping* (denoted $Anti\,Pal$) of [16], Theorem 7.1: an anti-palindrome in $\{a, b\}^*$ is a word which is equal to the word obtained by exchanging a and b in its reversal; for example $abbaab$. One defines $Anti\,Pal$

similarly to Pal, replacing in its definition palindromes by anti-palindromes. Hence, the previous result says that the mapping $v \mapsto 2+|Pal \circ Anti\, Pal(v)|$ is a surjection form the free monoid onto the set of Markoff numbers (distinct form 1,2). The question of its injectivity is precisely the Frobenius conjecture (see [34]).

We use below the following result (see e.g. Corollary 3.2 in [4]).

let ν be the monoid homomorphism from the free monoid $\{a, b\}^*$ into the multiplicative monoid $SL_2(\mathbb{N})$ defined by

$$\nu(a) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \nu(b) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Proposition 9.1 *Let $w = aPal(v)b$ be a lower Christoffel word and $w = w_1 w_2$ be its standard factorization. Let $\nu(v) = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$. Then the lengths of the words w_1, w_2 and $w_1 w_2$ are respectively $p + r$, $q + s$ and $p + q + r + s$.*

Proof of Corollary 9.1 In view of the previous proposition, it is enough to show that $\mu(aPal(v)b)_{12} = S\nu(\theta(Pal(av))$ where S sends each matrix onto the sum of its entries.

Note that $\nu\theta$ is equal to ω. Indeed,

$$\nu\theta(a) = \nu(ab) = \nu(a)\nu(b) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \omega(a).$$

Similarly $\nu\theta(b) = \omega(b)$.

Hence we have $S\nu\theta Pal(av) = S\omega Pal(av)$. Now we use the formula of Justin [22]: $Pal(av) = G(Pal(v))a$. Thus

$$S\omega Pal(av) = S(\omega G(Pal(v))\omega(a)) = \begin{pmatrix} 1 & 1 \end{pmatrix} \omega G(Pal(v)) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Note that the product of the two last matrices is $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$. Now, for any word x, we have

$$\begin{pmatrix} 1 & 1 \end{pmatrix} \omega G(x) \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \end{pmatrix} PP^{-1} \omega G(x) PP^{-1} \begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

This is equal by Lemma 9.1 to

$$\begin{pmatrix} 2 & 1 \end{pmatrix} P^{-1} \omega G(x) P \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \end{pmatrix} \mu(x) \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Moreover,

$$\mu(aPal(v)b)_{12} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \mu Pal(v) \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 1 \end{pmatrix} \mu Pal(v) \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

□

Corollary 9.2 *Consider a lower Christoffel word $w = w_1 w_2$ with its standard factorization, and $(m, m_1, m_2) = (\mu(w)_{12}, \mu(w_1)_{12}, \mu(w_2)_{12})$ be the corresponding Markoff triple. Let $w = aPal(v)b$. Consider the Christoffel word $u = a(Pal \circ \theta \circ Pal(av))b$ and let its standard factorization be $u = u_1 u_2$. Then the unique solution $x \in \{0, 1, \ldots, m-1\}$ of the congruence $m_1 x \equiv m_2 \mod m$ (resp. $m_2 x \equiv m_1 \mod m$) is $x = |u_2|$ (resp $x = |u_1|$).*

Since the length of u is m, by Corollary 9.1, the sum of the two solutions of the congruences is m: modulo m they are opposite. Moreover, by squaring we obtain $m_i^2 x^2 \equiv m_j^2 \mod m$, $\{i, j\} = \{1, 2\}$. By the Markoff equation we have $m_1^2 \equiv -m_2^2$, so that x is a square root of -1 modulo m.

An example is the following: $w = abb$, $w_1 = ab$, $w_2 = b$, $(m, m_1, m_2) = (29, 5, 2)$, $29^2 + 5^2 + 2^2 = 870 = 3 \cdot 29 \cdot 5 \cdot 2$, $v = b$, $Pal(ab) = aba$, $\theta Pal(ab) = abbaab$, $Pal\theta Pal(ab) = ababaababaababababaababaababa$, $u = aPal\theta Pal(ab)b = aababaababaababababaababaababab$, $u_1 = aababaababaababab$, $u_2 = aababaababab$, $|u_1| = 17$, $|u_2| = 12$, $m_1 \cdot |u_2| = 5 \cdot 12 = 60 = 2 \cdot 29 + 2 \equiv 2 = m_2 \mod 29$, $m_2 \cdot |u_1| = 2 \cdot 17 = 34 = 29 + 5 \equiv 5 = m_1 \mod 29$, $|u_1|^2 = 289 = -1 + 10 \cdot 29 \equiv -1 \mod 29$, $|u_2|^2 = 144 = -1 + 5 \cdot 29 \equiv -1 \mod 29$.

Proof Uniqueness follows from the fact that in a Markoff triple, the numbers are pairwise relatively prime ([1] Corollary 3.4). Let $\omega(Pal(av)) = \begin{pmatrix} h & i \\ j & k \end{pmatrix}$. It follows from the proposition and from the equality $\nu\theta = \omega$ that the lengths of u_1 and u_2 are respectively $h + j$ and $i + k$. From Lemma 9.1, we have $\omega(t) = P\mu G^{-1}(t)P^{-1}$ for any word t. By Justin's formula $G^{-1}(Pal(av)) = G^{-1}G(Pal(v))G^{-1}(a) = Pal(v)a$. Thus

$$\omega(Pal(av)) = P\mu(Pal(v))\mu(a)P^{-1} = P\mu(a)^{-1}\mu(aPal(v)b)\mu(b)^{-1}\mu(a)P^{-1}.$$

Since $\mu(a) = P^2$, we have $P\mu(a)^{-1} = P^{-1}$ and $\mu(b)^{-1}\mu(a)P^{-1} = \mu(b)^{-1}P = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 3 & -2 \end{pmatrix}$. Let $\mu(w) = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$. Thus

$$\omega(Pal(av)) = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} p & q \\ r & s \end{pmatrix}\begin{pmatrix} -1 & 1 \\ 3 & -2 \end{pmatrix}$$

$$= \begin{pmatrix} r & s \\ p-r & q-s \end{pmatrix}\begin{pmatrix} -1 & 1 \\ 3 & -2 \end{pmatrix} = \begin{pmatrix} -r+3s & r-2s \\ -p+r+3q-s & p-r-2q+2s \end{pmatrix}.$$

It follows that the length of u_1 is $-p+3q$ and that of u_2 is $p-2q$. Now, let $\mu(w_1) = \begin{pmatrix} p_1 & q_1 \\ r_1 & s_1 \end{pmatrix}$ and $\mu(w_2) = \begin{pmatrix} p_2 & q_2 \\ r_2 & s_2 \end{pmatrix}$. We have $\mu(w) = \mu(w_1)\mu(w_2)$ and thus (since the determinants are equal to 1) $\mu(w_1) = \mu(w)\mu(w_2)^{-1} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}\begin{pmatrix} s_2 & -q_2 \\ -r_2 & p_2 \end{pmatrix}$, from which follows $q_1 = -pq_2 + qp_2$. Similarly $q_2 = -p_1q + q_1p$. Note that $m = q, m_1 = q_1, m_2 = q_2$. We deduce that $m_1|u_2| - m_2 = q_1(p-2m) - q_2 = p_1m - 2q_1m = m(p_1 - 2q_1)$. Likewise $m_2|u_1| - m_1 = q_2(-p+3m) - q_1 = -mp_2 + 3mq_2 = m(3q_2 - p_2)$. □

References

1. Aigner, M.: Markov's Theorem and 100 Years of the Uniqueness Conjecture. Springer, Berlin (2013)
2. Berstel, J., de Luca, A.: Sturmian words, Lyndon words and trees. Theor. Comput. Sci. **178**, 171–203 (1997)
3. Berstel, J., Lauve, A., Reutenauer, C., Saliola, F.: Combinatorics on Words: Christoffel Words and Repetitions in Words. CRM Monograph Series. American Mathematical Society, Providence (2008)
4. Berthé, V., de Luca, A., Reutenauer, C.: On an involution of Christoffel words and Sturmian morphisms. Eur. J. Comb. **29**, 535–553 (2008)
5. Bombieri, E.: Continued fractions and the Markoff tree. Expo. Math. **25**, 187–213 (2007)
6. Borel, J.-P., Laubie, F.: Quelques mots sur la droite projective réelle. Journal de Théorie des Nombres de Bordeaux **5**, 23–51 (1993)
7. Brlek, S., Lachaud, J.O., Provençal, X., Reutenauer, C.: Lyndon + Christoffel = digitally convex. Pattern Recogn. **42**, 2239–2246 (2009)
8. Calkin, N., Wilf, H.S.: Recounting the rationals. Am. Math. Mon. **107**, 360–363 (2000)
9. Cassels, J.W.S.: The Markoff chain. Ann. Math. **50**, 676–685 (1949)
10. Cassels, J.W.S.: An Introduction to Diophantine Approximation. Cambridge University Press, Cambridge (1957)
11. Christoffel, E.B.: Observatio arithmetica. Annali di Matematica Pura ed Applicata **6**, 145–152 (1875)
12. Cohn, H.: Approach to Markoff's minimal forms through modular functions. Ann. Math. **61**, 1–12 (1955)

13. Cohn, H.: Growth types of Fibonacci and Markoff. Fibonacci Quart. **17**, 178–183 (1979)
14. Cusick, T.W., Flahive, M.E.: The Markoff and Lagrange Spectra. American Mathematical Society, Providence (1989)
15. de Luca, A.: Sturmian words: structure, combinatorics, and their arithmetics. Theor. Comput. Sci. **183**, 45–82 (1997)
16. de Luca, A., de Luca, A.: Pseudopalindrome closure operators in free monoids. Theor. Comput. Sci. **362**(1–3), 282–300 (2006)
17. Dickson L.E.: Studies in the Theory of Numbers. Chelsea, New York (1957) (first edition 1930)
18. Fricke, R.: Ueber die Theorie der automorphen Modulgruppen. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen 91–101 (1896)
19. Frobenius, G.F.: Über die Markoffschen Zahlen. Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin **26**, 458–487 (1913)
20. Graham, R.L., Knuth, D., Patashnik, O.: Concrete Mathematics, 2nd edn. Addison Wesley, Reading (1994)
21. Hurwitz, A.: Ueber die angenäherte Darstellung der Irrationalzahlen durch rational Brüche. Mathematische Annalen **39**, 279–284 (1891)
22. Justin, J.: Episturmian morphisms and a Galois theorem on continued fractions. Theor. Inform. Appl. **39**, 207–215 (2005)
23. Klette, R., Rosenfeld, A.: Digital straightness – a review. Discret. Appl. Math. **139**, 197–230 (2004)
24. Korkine, A., Zolotareff, G.: Sur les formes quadratiques. Mathematische Annalen **6**, 366–389 (1873)
25. Lothaire, M.: Algebraic Combinatorics on Words. Cambridge University Press, Cambridge (2002)
26. Markoff, A.A.: Sur les formes quadratiques binaires indéfinies. Mathematische Annalen **15**, 381–496 (1879)
27. Markoff, A.A.: Sur les formes quadratiques binaires indéfinies (second mémoire). Mathematische Annalen **17**, 379–399 (1880)
28. Morse, M., Hedlund, G.A.: Symbolic dynamics II: sturmian trajectories. Am. J. Math. **62**, 1–42 (1940)
29. Perrine, S.: La théorie de Markoff et ses développements. Tessier et Ashpool, Chantilly (2002)
30. Perron, O.: Über die Approximationen irrationaler Zahlen durch rationale II. Sitzungsbereich der Heidelberger Akademie der Wissenschaften **8**, 2–12 (1921)
31. Remak, R.: Über indefinite binäre quadratische Minimalformen. Mathematische Annalen **92**, 155–182 (1924)
32. Reutenauer, C.: Christoffel words and Markoff triples. Integers **9**, 327–332 (2009)
33. Reutenauer, C.: From Christoffel Words to Markoff Numbers. Oxford University Press (2018, to appear)
34. Reutenauer, C., Vuillon, L.: Palindromic closures and Thue-Morse substitution for Markoff numbers. Unif. Distrib. Theory **12**, 25–35 (2017)
35. Series, C.: The geometry of Markoff numbers. Math. Intell. **7**, 20–29 (1985)
36. Smith, H.J.S.: Note on continued fractions. Messenger Math. **6**, 1–14 (1876)

An Algebraic Approach to Integration of Geometric Rough Paths

Danyu Yang

Abstract We build a connection between rough path theory and a non-commutative algebra, and interpret the integration of geometric rough paths as an example of a non-abelian Young integration. We identify a class of slowly-varying one-forms, and prove that the class is stable under basic operations.

Consider two Lie groups G_1 and G_2, and a differentiable function $f : G_1 \to G_2$. For a time interval $[S, T]$ and a differentiable path $X : [S, T] \to G_1$, the integration of the exact one-form df along X can be defined as:

$$\int_{r=S}^{T} df dX_r := \int_{X_S}^{X_T} df = f(X_S)^{-1} f(X_T) \in G_2.$$

When f and X are only continuous, $\int_{r=S}^{T} df dX_r$ can also be defined as $f(X_S)^{-1} f(X_T)$.

Consider a time-varying exact one-form $\left(df_t\right)_t$ with $f_t : G_1 \to G_2$ indexed by $t \in [S, T]$, and $X : [S, T] \to G_1$. If the following limit exists in G_2:

$$\lim_{|D| \to 0, D \subset [S,T]} \int_{r=t_0}^{t_1} df_{t_0} dX_r \int_{r=t_1}^{t_2} df_{t_1} dX_r \cdots \int_{r=t_{n-1}}^{t_n} df_{t_{n-1}} dX_r$$

The research was carried out at the Mathematical Institute University of Oxford and supported by the European Research Council grant nr 291244.

D. Yang (✉)
Department of Mathematical Sciences, NTNU, Trondheim, Norway
e-mail: danyu.yang@ntnu.no

E. Celledoni et al. (eds.), *Computation and Combinatorics in Dynamics, Stochastics and Control*, Abel Symposia 13,
https://doi.org/10.1007/978-3-030-01593-0_25

where $D = \{t_k\}_{k=0}^{n}, S = t_0 < \cdots < t_n = T, n \geq 1$ with $|D| := \max_{k=0}^{n-1} |t_{k+1} - t_k|$, then the integral $\int_{r=S}^{T} df_r dX_r$ is defined to be the limit. The integral can be viewed as an non-abelian counterpart of Young integral [30].

We interpret the integration of Lipschitz one-forms along geometric rough paths developed by Lyons [16] as an integration of time-varying exact one-forms along group-valued paths. The interpretation is in the language of Malvenuto–Reutenauer Hopf algebra of permutations [20, 21]. In particular, we view the dichotomization of the associative multiplication in a dendriform algebra [14] as an abstract integration by parts formula.

Notation

$S(x)$	the signature of a continuous bounded variation path x
K	a commutative ring with unit 1_K
V	a vector space over K
$T((V))$	formal tensor series of V; an associative algebra with the operation of tensor product
$\triangle$	homomorphism of K-algebras $T((V)) \to T((V)) \otimes T((V))$ given by $\triangle v = v \otimes 1_K + 1_K \otimes v$ for each $v \in V$ with the operation on $T((V))$ given by tensor product
$G(V)$	group-like elements in $T((V))$, i.e. the set of elements $a \in T((V))$ that satisfy $\triangle a = a \otimes a$
$BV(J, V)$	the set of continuous bounded variation paths $J \to V$
A	a (possibly infinite) set of non-commutative indeterminates
A^*	the free monoid generated by A
e	the empty sequence in A^*
$\|w\|$	the number of letters in $w \in A^*$
$K\langle A\rangle$	the set of non-commutative polynomials on A over K
$K\langle\langle A\rangle\rangle$	the set of formal series on A over K
$(s, w), (p, w)$	the coefficient of $w \in A^*$ in $s \in K\langle\langle A\rangle\rangle$ and in $p \in K\langle A\rangle$
$\sqcup\!\sqcup$	shuffle product $K\langle A\rangle \times K\langle A\rangle \to K\langle A\rangle$
δ'	deconcatenation coproduct $K\langle A\rangle \to K\langle A\rangle \otimes K\langle A\rangle$
$conc$	concatenation product $K\langle\langle A\rangle\rangle \times K\langle\langle A\rangle\rangle \to K\langle\langle A\rangle\rangle$
δ	homomorphism of K-algebras $K\langle\langle A\rangle\rangle \to K\langle\langle A\rangle\rangle \otimes K\langle\langle A\rangle\rangle$ given by $\delta(a) = a \otimes e + e \otimes a$ for each $a \in A$ with the operation on $K\langle\langle A\rangle\rangle$ given by concatenation
$G(A)$	group-like elements in $K\langle\langle A\rangle\rangle$, i.e. the set of elements $s \in K\langle\langle A\rangle\rangle$ that satisfy $\delta s = s \otimes s$
MR	Malvenuto–Reutenauer Hopf algebra of permutations
S_n	the symmetric group of order n for $n \geq 1$, and $S_0 = \{\lambda\}$
1_n	the identity element of S_n for $n \geq 1$, and $1_0 := \lambda \in S_0$
S	$\cup_{n\geq 0} S_n$ permutations

$*'$	$\mathbb{Z}S \times \mathbb{Z}S \to \mathbb{Z}S$ the product of MR
$\triangle'$	$\mathbb{Z}S \to \mathbb{Z}S \otimes \mathbb{Z}S$ the coproduct of MR
$\prec, \succ$	dendriform algebra operators on $K\langle A\rangle$ and on $\mathbb{Z}S$
$m_{\succ}(,\ldots,)$	$m_{\succ}(p_1) := p_1$,
	$m_{\succ}(p_1,\ldots,p_n) := \left(\cdots\left(p_1 \succ p_2\right) \succ \cdots \succ p_{n-1}\right) \succ p_n$,
	for $p_i, i = 1,\ldots,n$, in $K\langle A\rangle$ or in $\mathbb{Z}S$

1 Background: Rough Path Theory

The theory of rough path [16] provides a mathematical tool for modelling the interaction and evolution of highly oscillating systems that include but are not restricted to Brownian motion and semi-martingales.

In its basic form, the theory is related to the classical integration developed by Young [30]. Young proved that, for $x : [0,1] \to \mathbb{C}$ of finite p-variation and $y : [0,1] \to \mathbb{C}$ of finite q-variation, $p \geq 1$, $q \geq 1$, $p^{-1} + q^{-1} > 1$, the Stieltjes integral

$$\int_{t=0}^{1} x_t dy_t$$

is well defined[1] as the limit of Riemann sums. The definition of p-variation dates back to Wiener [28]:

$$\|x\|_{p-var,[0,1]} := \sup_{D\subset[0,1]} \left(\sum_{k,t_k\in D} \left\|x_{t_{k+1}} - x_{t_k}\right\|^p\right)^{\frac{1}{p}}$$

where the supremum is over all finite partitions $D = \{t_k\}_{k=0}^{n}, 0 = t_0 < \cdots < t_n = 1, n \geq 1$. The condition given by Young is sharp: the Riemann–Stieltjes integral $\int x dy$ does not necessarily exist when $p^{-1}+q^{-1} = 1$ [30]. Lyons [15] demonstrated that similar obstacles exist in stochastic integration, and one needs to consider x and y as a "pair" " in a fairly strong way".

From a different perspective, Chen [4–6] investigated the iterated integration of one-forms and developed a theory of cohomology for loop spaces. One of the major objects he studied is non-commutative formal series with coefficients given by the iterated integrals of a path. For a time interval $[S,T]$, let $x : [S,T] \to \mathbb{R}^d$ be a smooth path, and let $X_1,\ldots,X_d$ be non-commutative indeterminates. Consider the

[1] At least when x and y have no common jumps.

formal power series:

$$\theta(x) := 1 + \sum_{n \geq 1, i_j \in \{1,\ldots,d\}} w_{i_1 \cdots i_n} X_{i_1} \cdots X_{i_n}$$

where

$$w_{i_1 \cdots i_n} := \int \cdots \int_{S < u_1 < \cdots < u_n < T} dx_{u_1}^{i_1} \cdots dx_{u_n}^{i_n}.$$

The space of paths in $\mathbb{R}^d$ has an associative multiplication given by concatenation, and the set of formal series has an associative multiplication given by the bilinear extension of the concatenation of finite sequences. Chen [4] proved that θ is an algebra homomorphism from paths to formal series. Based on [5] θ takes values in a group whose elements are algebraic exponentials of Lie series, and the multiplication in the group is given by Campbell–Baker–Hausdorff formula.

In [16], Lyons developed the theory of rough paths. He observed that, in controlled systems, both the driving path and the solution path evolve in a group (denoted as G) rather than in a vector space. He also identified a family of metrics on group-valued paths such that the Itô map that sends a driving path to the solution path is continuous. Elements in G are algebraic exponentials of Lie series. A geometric p-rough path is a continuous path in G with finite p-variation.

For a continuous bounded variation path, there exists a canonical lift of the path to a geometric rough path given by the sequence of indefinite iterated integrals. The sequence of definite iterated integrals is called the signature of the continuous bounded variation path.

The following is Definition 1.1 by Hambly and Lyons [12].

Definition 1 (Signature) Let x be a path of bounded variation on $[S, T]$ with values in a vector space V. Then its signature is the sequence of definite iterated integrals

$$\begin{aligned} X_{S,T} &= \left(1 + X_{S,T}^1 + \cdots + X_{S,T}^k + \cdots\right) \\ &= \left(1 + \int_{S<u<T} dx_u + \cdots + \int_{S<u_1<\cdots<u_k<T} dx_{u_1} \otimes \cdots \otimes dx_{u_k} + \cdots\right) \end{aligned}$$

regarded as an element of an appropriate closure of the tensor algebra $T(V) = \bigoplus_{n=0}^{\infty} V^{\otimes n}$.

The signature is invariant under reparametrisations of the path.

Following [12] $X_{S,T}$ is also denoted by $S(x)$ where S is the signature mapping that sends a continuous bounded variation path x to the sequence of definite iterated integrals.

Notation 1 Denote by $S : x \mapsto S(x)$ the signature mapping.

The signature provides an efficient and effective description of the information encoded in paths, and two paths with the same signature have the same effect on all controlled systems. An important problem in the theory of rough path is the 'uniqueness of signature' problem: to describe the kernel of the signature mapping. This problem dates back to Chen [6] where he proved that the map θ (signature) provides a faithful representation for a family of paths that are piecewise regular, continuous and irreducible. In [12] Hambly and Lyons established quantitative estimates of a path in terms of its signature, and proved the uniqueness of signature result for continuous bounded variation paths. In [3] Boedihardjo, Geng, Lyons and Yang extended the result to weakly geometric rough paths over Banach spaces. There are also progresses in the direction of machine learning, where rough paths have been introduced as a new feature set for streamed data [19, 29].

Let K be a commutative ring with unit 1_K. Let V be a K-vector space.

Notation 2 Let $T((V))$ denote the formal tensor series on V over K.

$x \in T((V))$ if and only if $x = \sum_{k=0}^{\infty} x^k$ for $x^k \in V^{\otimes k}, k \geq 1$ and $x^0 \in K$. $T((V))$ is the completion of $T(V)$ in a distance [2, Chapter 1, Section 3]. For $x, y \in T((V))$, define their product $xy \in T((V))$ as $(xy)^k := \sum_{j=0}^{k} x^j \otimes y^{k-j}$. The product is associative with a unit. Let $\triangle$ denote the homomorphism of K-algebras $T((V)) \to T((V)) \otimes T((V))$ given by $\triangle v = 1_K \otimes v + v \otimes 1_K$ for each $v \in V$ with the operation on $T((V))$ given by tensor product.

Definition 2 The set of group-like elements in $T((V))$ is the set of elements $a \in T((V))$ that satisfy $\triangle a = a \otimes a$.

Notation 3 Let $G(V)$ denote the set of group-like elements in $T((V))$.

It is well known that $G(V)$ is a group [24, Theorem 3.2].

Let V be a Banach space. A time interval is an interval of the form $[S, T]$ for $0 \leq S \leq T < \infty$. For a time interval J, let $BV(J, V)$ denote the set of continuous bounded variation paths $J \to V$. For $x \in BV(J_1, V)$ and $y \in BV(J_2, V)$, let $x * y$ denote the concatenation of x with y obtained by translating the starting point of y to be the end point of x, and let $\overleftarrow{x}$ denote the time-reversal of x. Based on Chen [4], the image of $BV(J, V)$ under S is a subgroup of $G(V)$:

$$S(x * y) = S(x) S(y) \text{ and } S\left(\overleftarrow{x}\right) = S(x)^{-1}.$$

Suppose the tensor powers of V are equipped with admissible norms [18, Definition 1.25]. Following [3, Definition 2.1], we equip $G(V)$ with the metric

$$d(a, b) := \max_{k \in \mathbb{N}} \left\| \pi_k\left(a^{-1} b\right) \right\|^{\frac{1}{k}}$$

for $a, b \in G(V)$, where π_k denotes the projection of $T((V))$ to $V^{\otimes k}$. Based on [16, Definition 1.2.2] for a time interval J, $\omega : \{(s, t) | s \leq t, s \in J, t \in J\} \to \mathbb{R}$ is a

control if ω is non-negative, super-additive: $\omega(s,t) \leq \omega(s,u) + \omega(u,t)$ for every $s \leq u \leq t$, continuous and vanishes on the diagonal.

The following definition is based on [16, Definition 1.2.2].

Definition 3 (Geometric Rough Paths) For a time interval J and a Banach space V, $X : J \to G(V)$ is called a geometric p-rough path for some $p \geq 1$, if there exists a control $\omega : \left\{(s,t) \,|s \leq t, s \in J, t \in J\right\} \to \mathbb{R}$ such that

$$\|X\|^p_{p-var,[s,t]} \leq \omega(s,t)$$

for every $s \leq t$, where

$$\|X\|^p_{p-var,[s,t]} := \sup_{D \subset [s,t]} \sum_{k, t_k \in D} d\left(X_{t_k}, X_{t_{k+1}}\right)^p,$$

with the supremum over all $D = \{t_k\}_{k=0}^n, s = t_0 < \cdots < t_n = t, n \geq 1$.

2 Background: Shuffle Algebra and Malvenuto–Reutenauer Algebra

The introduction of shuffles dates back to Eilenberg and MacLane [8]. The shuffle product can be expressed in terms of permutations based on the Malvenuto–Reutenauer Hopf algebra (denoted by MR) introduced in [20, 21]. MR is a $\mathbb{Z}$-Hopf algebra on permutations $S := \cup_{n \geq 0} S_n$ with $S_0 = \{\lambda\}$, and is non-commutative.

We first review the shuffle Hopf algebra and its dual space based on Reutenauer [24, Section 1.5]. Let A be a (possibly infinite) set, and let K be a commutative $\mathbb{Q}$-algebra. Let $K\langle A\rangle$ denote the set of non-commutative polynomials on A over K. Let A^* denote the free monoid generated by A. A^* is the set of finite sequences of elements in A including the empty sequence denoted by e. The operation on A^* is given by associative concatenation:

$$(a_1 \cdots a_n)\left(a_{n+1} \cdots a_{n+m}\right) := a_1 \cdots a_{n+m}$$

for $a_i \in A$, with $e \in A^*$ the identity element. There is a natural embedding $A \hookrightarrow A^*$. Based on Ree [23] and Schützenberger [26], define $\sqcup\!\sqcup : K\langle A\rangle \times K\langle A\rangle \to K\langle A\rangle$ as the K-bilinear map given recursively by

$$w_1 a_1 \sqcup\!\sqcup w_2 a_2 := (w_1 a_1 \sqcup\!\sqcup w_2)\, a_2 + (w_1 \sqcup\!\sqcup w_2 a_2)\, a_1$$

for $w_i \in A^*, a_i \in A$, where wa denotes the concatenation of $w \in A^*$ with $a \in A$, and $e \sqcup\!\sqcup w = w \sqcup\!\sqcup e := w$ for $w \in A^*$. The product $\sqcup\!\sqcup$ is associative and commutative, with unit $u(k) = ke$ for $k \in K$. Let $\delta' : K\langle A\rangle \to K\langle A\rangle \otimes K\langle A\rangle$ denote the

deconcatenation coproduct, that is the K-linear map given by

$$\delta'(a_1 \cdots a_n) := \sum_{k=0}^{n} a_1 \cdots a_k \otimes a_{k+1} \cdots a_n$$

for $a_i \in A$, $n \geq 1$, and $\delta'(e) := e \otimes e$. The counit ϵ is the projection of $K\langle A\rangle$ to the space spanned by $e \in A^*$. $\left(K\langle A\rangle, \sqcup\!\sqcup, u, \delta', \epsilon\right)$ is a Hopf algebra [24, p.31] which we call the shuffle Hopf algebra.

Let $K\langle\langle A\rangle\rangle$ denote the set of formal series on A over K. An element $s \in K\langle\langle A\rangle\rangle$ can be written in the form $s = \sum_{w \in A^*} (s, w)\, w$ for $(s, w) \in K$. The concatenation product $conc : K\langle\langle A\rangle\rangle \otimes K\langle\langle A\rangle\rangle \to K\langle\langle A\rangle\rangle$ is the K-bilinear map given by

$$\left(conc \circ (s \otimes t), w\right) := \sum_{uv=w} (s, u)\,(t, v)$$

for $w \in A^*$. The map $\delta : K\langle\langle A\rangle\rangle \to K\langle\langle A\rangle\rangle \otimes K\langle\langle A\rangle\rangle$ is the homomorphism of K-algebras given by $\delta(a) := e \otimes a + a \otimes e$ for each $a \in A$ with the operation on $K\langle\langle A\rangle\rangle$ given by concatenation, and is of the explicit form [24, p.25]

$$\delta(s) = \sum_{w_1, w_2 \in A^*} (s, w_1 \sqcup\!\sqcup w_2)\, w_1 \otimes w_2$$

for $s \in K\langle\langle A\rangle\rangle$. $K\langle\langle A\rangle\rangle$ is nearly a Hopf algebra, but the map δ does not necessarily take values in $K\langle\langle A\rangle\rangle \otimes K\langle\langle A\rangle\rangle$ and the sum in $\delta(s) = \sum_{(s)} s_{(1)} \otimes s_{(2)}$ can be infinite [24, p.38].

There is a pairing between $K\langle\langle A\rangle\rangle$ and $K\langle A\rangle$ as vector spaces given by

$$\left(s, p\right) := \sum_{w \in A^*} (s, w)\left(p, w\right)$$

for $s \in K\langle\langle A\rangle\rangle$ and $p \in K\langle A\rangle$. Then the following identities hold [24, p.26]:

$$\left(s, p \sqcup\!\sqcup q\right) = \left(\delta s, p \otimes q\right)$$
$$\left(conc \circ (s \otimes t), p\right) = \left(s \otimes t, \delta' p\right)$$

for $s, t \in K\langle\langle A\rangle\rangle$ and $p, q \in K\langle A\rangle$.

Definition 4 Group-like elements in $K\langle\langle A\rangle\rangle$ are elements $s \in K\langle\langle A\rangle\rangle$ that satisfy $\delta s = s \otimes s$.

Notation 4 Let $G(A)$ denote the set of group-like elements in $K\langle\langle A\rangle\rangle$.

$G(A)$ is the set of algebraic exponentials of Lie series on A over K [24, Theorem 3.2], and $G(A)$ is a group [24, Corollary 3.3]. When a vector space V has a basis A, $G(V)$ can be represented as $G(A)$. The set of group-like elements

in $K\langle\langle A\rangle\rangle$ is the set of characters of the shuffle algebra on $K\langle A\rangle$ [24, Theorem 3.2]: $s \in G(A)$ if and only if $(s, p)(s, q) = (s, p \sqcup\!\sqcup q)$ for every $p, q \in K\langle A\rangle$.

Let $\mathrm{End}(K\langle A\rangle)$ denote the K-module of linear endomorphisms of $K\langle A\rangle$. Based on Reutenauer [24, Proposition 1.10], $\mathrm{End}(K\langle A\rangle)$ becomes a K-associative algebra with the convolution product $*'$ given by

$$f *' g := \sqcup\!\sqcup \circ (f \otimes g) \circ \delta'$$

for $f, g \in \mathrm{End}(K\langle A\rangle)$. There is an embedding of permutations $\mathbb{Z}S = \bigoplus_{n\geq 0} \mathbb{Z}S_n$ in $\mathrm{End}(K\langle A\rangle)$ given by

$$\sigma \cdot (a_1 \cdots a_n) := \delta_{n,m} a_{\sigma(1)} \cdots a_{\sigma(n)}$$

for $\sigma \in S_m$ and $a_i \in A$. The product $*'$ is closed on $\mathbb{Z}S$. The multiplication of MR is the $\mathbb{Z}$-bilinear map $*' : \mathbb{Z}S \times \mathbb{Z}S \to \mathbb{Z}S$, and is of the explicit form [21]:

$$\sigma *' \rho = \sigma \sqcup\!\sqcup \bar{\rho}$$

for $\sigma \in S_n$ and $\rho \in S_m$, where permutations are considered as words with $\bar{\rho}(i) := n + \rho(i)$. The product $*'$ is associative with identity element $\lambda \in S_0$. The coproduct on MR is the $\mathbb{Z}$-linear map $\triangle' : \mathbb{Z}S \to \mathbb{Z}S \otimes \mathbb{Z}S$ given by

$$\triangle' := (\mathrm{st} \otimes \mathrm{st}) \circ \delta'$$

where 'st' denotes the unique increasing map that sends a sequence of k non-repeating integers to $\{1, 2, \ldots, k\}$ for $k \geq 1$, e.g. $\mathrm{st}(283) = 132$. The counit is the projection of $\mathbb{Z}S$ to the space $\mathbb{Z}S_0$. Based on [20–22] MR is a Hopf algebra that is self-dual, free and cofree, so is neither commutative nor cocommutative.

For $f \in \mathrm{End}(K\langle A\rangle)$, define the adjoint map $f^* \in \mathrm{End}(K\langle\langle A\rangle\rangle)$ as

$$(f^* s, p) := (s, fp)$$

for every $s \in K\langle\langle A\rangle\rangle$ and $p \in K\langle A\rangle$. As a sub-algebra of $\mathrm{End}(K\langle A\rangle)$, MR induces a sub-algebra of $\mathrm{End}(K\langle\langle A\rangle\rangle)$. Proposition 1 below states that $G(A)$ is a subgroup of the group of characters of MR: for $s \in G(A)$,

$$\begin{aligned} \widehat{s} : (\mathbb{Z}S, *') &\to (K\langle\langle A\rangle\rangle, conc) \\ \sigma &\mapsto \sigma^* s \end{aligned}$$

is an algebra homomorphism. Proposition 1 is closely related to the cotensor algebra [25, p.248].

For $w \in A^*$, let $|w|$ denote the number of letters in w.

Proposition 1 *For $s \in G(A)$, define a $\mathbb{Z}$-linear map $\widehat{s} : \mathbb{Z}S \to K\langle\langle A\rangle\rangle$ as*

$$\widehat{s}(\sigma) := \sum_{w \in A^*} (s, \sigma \cdot w)\, w. \tag{1}$$

Then

$$conc \circ \left(\widehat{s}(\sigma) \otimes \widehat{s}(\rho)\right) = \widehat{s}\left(\sigma *' \rho\right)$$

for $\sigma, \rho \in \mathbb{Z}S$.

Proof Since $\sigma *' \rho := \sqcup\!\sqcup \circ \left(\sigma \otimes \rho\right) \circ \delta'$,

$$\left(\sigma *' \rho\right) \cdot w = (\sigma \cdot u) \sqcup\!\sqcup \left(\rho \cdot v\right)$$

for $\sigma \in S_n, \rho \in S_m, w \in A^*, w = uv, |u| = n, |v| = m$. Since $s \in G(A)$ is a character of the shuffle algebra on $K\langle A\rangle$,

$$\left(s, \left(\sigma *' \rho\right) \cdot w\right) = \left(s, (\sigma \cdot u) \sqcup\!\sqcup \left(\rho \cdot v\right)\right) = (s, \sigma \cdot u)\left(s, \rho \cdot v\right).$$

The statement follows.

3 Dendriform Algebra and Iterated Integration

In a study of Hall basis [26], Schützenberger introduced a binary operator T, and expressed shuffle product as the symmetrization of T. Based on Chen's representation of paths [5], Ree [23] introduced the algebra for study of Lie polynomials. Ree gave a recursive definition of the shuffle product as the sum of two operations (operator T). The dichotomization of shuffle product has been applied in control theory by Kawski and Sussmann [13] and by others.

The dichotomization of the shuffle product is incorporated in an algebraic framework developed by Loday [14] in his work on cohomology for dialgebras. A dual dialgebra is a dendriform algebra [14]. A dendriform algebra is an associative algebra whose multiplication allows a consistent dichotomization [14, $(i)\,(ii)\,(iii)$ p.8]. Ebrahimi-Fard [7] studied a relation between dendriform algebra and Rota-Baxter algebra (see Aguiar [1] for the weight zero case), and built a link between dendriform algebra and abstract integration.

The shuffle algebra has a dendriform algebra structure[2] (with $\sqcup\!\sqcup = \prec + \succ$), and is the free commutative dendriform algebra [14, Section 7], i.e. $a \succ b := b \prec a$ for

[2] As Loday commented, the dendriform structure of shuffle algebra had been previously remarked by Rota.

all elements a, b. We view the dichotomization of the associative multiplication of a dendriform algebra as an abstract integration by parts formula, and view dendriform operations as abstract iterated integrations.

Notation 5 Denote $\succ$: $(K\langle A\rangle \times K\langle A\rangle)\setminus(Ke\times Ke)\to K\langle A\rangle$ as the K-bilinear map given by

$$\begin{aligned}&(a_1\cdots a_n)\succ\left(a_{n+1}\cdots a_{n+m}\right)\\ :=\ &\left((a_1\cdots a_n)\sqcup\!\sqcup\left(a_{n+1}\cdots a_{n+m-1}\right)\right)a_{n+m}\end{aligned}$$

for $a_i\in A, m\geq 1$, where for $w\in A^*$ and $a\in A$, wa denotes the concatenation of $w\in A^*$ with $a\in A$; $(a_1\cdots a_n)\succ e:=0\in K$ for $a_i\in A, n\geq 1$.

In defining the integration of geometric rough paths, Lyons considered an ordered shuffle to define almost multiplicative functionals [16, p.285, Definition 3.2.2]. The ordered shuffle can be viewed as iterated applications of $\succ$. Consider $p_1,\ldots,p_n\in K\langle A\rangle$ that satisfy $\left(p_i,e\right)=0, i=1,\ldots,n$, where $p=\sum_{w\in A^*}\left(p,w\right)w$ and e is the empty sequence in A^*. Define [25, Notation 2.3]

$$\begin{aligned}m_\succ\left(p_1\right):&=p_1\\ m_\succ\left(p_1,\ldots,p_n\right):&=\left(\cdots\left(p_1\succ p_2\right)\succ\cdots\succ p_{n-1}\right)\succ p_n.\end{aligned}$$

The following Lemma helps to prove that indefinite integrals of one-forms along geometric rough paths are geometric rough paths.

Let 1_n denote the identity element of S_n for $n\geq 1$, and $1_0:=\lambda\in S_0$.

Lemma 1 *Let $p_1,\ldots,p_{n+m}\in K\langle A\rangle$ satisfy $\left(p_i,e\right)=0, i=1,\ldots,n+m$. Then*

$$\begin{aligned}&m_\succ\left(p_1,\ldots,p_n\right)\sqcup\!\sqcup m_\succ\left(p_{n+1},\ldots,p_{n+m}\right)\\ =&\sum_{\rho\in 1_n *' 1_m} m_\succ\left(p_{\rho(1)},\ldots,p_{\rho(n+m)}\right).\end{aligned}$$

Lemma 1 holds for a general commutative dendriform algebra. The proof is based on a recursive definition of the shuffle product for elements of the form $m_\succ\left(p_1,\ldots,p_n\right)$: for $a_i, p_i\in K\langle A\rangle, \left(p_i,e\right)=0$,

$$\begin{aligned}&\left(a_1\succ p_1\right)\sqcup\!\sqcup\left(a_2\succ p_2\right)\\ =&\left(\left(a_1\succ p_1\right)\sqcup\!\sqcup a_2\right)\succ p_2+\left(\left(a_2\succ p_2\right)\sqcup\!\sqcup a_1\right)\succ p_1.\end{aligned}\tag{2}$$

Proof Suppose $n=1$. When $n=1, m=1$, the statement holds. Suppose the statement holds when $n=1, m=k$. Then when $n=1, m=k+1$ the statement holds based on (2) with $a_1=e$, and based on the expression of $*'$ as a shifted shuffle product and the recursive definition of shuffle product. Since shuffle product is commutative, the statement holds when $n=1$ or $m=1$.

Suppose the equality holds when $n \leq i$ or $m \leq i$ for some $i \geq 1$. Further suppose that the statement holds when $n = i+1, m \leq j$ for some $j \geq i$ (it holds when $j = i$ since $m \leq i$). Then when $n = i+1, m = j+1$, the statement holds based on (2) and the inductive hypothesis ($n = i, m = j+1$ and $n = i+1, m = j$), and based on the expression of $*'$ as a shifted shuffle product and the recursive definition of shuffle product. Hence the statement holds when $n = i+1, m \geq 1$. Since shuffle product is commutative, the statement holds when $n \leq i+1$ or $m \leq i+1$.

The operation $\succ\colon K\langle A\rangle \times K\langle A\rangle \to K\langle A\rangle$ induces an operation on $\mathrm{End}\,(K\langle A\rangle)$ given by, for f and g in $\mathrm{End}\,(K\langle A\rangle)$,

$$f \succ g := \succ \circ \left(f \otimes g\right) \circ \delta',$$

where δ' denotes the deconcatenation coproduct. Permutations induce elements in $\mathrm{End}\,(K\langle A\rangle)$, and $\succ$ is closed on permutations [25]:

Notation 6 Let $\succ\colon \left(\mathbb{Z}S \times \mathbb{Z}S\right) \setminus \left(\mathbb{Z}S_0 \times \mathbb{Z}S_0\right) \to \mathbb{Z}S$ denote the $\mathbb{Z}$-bilinear map given by

$$\sigma \succ \rho := \sigma \succ \bar{\rho}$$

for $\sigma \in S_n$, $\rho \in S_m$, where permutations are viewed as words with $\bar{\rho}\,(i) := n + \rho\,(i)$.

Malvenuto–Reutenauer algebra is a dendriform algebra (with $*' = \prec + \succ$) [14]. With $(1) \in S_1$, define

$$\begin{aligned} \mathscr{I}\colon \mathbb{Z}S &\to \mathbb{Z}S \\ \sigma &\mapsto \sigma \succ (1)\,. \end{aligned}$$

$\mathscr{I}$ can be viewed as an abstract integration. Since $\triangle'\,(1) = (1) \otimes \lambda + \lambda \otimes (1)$ for $\lambda \in S_0$, based on [25, p.248, Definition 1.2 (b)], for $\sigma \in \mathbb{Z}S$,

$$\triangle'\mathscr{I}(\sigma)\,\mathscr{I}(\sigma) \otimes \lambda + \sum_{(\sigma)} \sigma_{(1)} \otimes \mathscr{I}\left(\sigma_{(2)}\right).$$

Proposition 2 below helps to prove that slowly-varying one-forms are closed under iterated integration (Proposition 5).

Proposition 2 *For $s \in G\,(A)$, define $\widehat{s}\colon \mathbb{Z}S \to K\langle\langle A\rangle\rangle$ as in Proposition 1. Then for $n_1 \geq 0, n_2 \geq 1$,*

$$\begin{aligned} &\widehat{conc \circ (s \otimes t)}\left(1_{n_1} \succ 1_{n_2}\right) - \widehat{s}\left(1_{n_1} \succ 1_{n_2}\right) \qquad (3) \\ = &\sum_{\substack{k_1=0,\ldots,n_1 \\ k_2=0,\ldots,n_2-1}} \rho_{k_1,k_2} \cdot \left(conc \circ \left(\widehat{s}\left(1_{k_1} *' 1_{k_2}\right) \otimes \widehat{t}\left(1_{n_1-k_1} \succ 1_{n_2-k_2}\right)\right)\right) \end{aligned}$$

for $s,t \in G(A)$*, where* $\sigma \cdot (a_1 \cdots a_n) := a_{\sigma(1)} \cdots a_{\sigma(n)}$ *for* $\sigma \in S_n$ *and* $a_i \in A$*;* $\rho_{k_1,k_2} \in S_{n_1+n_2}$ *is given by changing the order of two sub-sequences* $(k_1+1,\ldots,k_1+k_2)$ *and* $(k_1+k_2+1,\ldots,n_1+k_2)$ *in* $(1,\ldots,n_1+n_2)$.

Remark 1 The equality (3) can be expressed as

$$\left(\int x^{n_1} dx^{n_2}\right)\Big|_{x=s}^{st} = \left(\int (sx)^{n_1}\, d\,(sx)^{n_2}\right)\Big|_{x=1}^{t}, \tag{4}$$

where $x^n := \widehat{x}(1_n)$, and is a change of variable formula for iterated integration. When s and t are signature of continuous bounded variation paths, the iterated integrals in (4) can be defined classically.

Proof Recall [25, Definition 1.2]:

$$\delta'(u \succ v) = (u \succ v) \otimes e + \sum_{|v_{(2)}| \geq 1} \left(u_{(1)} \sqcup\!\sqcup v_{(1)}\right) \otimes \left(u_{(2)} \succ v_{(2)}\right) \tag{5}$$

for $u, v \in A^*$, $|v| \geq 1$, where $\delta'(u) = \sum_{(u)} u_{(1)} \otimes u_{(2)}$ and $\delta'(v) = \sum_{(v)} v_{(1)} \otimes v_{(2)}$.

Since $1_{n_1} \succ 1_{n_2} := \succ \circ \left(1_{n_1} \otimes 1_{n_2}\right) \circ \delta'$, for $w \in A^*$, $w = uv$, $|u| = n_1$, $|v| = n_2$,

$$\begin{aligned}
&\left(conc \circ (s \otimes t)\,, \left(1_{n_1} \succ 1_{n_2}\right) \cdot w\right) \\
&= \left(conc \circ (s \otimes t)\,, u \succ v\right) \\
&= \left(s \otimes t, \delta'(u \succ v)\right) \text{ (duality between } conc \text{ and } \delta') \\
&= (s, u \succ v) + \sum_{|v_{(2)}| \geq 1} \left(s, u_{(1)} \sqcup\!\sqcup v_{(1)}\right)\left(t, u_{(2)} \succ v_{(2)}\right) \text{ (based on (5)) } \\
&= \left(s, \left(1_{n_1} \succ 1_{n_2}\right) \cdot w\right) \\
&\quad + \sum_{|v_{(2)}| \geq 1} \left(s, \left(1_{|u_{(1)}|} *' 1_{|v_{(1)}|}\right) \cdot \left(u_{(1)} v_{(1)}\right)\right)\left(t, \left(1_{|u_{(2)}|} \succ 1_{|v_{(2)}|}\right) \cdot \left(u_{(2)} v_{(2)}\right)\right)
\end{aligned}$$

for $s, t \in G(A)$.

The proof of Proposition 2 is based on a consistent relation between dendriform operations and the coproduct as at (5). A dendriform Hopf algebra [25, Definition 1.2] is a dendriform algebra that has a consistent coproduct, i.e. dendriform operations and the coproduct satisfy conditions of the form (5). We view the consistent relations in the form of (5) as an abstract change of variable formula for iterated integration.

Although expressed in terms of shuffles/permutations, core arguments in this section (in particular Proposition 2) can be applied to a general dendriform Hopf algebra.

4 Integration of Geometric Rough Paths

We first consider an example that is simple and important.

For two Banach spaces V and U, let $L(V, U)$ denote the set of continuous linear mappings from V to U. Consider a (degree-n) polynomial one-form $p : V \to L(V, U)$, which is a polynomial taking values in $L(V, U)$. For $v, w, v_0 \in V$,

$$p(v)(w) = \sum_{k=0}^{n} \left(D^k p\right)(v_0) \frac{(v - v_0)^{\otimes k}}{k!}(w),$$

where $p(v) \in L(V, U)$ and $p(v)(w) \in U$. The value of $p(v)(w)$ does not depend on v_0.

We lift the polynomial one-form p to an exact one-form $G(V) \to G(U)$, where $G(U)$ denotes the set of group-like elements in $T((U))$. This idea comes from Lyons and the author [17]. In [17, Theorem 5, Theorem 6], polynomial one-forms are lifted to closed one-forms from a group to a vector space, and the first level rough integral is reduced to an inhomogeneous Young integral. Here we consider exact one-forms between two Lie groups, and interpret the full integration of geometric rough paths. The interpretation is in the language of MR dendriform algebra.

Let $x \in BV([S, T], V)$ such that $x_S = 0$. Then

$$\begin{aligned}
&\int_{r=S}^{T} p(x_r)\, dx_r \\
&= \sum_{k=0}^{n} \left(D^k p\right)(0) \int_{r=S}^{T} \frac{(x_r)^{\otimes k}}{k!} \otimes dx_r \\
&= \sum_{k=0}^{n} \left(D^k p\right)(0) \int_{S<u_1<\cdots<u_{k+1}<T} dx_{u_1} \otimes \cdots \otimes dx_{u_{k+1}} \\
&=: \sum_{k=0}^{n} \left(D^k p\right)(0) X_{S,T}^{k+1}
\end{aligned}$$

where the first equality is the Taylor expansion of p; the second equality is based on the integration by parts formula and based on the symmetry of $D^k p$. Then $\int p(x)\, dx$ is expressed as a finite linear combination of iterated integrals of x.

Notation 7 For a polynomial one-form $p : V \to L(V, U)$ of degree n, define $f_p : G(V) \to U$ as

$$f_p(g) := \sum_{k=0}^{n} \left(D^k p\right)(0) g^{k+1}$$

where $g = \sum_{k \geq 0} g^k$ with $g^k \in V^{\otimes k}$.

f_p is defined on $T((V))$ not only on group-like elements. For $x \in BV([S,T],V)$, denote $X_t := \exp(x_S)\, S\left(x|_{[S,t]}\right)$, where $\exp(x_S)$ denotes the algebraic exponential of x_S in $T((V))$, $x|_{[S,t]}$ denotes the restriction of x to $[S,t]$, and $S\left(x|_{[S,t]}\right)$ denotes the signature of $x|_{[S,t]}$. Then

$$\int_{r=S}^{T} p(x_r)\, dx_r = f_p(X_T) - f_p(X_S). \tag{6}$$

When $x_S = 0$, $f_p(X_S) = 0$ and the equality holds based on the calculation above. When $x_S \neq 0$, let η be the straight line path that joins 0 to x_S, represented as $\eta(t) = t x_S, t \in [0,1]$. By using $S(\eta) = \exp(x_S)$ [4] and the additive property of integrals, the equality (6) still holds. The left hand side of (6) does not depend on the continuous bounded variation path η that joins 0 to x_S.

Then based on fundamental theorem of calculus and change of variable formula,

$$f_p(X_T) - f_p(X_S) = \int_{X_S}^{X_T} df_p = \int_{r=S}^{T} df_p dX_r.$$

As a result, the polynomial one-form $p : V \to L(V,U)$ is lifted to an exact one-form df_p for $f_p : G(V) \to U$ such that

$$\int_{r=S}^{T} p(x_r)\, dx_r = \int_{r=S}^{T} df_p dX_r$$

for each $x \in BV([S,T],V)$.

The lifting of a path to a rough path is necessary. A rough path can be viewed as a basis of controlled systems, and integration/differential equation can be viewed as a transformation between bases of controlled systems. The metric on rough path space can be p-variation for $p < \infty$, which requires much less than the metric needed to define iterated integrals that is $p < 2$. When $p \geq 2$, basis systems of order $1, \ldots, [p]$ are selected to postulate iterated integrals and satisfy an abstract 'integration by parts formula' (defines a character of shuffle algebra for each fixed time). The algebraic structure is important to interpret the limit behavior of controlled systems. For example, physical Brownian motion in a magnetic field can be described by Brownian motion with a 'non-canonical' Lévy area [10].

Consider the lift of the classical integral $\int p(x)\, dx$ to $f_p : G(V) \to U$. The function f_p takes values in Banach space U same as the integral $\int p(x)\, dx$. The full rough integral is a mapping between group-valued paths, and $f_p : G(V) \to U$ can be lifted to a function $F_p : G(V) \to G(U)$ such that

$$S\left(\int_{r=S}^{\cdot} p(x_r)\, dx_r \Big|_{[S,T]}\right) = F_p(X_S)^{-1} F_p(X_T)$$

for each $x \in BV([S,T],V)$, where $\int_{r=S}^{\cdot} p(x_r)\,dx_r$ denotes the integral path $t \mapsto \int_{r=S}^{t} p(x_r)\,dx_r$ for $t \in [S,T]$ and $X_t := \exp(x_S)\, S\left(x|_{[S,t]}\right)$. The lift of f_p to F_p is closely related to Definition 3.2.2 Lyons [16, p.285] recalled in Definition 5 below. The lift can also be interpreted in terms of iterated integrals of controlled paths introduced by Gubinelli [11, p.101, Theorem 1]. We view f_p and F_p as functions that only depend on p, with a consistent parallel translation (see (13) below).

We construct F_p from f_p based on MR dendriform algebra of permutations (with $*' = \prec + \succ$). For $l = 1, 2, \ldots$, denote

$$\sigma_l := \sum_{\substack{k_i=0,\ldots,n \\ i=1,\ldots,l}} \left(D^{k_1} p\right)(0) \otimes \cdots \otimes \left(D^{k_l} p\right)(0)\, m_{\succ}\left(1_{k_1+1}, \ldots, 1_{k_l+1}\right).$$

For simplicity, we assume that V has a (possibly infinite) basis given by a set A, and let $G(V)$ be the set of group-like elements in $K\langle\langle A\rangle\rangle$.

Notation 8 For a polynomial one-form $p : V \to L(V,U)$ of degree n, define $F_p : G(V) \to T\left((U)\right)$ as

$$F_p(s) := 1 + \sum_{l=1}^{\infty} F_p(s)^l,\ s \in G(V), \tag{7}$$

where $F_p(s)^l \in U^{\otimes l}$ is given by

$$F_p(s)^l := \sum_{\substack{k_i=0,\ldots,n \\ i=1,\ldots,l}} \left(D^{k_1} p\right)(0) \otimes \cdots \otimes \left(D^{k_l} p\right)(0)\, \widehat{s}\left(m_{\succ}\left(1_{k_1+1}, \ldots, 1_{k_l+1}\right)\right)$$

with $\widehat{s}$ defined at *(1)*.

The following proposition helps to prove that the indefinite integral of a polynomial one-form along a geometric rough path is again a geometric rough path.

Proposition 3 *$F_p : G(V) \to T\left((U)\right)$ is a lift of $f_p : G(V) \to U$, and F_p takes values in $G(U)$ the group-like elements in $T\left((U)\right)$.*

Proof F_p is a lift of f_p because $f_p(s) = F_p(s)^1$ for $s \in G(V)$.

Let $\succ$ resp. $\succ'$ denote the dendriform operation of MR resp. shuffle algebra. For $\rho \in S_j$ and integers $n_1 \geq 1, \ldots, n_j \geq 1$, denote $\overline{1_{n_1}}(i) := i$, $\overline{1_{n_{l+1}}}(i) := \sum_{r=1}^{l} n_r + i$ for $l = 0, \ldots, j-1$, and denote

$$\rho \cdot m_{\succ}\left(1_{n_1}, \ldots, 1_{n_j}\right) := m_{\succ'}\left(\overline{1_{n_{\rho(1)}}}, \ldots, \overline{1_{n_{\rho(j)}}}\right).$$

For integers $n_1 \geq 1, \ldots, n_{k+j} \geq 1$, denote $\overline{1_{n_1}}(i) := i$, $\overline{1_{n_{l+1}}}(i) := \sum_{r=1}^{l} n_r + i$ for $l = 0, \ldots, k+j-1$. Based on Lemma 1 and Proposition 1, for $s \in G(V)$, $\widehat{s}$ defined at (1) satisfies

$$
\begin{aligned}
&conc \circ \left(\widehat{s}\left(m_{\succ}\left(1_{n_1}, \ldots, 1_{n_k}\right)\right) \otimes \widehat{s}\left(m_{\succ}\left(1_{n_{k+1}}, \ldots, 1_{n_{k+j}}\right)\right)\right) \\
&= \widehat{s}\left(m_{\succ}\left(1_{n_1}, \ldots, 1_{n_k}\right) *' m_{\succ}\left(1_{n_{k+1}}, \ldots, 1_{n_{k+j}}\right)\right) \quad \text{(Lemma 1)} \\
&= \widehat{s}\left(m_{\succ'}\left(\overline{1_{n_1}}, \ldots, \overline{1_{n_k}}\right) \sqcup\!\sqcup m_{\succ'}\left(\overline{1_{n_{k+1}}}, \ldots, \overline{1_{n_{k+j}}}\right)\right) \ (*' \text{ as shifted shuffle}) \\
&= \widehat{s}\left(\sum_{\rho \in 1_k *' 1_j} m_{\succ'}\left(\overline{1_{n_{\rho(1)}}}, \ldots, \overline{1_{n_{\rho(k+j)}}}\right)\right) \quad \text{(Proposition 1)} \\
&=: \widehat{s}\left(\left(1_k *' 1_j\right) \cdot m_{\succ}\left(1_{n_1}, \ldots, 1_{n_{k+j}}\right)\right).
\end{aligned}
$$

The projection of $F_p(s)$ to $U^{\otimes l}$ is $F_p(s)^l$. For $\rho \in S_l$, based on the equality $\sum_w (s, w)\, \rho^{-1} \cdot w = \sum_w \left(s, \rho \cdot w\right) w$, we have

$$
\rho^{-1} \cdot \left(F_p(s)^l\right) = \sum_{\substack{k_i = 0, \ldots, n \\ i = 1, \ldots, l}} \left(D^{k_1} p\right)(0) \otimes \cdots \otimes \left(D^{k_l} p\right)(0)\, \widehat{s}\left(\rho \cdot m_{\succ}\left(1_{k_1+1}, \ldots, 1_{k_l+1}\right)\right)
$$

Then based on the calculation above,

$$
F_p(s)^k \otimes F_p(s)^j = \sum_{1_k *' 1_j} \rho^{-1} \cdot \left(F_p(s)^{k+j}\right), \tag{8}
$$

where $\otimes$ is the tensor product in $T\left((U)\right)$. That the equality (8) holds for all $k \geq 0$, $j \geq 0$ is equivalent to $F_p(s) \in G(U)$.

For $x \in BV([S, T], V)$, denote $X_t := \exp(x_S)\, S\left(x|_{[S,t]}\right)$. Define $y \in BV([S, T], U)$ as $y_t := \int_{r=S}^{t} p(x_r)\, dx_r$, and denote $Y_t := S\left(y|_{[S,t]}\right)$. The exact one-form dF_p is a lift of the polynomial one-form p:

$$
Y_S^{-1} Y_T = F_p(X_S)^{-1} F_p(X_T) = \int_{r=S}^{T} dF_p dX_r.
$$

When $x_S = 0$, $F_p(X_S) = 1$ and the equality holds. When $x_S \neq 0$, let η denote the straight line path that connects 0 to x_S. Based on Chen's identity, the equality $Y_S^{-1} Y_T = F_p(X_S)^{-1} F_p(X_T)$ still holds, and $Y_S^{-1} Y_T$ does not depend on the choice of continuous bounded variation path η that connects 0 to x_S.

Generally, for a continuous path $X : [S, T] \to G(V)$, the integral of p along X can be defined as

$$\int_{r=S}^{T} p(x_r)\, dX_r := F_p(X_S)^{-1} F_p(X_T),$$

where x denotes the projection of X to a path in V. In particular, when X is a geometric rough path, the integral can be defined this way. This integral is related to the definition of almost multiplicative functional defined by Lyons [16, Definition 3.2.2].

The integration of polynomial one-forms provides basic ingredients for the integration of general regular one-forms. Classically the smoothness of a function is expressed in terms of polynomials. Based on Stein [27, Chapter VI], a function θ on a closed subset $F \subseteq \mathbb{R}^n$ is $\mathrm{Lip}(\gamma)$ for some $\gamma \in (k, k+1]$ if there exists a family of functions θ^j, $j = 1, \ldots, k+1$, with $\theta^1 = \theta$, such that if

$$\theta^j(x) = \sum_{|j+l| \le k+1} \frac{\theta^{j+l}(y)}{l!} (x-y)^l + R_j(x, y)$$

then $|\theta^j(x)| \le M$ and $\left|R_j(x, y)\right| \le M |x-y|^{\gamma+1-|j|}$ for all $x, y \in F$, $|j| \le k+1$. Based on Lyons [16] Lipschitz one-forms are Lipschitz functions in the sense of Stein, taking values in continuous linear mappings.

The following is Definition 3.2.2 [16, p.285].

Definition 5 (Lyons) For any multiplicative functional $X_{s,t}$ in $\Omega G(V)^p$ define

$$Y^i_{s,t} = \sum_{l_1, \ldots, l_i = 1}^{[p]} \theta^{l_1}(x_s) \otimes \cdots \otimes \theta^{l_i}(x_s) \sum_{\pi \in \Pi_{\underline{l}}} \pi\left(X^{\|\underline{l}\|}_{s,t}\right)$$

$\Omega G(V)^p$ denotes the set of multiplicative functionals of finite p-variation taking values in the step-$[p]$ truncation of $G(V)$ [16, p.258, Definition 2.3.1]. Based on [16, p.284], $\underline{l} = (l_1, \ldots, l_i)$ and $\|\underline{l}\| = \sum_{j=1}^{i} l_j$. Denote $K_{j+1} := K_j + l_{j+1}$ with $K_0 := 0$. $\Pi_{\underline{l}}$ is the set of permutation π of order $\|\underline{l}\|$ that satisfy $\pi^{-1}(K_j + 1) < \cdots < \pi^{-1}(K_j + l_{j+1})$ and $\pi^{-1}(K_j) < \pi^{-1}(K_{j+1})$ for $j = 0, \ldots, i-1$. Each $\pi \in S_n$ acts linearly on $V^{\otimes n}$ given by $\pi(v_1 \otimes \cdots \otimes v_n) := v_{\pi(1)} \otimes \cdots \otimes v_{\pi(n)}$ for $v_i \in V$, $i = 1, 2, \ldots, n$. $X^{\|\underline{l}\|}_{s,t}$ denotes the $\|\underline{l}\|$th component of $X_{s,t}$, and is an element in $V^{\otimes \|\underline{l}\|}$.

The following is Theorem 3.2.1 [16].

Theorem 9 (Lyons, Existence of Integral) *For any multiplicative functional $X_{s,t}$ in $\Omega G(V)^p$ and any one-form $\theta \in \mathrm{Lip}(\gamma-1, \{X_u, u \in [s,t]\})$ with $\gamma > p$ the sequence $Y_{s,t} = (1, Y^1_{s,t}, \ldots, Y^{[p]}_{s,t})$ defined above is almost multiplicative and of finite p-variation; if $X_{s,t}$ is controlled by ω on J where ω is bounded by L, and the $\mathrm{Lip}(\gamma-1)$ norm of θ is bounded by M, then the almost multiplicative and p-variation properties of Y are controlled by multiples of ω which depend only on γ, p, L, M.*

The multiplicative functional associated with Y obtained in Theorem 9 is defined to be the integral of the one-form θ along geometric rough path X [16, p.274, Theorem 3.3.1, and p.288, Definition 3.2.3]. Denote the integral as $\int \theta(x)\, dX$. Based on [16, p.274, Theorem 3.3.1],

$$\int_{r=0}^{1} \theta(x_r)\, dX_r := \lim_{|D|\to 0, D\subset[0,1]} Y_{t_0,t_1} \cdots Y_{t_{n-1},t_n}. \tag{9}$$

Based on the lift of polynomial one-form p to exact one-form dF_p, the integral $\int \theta(x)\, dX$ can be interpreted in terms of time-varying exact one-forms.

Let $X : [0,1] \to G(V)$ be a geometric p-rough path, and let $\theta : V \to L(V,U)$ be a $\mathrm{Lip}(\gamma)$ one-form for $\gamma > p-1$. Let x denote the projection of X to a path in V. For $x_s \in V$, define polynomial one-form $p_{x_s} : V \to L(V,U)$ as

$$p_{x_s}(v)(w) = \sum_{k=0}^{[p]-1} \theta^k(x_s) \frac{(v-x_s)^{\otimes k}}{k!}(w) \tag{10}$$

for $v, w \in V$. For the polynomial one-form p_{x_s}, define $F_{p_{x_s}}$ as at (7). The almost multiplicative functional $Y_{s,t}$ in Definition 5 can be expressed as

$$Y_{s,t} = \prod\nolimits_{\le[p]} \left(F_{p_{x_s}}(X_s)^{-1} F_{p_{x_s}}(X_t) \right) \tag{11}$$

for every $s \le t$, where $\prod_{\le[p]}$ denotes the step-$[p]$ truncation of elements in $T((U))$. The equality (11) follows from a generalized Chen's identity about the multiplicativity of rough path liftings of controlled paths/effects (see Sect. 6 for details). The generalized Chen's identity can be proved based on the uniqueness of the continuous lifting (Proposition 5).

Theorem 10 *For a geometric p-rough path $X : [0,1] \to G(V)$ and a $\mathrm{Lip}(\gamma-1)$ one-form $\theta : V \to L(V,U)$ for $\gamma > p$, let $\int_{r=0}^{1} \theta(x_r)\, dX_r$ denote the integral*

defined by Lyons in [16]. *Then with* p_{x_s} *defined at* (10) *and* $F_{p_{x_s}}$ *defined at* (7),

$$\begin{aligned}
&\int_{r=0}^{1} \theta\left(x_r\right) dX_r \\
&= \lim_{|D|\to 0, D\subset[0,1]} \int_{r=t_0}^{t_1} dF_{p_{x_{t_0}}} dX_r \cdots \int_{r=t_{n-1}}^{t_n} dF_{p_{x_{t_{n-1}}}} dX_r \qquad (12) \\
&=: \int_{r=0}^{1} dF_{p_{x_r}} dX_r
\end{aligned}$$

where $D = \{t_k\}_{k=0}^{n}, 0 = t_0 < \cdots < t_n = 1, n \geq 1$ *with* $|D| := \max_k \left|t_{k+1} - t_k\right|$.

The equality is based on (9) and (11). By applying the neo-classical inequality [16, Lemma 2.2.2] to the multiplicative functional $r \mapsto F_{p_{x_s}}\left(X_r\right), r \in [s,t]$, it can be proved that the difference between $F_{p_{x_s}}\left(X_s\right)^{-1} F_{p_{x_s}}\left(X_t\right)$ and $Y_{s,t}$ is bounded by a term of the form $\omega\left(s,t\right)^{\kappa}$ for a control ω and $\kappa > 1$. Based on the Lipscthiz condition on θ, ω and κ can be chosen to be independent of the interval $[s,t]$. Then the existence of the limit (12) follows from the existence of the integral $\int \theta\left(x\right) dX$ in (9), which is obtained in Theorem 3.2.1 [16] i.e. Theorem 9.

There is a minor difference between geometric rough paths $\Omega G\left(V\right)^{p}$ in [16] and that defined in Definition 3 (paths in Definition 3 are also called weakly geometric rough paths). The integration $\int_{r=0}^{1} dF_{p_{x_r}} dX_r$ in Theorem 10 can be applied to both classes of rough paths.

The integration of time-varying exact one-forms exists in a general setting. Consider two Lie groups G_1 and G_2, and a path $X : [S,T] \to G_1$. Suppose $f_t : G_1 \to G_2$ is a family of functions indexed by $t \in [S,T]$. If the limit exists in G_2:

$$\lim_{|D|\to 0, D=\{t_k\}_{k=0}^{n}\subset[S,T]} \int_{r=t_0}^{t_1} df_{t_0} dX_r \int_{r=t_1}^{t_2} df_{t_1} dX_r \cdots \int_{r=t_{n-1}}^{t_n} df_{t_{n-1}} dX_r$$

where $\int_{r=s}^{t} df_s dX_r := f_s\left(X_s\right)^{-1} f_s\left(X_t\right)$, then the integral

$$\int_{r=S}^{T} df_r dX_r$$

is defined to be the limit. A sufficient condition for the existence of the integral is given in Theorem 12 below based on Feyel, de la Pradelle and Mokobodzki [9]. The integral can be viewed as a non-abelian analogue of Young's integral [30].

The integration of time-varying exact one-forms can be explained by the parallel translation on a principal bundle of functions between two Lie groups. Consider a principal bundle P on G_1 that associates each $a \in G_1$ with the set of functions

$P_a := \left\{ f | f : G_1 \to G_2, f\left(1_{G_1}\right) = 1_{G_2} \right\}$. For $f \in P_a$ and $b \in G_1$, define

$$\begin{aligned} f_b &\in P_{ab} \\ g \in G_1 &\mapsto f(b)^{-1} f\left(bg\right) \in G_2. \end{aligned} \tag{13}$$

The $\left\{ f_b | b \in G_1 \right\}$ defined are consistent:

$$\left(f_b\right)_c = f_{bc}$$

for $b, c \in G_1$. The parallel translation is consistent with the integration of exact one-forms: $\int_{r=S}^{T} df dX_r = f(X_S)^{-1} f(X_T) = f_{X_S}\left(X_S^{-1} X_T\right)$ for $X : [S, T] \to G_1$. In the integration of time-varying exact one-forms $\int_{r=S}^{T} df_r dX_r$, $\{f_t\}_{t\in[S,T]}$ are compared after parallel translation.

For a fixed continuous path $X : [S, T] \to G_1$, consider a condition on exact one-forms $\left(df_t\right)_t$ for $f_t : G_1 \to G_2$ that guarantees the existence of the integral $\int_{r=S}^{T} df_r dX_r$. Theorem 12 below gives a condition that roughly states that, if one-step discrete approximations are comparable to two-steps discrete approximations up to a small error, then the integral exists as the limit of Riemann products. When X is a geometric p-rough path, the condition on $\left(df_t\right)_t$ can be further specified, and the condition can be viewed as an inhomogeneous analogue of Young's condition [30, p.264]. Such a condition is closely related to the notion of weakly controlled paths introduced by Gubinelli [11]. The following is Definition 1 [11].

Definition 6 (Gubinelli, weakly controlled paths) Fix an interval $I \subseteq \mathbb{R}$ and let $X \in \mathscr{C}^{\gamma}(I, V)$. A path $Z \in \mathscr{C}^{\gamma}(I, V)$ is said to be weakly controlled by X in I with a reminder of order η if there exists a path $Z' \in \mathscr{C}^{\eta-\gamma}\left(I, V \otimes V^*\right)$ and a process $R_Z \in \Omega C^{\eta}(I, V)$ with $\eta > \gamma$ such that

$$\delta Z^{\mu} = Z'^{\mu\nu} \delta X^{\nu} + R_Z^{\mu}.$$

If this is the case we will write $\left(Z, Z'\right) \in \mathscr{D}_X^{\gamma,\eta}(I, V)$ and we will consider on the linear space $\mathscr{D}_X^{\gamma,\eta}(I, V)$ the semi-norm

$$\|Z\|_{\mathscr{D}(X,\gamma,\eta),I} := \left\|Z'\right\|_{\infty,I} + \left\|Z'\right\|_{\eta-\gamma,I} + \|R_Z\|_{\eta,I} + \|Z\|_{\gamma,I}.$$

$\mathscr{C}^{\gamma}(I, V)$ denotes the set of γ-Hölder paths $I \to V$ for a Banach space V. $\Omega C^{\eta}(I, V)$ denotes the set of maps $R : \left\{(s, t) | s \in I, t \in I\right\} \to V$ that satisfy

$$\|R\|_{\eta,I} := \sup_{s\in I, t\in I} \frac{\left\|R_{s,t}\right\|}{|t - s|^{\eta}} < \infty.$$

$\|R\|_{\infty,I}$ denotes the supremum norm of R on I. For $Z \in \mathscr{C}^{\gamma}(I,V)$, $\delta Z \in \Omega C^{\gamma}(I,V)$ is given by $(\delta Z)_{s,t} := Z_t - Z_s$, and denote $\|Z\|_{\gamma,I} := \|\delta Z\|_{\gamma,I}$. μ and ν are indices of vectors in V. V^* denotes the linear dual of V. For $Z' \in \mathscr{C}^{\eta-\gamma}\left(I, V \otimes V^*\right)$ and $X \in C^{\gamma}(I,V)$, $Z'\delta X \in \Omega C^{\eta}(I,V)$ is given by $\left(Z'\delta X\right)_{s,t} := Z'_s (\delta X)_{s,t}$.

For a fixed geometric rough path X, we consider a class of integrable one-forms whose indefinite integrals are called effects (see Sect. 6 for more details). The set of effects of X is a subset of the paths controlled by X. The relationship between controlled paths and effects is comparable to that between the integrand and the integral. In the integration of one-form $\int \alpha(x)\, dX$, $t \mapsto \alpha(x_t)$ is a controlled path; $t \mapsto \int_{r=0}^{t} \alpha(x_r)\, dX_r$ is an effect so a controlled path. A controlled path can also be interpreted as a time-varying exact one-form, and its varying speed can be a little quicker than that of an effect. One benefit of working with effects is that basic operations (multiplication, composition with regular functions, integration, iterated integration) are continuous operations in the space of one-forms in operator norm. In particular, the lifting of an effect to a geometric rough path is continuous. In [17] effects are employed to give a short proof of the unique solvability and stability of the solution to differential equations driven by rough paths, and the differences between adjacent Picard iterations decay factorially in operator norm.

5 Integration of Time-Varying Exact One-Forms

For continuous $x : [0,1] \to \mathbb{C}$ of finite p-variation and continuous $y : [0,1] \to \mathbb{C}$ of finite q-variation $p^{-1} + q^{-1} > 1, p \geq 1, q \geq 1$, Young [30] defined the Stieltjes integral $\int_{r=0}^{1} x_r dy_r$ as the limit of Riemann sums. Lyons [16] defined the integration of one-forms along geometric rough paths by constructing a multiplicative functional from an almost multiplicative functional.

Based on Feyel, La Pradelle and Mokobodzki [9], let M be a monoid with a unit element I, and M is complete under a distance d that satisfies

$$d\left(xz, yz\right) \leq |z|\, d\left(x, y\right),\ d\left(zx, zy\right) \leq |z|\, d\left(x, y\right) \tag{14}$$

for $x, y, z \in M$, where $z \mapsto |z|$ is a Lipschitz function on M with $|I| = 1$.

Let L denote the Lipschitz constant of $z \mapsto |z|$.

Suppose $V : [0,T) \to \mathbb{R}$ is a *strong control function,* i.e. $V(0) = 0$, non-decreasing, and there exists a $\theta > 2$ such that for every t

$$\overline{V}_{\theta}(t) := \sum_{n \geq 0} \theta^n V\left(t 2^{-n}\right) < \infty.$$

For example, $V(t) = t^{\alpha}$ when $\alpha > 1$ is a strong control function.

Suppose $\mu : \{(s,t) \,|0 \le s \le t < T\} \to (M,d)$ is continuous, $\mu(t,t) = I$ for every t, and

$$d\left(\mu(s,t), \mu(s,u)\,\mu(u,t)\right) \le V(t-s) \tag{15}$$

for every $s \le u \le t$. A map $\nu : \{(s,t) \,|0 \le s \le t < T\} \to (M,d)$ is called *multiplicative* if $\nu(s,t) = \nu(s,u)\,\nu(u,t)$ for every $s \le u \le t$.

Theorem 11 (Feyel, La Pradelle, Mokobodzki) *There exists a unique continuous multiplicative function ν such that $d\left(\mu(s,t), \nu(s,t)\right) \le C_{\theta,L}\overline{V}_\theta(t-s)$ for every $s \le t$.*

Let G_1 and G_2 be two groups, and suppose G_2 is complete under a distance d that satisfies (14). Let $X : [0,T] \to G_1$. Suppose $f_t : G_1 \to G_2$ is a family of functions indexed by $t \in [0,T]$. Define $\mu : \{(s,t) \,|0 \le s \le t < T\} \to (G_2,d)$ as

$$\mu(s,t) := f_s(X_s)^{-1} f_s(X_t) \in G_2 \tag{16}$$

for $s \le t$. Suppose (15) holds for this μ and a strong control function V.

Let $\nu : \{(s,t) \,|0 \le s \le t < T\} \to (G_2,d)$ denote the multiplicative function associated with μ at (16) obtained by Theorem 11.

Theorem 12 *Define $\int_{r=0}^{\cdot} df_r dX_r : [0,T] \to G_2$ as $\int_{r=0}^{t} df_r dX_r := \nu(0,t)$ for every t. Then $\int_{r=0}^{\cdot} df_r dX_r$ is the unique continuous path $y : [0,T] \to G_2$ such that $y_0 = 1_{G_2}$ and $d(y_s^{-1} y_t, \int_{r=s}^{t} df_s dX_r) \le C_{\theta,L}\overline{V}_\theta(t-s)$ for every $s \le t$.*

6 Effects of a Geometric Rough Path

6.1 Definition

The set of effects of a geometric rough path is a subset of the paths controlled by the geometric rough path. Similar to controlled paths, effects are stable under basic operations. Integrals of one-forms and solutions to differential equations are effects so are controlled paths.

For $k = 1, \ldots, [p]$, let $L\left(V^{\otimes k}, U\right)$ denote the set of continuous linear maps from $V^{\otimes k}$ to U.

Notation 13 Let E^U denote the vector bundle on $G(V)$ that associates each $a \in G(V)$ with the vector space:

$$E_a^U := \left\{ \phi \middle| \phi : G(V) \to U, \phi = \sum_{k=1}^{[p]} \phi^k, \phi^k \in L\left(V^{\otimes k}, U\right) \right\}$$

where $\phi^k(x) := \phi^k x^k$ for $x \in G(V)$, $x = \sum_{k \ge 0} x^k$, $x^k \in V^{\otimes k}$.

E_a^U can be considered as the space of 'polynomials' up to degree-$\lfloor p \rfloor$ that has no 'constant', and can be viewed as a polynomial approximation to the 'tangent space' of functions $\left\{ f_a | f : G(V) \to U \right\}$ with $f_a(x) := f(ax) - f(a)$, $x \in G(V)$.

The parallel translation on E^U is given by

Notation 14 For $p \in E_a^U$ and $b \in G(V)$, define $p_b \in E_{ab}^U$ as

$$p_b(x) := p(bx) - p(b)$$

for $x \in G(V)$.

Then $\left(p_b\right)_c = p_{bc}$ for $b, c \in G(V)$.

For $\phi \in E_a$, $\phi = \sum_{k=1}^{[p]} \phi^k$, denote

$$\|\phi\| := \max_{k=1,\ldots,[p]} \left\|\phi^k\right\| \text{ and } \|\phi\|_k := \left\|\phi^k\right\|$$

where $\left\|\phi^k\right\|$ denotes the norm of ϕ^k as a linear operator.

Definition 7 (Operator Norm) Let $X : [0,1] \to G(V)$ be a geometric p-rough path for some $p \geq 1$, and let U be a Banach space. Suppose

$$\beta \in \left(X, E_X^U\right) \text{ i.e. } \beta : X_t \mapsto \beta(X_t) \in E_{X_t}^U.$$

For $t \in [0,1]$ and $a \in G(V)$, define

$$\begin{aligned} \left(\beta(X_t)\right)_a &\in E_{X_t a}^U \\ x &\mapsto \beta(X_t)(ax) - \beta(X_t)(a), x \in G(V). \end{aligned}$$

For a control ω and $\theta > 1$, define the operator norm

$$\|\beta\|_\theta^\omega := \sup_{t \in [0,1]} \|\beta(X_t)\| + \max_{k=1,\ldots,[p]} \sup_{0 \leq s < t \leq 1} \frac{\left\|\beta(X_t) - \left(\beta(X_s)\right)_{X_s^{-1} X_t}\right\|_k}{\omega(s,t)^{\theta - \frac{k}{p}}}.$$

Definition 8 (Slowly-Varying One-Form) Let $X : [0,1] \to G(V)$ be a geometric p-rough path for some $p \geq 1$, and let U be a Banach space. Then $\beta \in \left(X, E_X^U\right)$ is called a slowly varying one-form, if there exists a control ω and $\theta > 1$ such that $\|\beta\|_\theta^\omega < \infty$.

For each $t \in [0,1]$, $\beta(X_t)$ can be viewed as a continuous linear mapping from monomials (components of rough paths) to the vector space U.

For a control ω and $\theta > 1$, the set of slowly varying one-forms along X with finite operator norm $\|\cdot\|_\theta^\omega$ forms a Banach space.

Suppose $\beta \in \left(X, E_X^U\right)$ is a slowly-varying one-form. Let $\beta\left(X_t\right)_{X_t^{-1}}$, $t \in [0,1]$ be considered as functions $G(V) \to U$ index by t. Define

$$\begin{aligned}
&\int_{r=0}^{1} \beta\left(X_r\right) dX_r \\
:&= \int_{r=0}^{1} d\left(\beta\left(X_r\right)_{X_r^{-1}}\right) dX_r \\
:&= \lim_{|D|\to 0, D\subset[0,1]} \sum_{k, t_k \in D} \left(\beta\left(X_{t_k}\right)_{X_{t_k}^{-1}}\left(X_{t_{k+1}}\right) - \beta\left(X_{t_k}\right)_{X_{t_k}^{-1}}\left(X_{t_k}\right)\right) \\
=&\lim_{|D|\to 0, D\subset[0,1]} \sum_{k, t_k \in D} \beta\left(X_{t_k}\right)\left(X_{t_k, t_{k+1}}\right)
\end{aligned}$$

where the last equality is based on $\beta\left(X_{t_k}\right)\left(1_{G(V)}\right) = 0$. The integral exists based on the slowly varying condition on β.

Definition 9 (Effects) Let $X : [0,1] \to G(V)$ be a geometric p-rough path for some $p \geq 1$, and let $\beta \in \left(X, E_X^U\right)$ be a slowly varying one-form. Then for $\xi \in U$, the integral path

$$t \mapsto \xi + \int_{r=0}^{t} \beta\left(X_r\right) dX_r, t \in [0,1]$$

is called an effect of X.

Theorem 15 *Suppose $\beta \in \left(X, E_X^U\right)$ is a slowly-varying one-form such that $\|\beta\|_\theta^\omega < \infty$ for a control ω and $\theta > 1$. Define $h : [0,1] \to U$ as*

$$h_t := \int_{r=0}^{t} \beta\left(X_r\right) dX_r, t \in [0,1].$$

Then with control $\hat{\omega} := \omega + \|X\|_{p-var}^p$,

$$\left\| h_t - h_s - \beta_s\left(X_s\right)\left(X_{s,t}\right)\right\| \leq C_{p,\theta,\hat{\omega}(0,T)} \|\beta\|_\theta^\omega \hat{\omega}(s,t)^\theta$$

for every $s \leq t$, and

$$\|h\|_{p-var,[0,T]} \leq C_{p,\theta,\hat{\omega}(0,T)} \|\beta\|_\theta^\omega.$$

The first estimate can be proved similarly to Young [30, p.254, result 5]; the second follows from the first.

6.2 *Stability of Effects Under Basic Operations*

Consider the set of effects of a geometric rough path. Effects are closed under basic operations (multiplication, composition with regular functions, integration, iterated integration); the proof is similar to that for controlled paths as Proposition 4 on p.100 and Theorem 1 on p.101 by Gubinelli [11]. For effects these operations are continuous in the space of one-forms in operator norm.

Fix a geometric p-rough path $X : \left[0, 1\right] \to G\left(V\right)$ for some $p \geq 1$.

6.2.1 Composition with Regular Functions

The stability of effects under composition with regular functions follows from the fact that polynomials are closed under composition.

For Banach spaces U and W, denote by $C^{\gamma}\left(U, W\right)$ the set of functions $\varphi : U \to W$ that are $\lfloor\gamma\rfloor$-times Fréchet differentiable ($\lfloor\gamma\rfloor := \max\left\{n | n \in \mathbb{N}, n < \gamma\right\}$) with the $\lfloor\gamma\rfloor$th derivative $\left(\gamma - \lfloor\gamma\rfloor\right)$-Hölder, uniformly on any bounded set. For $R > 0$, denote

$$\left\|\varphi\right\|_{\gamma, R} := \max_{k=0,1,\ldots,\lfloor\gamma\rfloor} \left\|\left(D^{k}\varphi\right)\right\|_{\infty, R} + \left\|\left(D^{\lfloor\gamma\rfloor}\varphi\right)\right\|_{\left(\gamma-\lfloor\gamma\rfloor\right)\text{-Höl}, R}$$

where $\left\|\cdot\right\|_{\infty, R}$ resp. $\left\|\cdot\right\|_{\left(\gamma-\lfloor\gamma\rfloor\right)\text{-Höl}, R}$ is the uniform resp. Hölder norm on $\left\{u \in U \,|\, \left\|u\right\| \leq R\right\}$.

For $\phi\left(X_t\right) \in E^{U}_{X_t}$ and $l = 1, \ldots, \left[p\right]$, define the 'truncated polynomial':

$$\prod\nolimits_{\leq[p]} \left(\phi\left(X_t\right)^{\otimes l}\right) \in E^{U^{\otimes l}}_{X_t}$$

$$x \mapsto \sum_{\substack{k_1+\cdots+k_l \leq [p] \\ k_i = 1,\ldots,[p]}} \left(\phi^{k_1}\left(X_t\right) \otimes \cdots \otimes \phi^{k_l}\left(X_t\right)\right) \left(\left(1_{k_1} *' \cdots *' 1_{k_l}\right) \cdot \left(x^{k_1+\cdots+k_l}\right)\right)$$

for $x \in G\left(V\right)$, where $\phi = \sum_{k=1}^{[p]} \phi^{k}$, $\phi^{k} \in L\left(V^{\otimes k}, U\right)$ and $x = \sum_{k\geq 0} x^{k}$, $x^{k} \in V^{\otimes k}$.

Proposition 4 *Suppose $\beta_1 \in \left(X, E^{U}_{X}\right)$ is a slowly-varying one-form and $\left\|\beta_1\right\|^{\theta_1}_{\omega_1} < \infty$ for a control ω_1 and $\theta_1 > 1$. Denote $h_t := \int_0^t \beta_1\left(X_t\right) dX_t, t \in \left[0, 1\right]$. For $\varphi \in C^{\gamma}\left(U, W\right)$, $\gamma > p$, define $\beta \in \left(X, E^{W}_{X}\right)$ as*

$$\beta(X_t)(x) := \sum_{l=1}^{[p]} \frac{1}{l!} \left(D^l \varphi\right)(h_t) \left(\prod\nolimits_{\leq [p]} \left(\beta_1 (X_t)^{\otimes l}\right)\right)(x)$$

for $x \in G(V)$. *Then with* $\omega := \omega_1 + \|X\|_{p-var}^p$ *and* $\theta := \min\left(\theta_1, \frac{\gamma}{p}, \frac{[p]+1}{p}\right)$,

$$\|\beta\|_\theta^\omega \leq C_{p,\theta,\omega(0,1)} \|\varphi\|_{\gamma,\|h\|_\infty} \max\left(\|\beta_1\|_{\theta_1}^{\omega_1}, \left(\|\beta_1\|_{\theta_1}^{\omega_1}\right)^{[p]}\right), \tag{17}$$

where $\|h\|_\infty := \sup_{t\in[0,1]} \|h_t\|$, *and*

$$\int_{r=0}^{t} \beta(X_r)\, dX_r = \varphi(h_t) - \varphi(h_0) \text{ for } t \in [0,1].$$

Remark 2 Effects are closed under pointwise tensor multiplication and form an algebra. For Banach spaces $U_i, i = 1, 2$, consider $\varphi : (U_1, U_2) \to U_1 \otimes U_2$ given by $(u_1, u_2) \mapsto u_1 \otimes u_2$. Then $D^3\varphi \equiv 0$.

Proof For $l = 1, \ldots, [p]$,

$$\left(\phi^{\otimes l}\right)_a (x) = \left(\phi(a) + \phi_a(x)\right)^{\otimes l} - \phi^{\otimes l}(a)$$

for $x \in G(V)$.

We rescale φ by $\|\varphi\|_{\gamma,\|h\|_\infty}^{-1}$ and assume $\|\varphi\|_{\gamma,\|h\|_\infty} = 1$. Denote $X_{s,t} := X_s^{-1} X_t$. For $s \leq t$,

$$\begin{aligned}
&\left(\beta(X_t) - \left(\beta(X_s)\right)_{X_{s,t}}\right)(x) \\
&= \sum_{l=1}^{[p]} \frac{1}{l!} \left(\left(D^l\varphi\right)(h_t) - \sum_{j=0}^{[p]-l} \frac{1}{j!} \left(D^{j+l}\varphi\right)(h_s)(h_t - h_s)^{\otimes l}\right) \\
&\quad \times \left(\prod\nolimits_{\leq [p]} \left(\beta_1(X_t)^{\otimes l}\right)\right)(x) \\
&\quad + \sum_{l=1}^{[p]} \sum_{j=0}^{[p]-l} \frac{1}{l!}\frac{1}{j!} \left(D^{l+j}\varphi\right)(h_s) \left((h_t - h_s)^{\otimes j} - \beta(X_s)\left(X_{s,t}\right)^{\otimes j}\right) \\
&\quad \times \left(\prod\nolimits_{\leq [p]} \left(\beta_1(X_t)^{\otimes l}\right)\right)(x)
\end{aligned}$$

$$+\sum_{l=1}^{[p]}\frac{1}{l!}\left(D^{l}\varphi\right)(h_s)\prod\nolimits_{\leq[p]}\left(\left(\beta_1(X_s)\left(X_{s,t}\right)+\beta_1(X_t)\right)^{\otimes l}\right.$$
$$\left.-\left(\beta_1(X_s)\left(X_{s,t}\right)+\left(\beta_1(X_s)\right)_{X_{s,t}}\right)^{\otimes l}\right)(x)$$

for $x \in G(V)$. Then the estimate (17) follows from Theorem 15. Based on comparison of local expansions,

$$\int_{r=0}^{t}\beta(X_r)\,dX_r=\varphi(h_t)-\varphi(h_0),\ t\in[0,1].$$

6.2.2 Iterated Integration

Let $U_i, i=1,2$ be two Banach spaces. For $\phi_i \in E_{X_t}^{U_i}, i=1,2$, define the 'truncated iterated integration':

$$\prod\nolimits_{\leq[p]}(\phi_1\succ\phi_2)\in E_{X_t}^{U_1\otimes U_2}$$
$$x\mapsto\sum_{\substack{k_1+k_2\leq[p]\\ k_i=1,\ldots,[p]}}\left(\phi_1^{k_1}\otimes\phi_2^{k_2}\right)\left(\left(1_{k_1}\succ 1_{k_2}\right)\cdot\left(x^{k_1+k_2}\right)\right)$$

for $x \in G(V)$, where $\phi_i=\sum_{k=1}^{[p]}\phi_i^k$, $\phi_i^k \in L\left(V^{\otimes k},U_i\right)$ and $x=\sum_{k\geq 0}x^k$, $x^k \in V^{\otimes k}$.

Proposition 5 *For $i=1,2$, let $\beta_i \in \left(X, E_X^{U_i}\right)$ be a slowly-varying one-form such that $\|\beta_i\|_{\theta_i}^{\omega_i}<\infty$ for a control ω_i and $\theta_i>1$. Define $\beta \in \left(X, E_X^{U_1\otimes U_2}\right)$ as*

$$\beta(X_t):=\left(\int_{r=0}^{t}\beta_1(X_r)\,dX_r\right)\otimes\beta_2(X_t)+\prod\nolimits_{\leq[p]}\left(\beta_1(X_t)\succ\beta_2(X_t)\right)$$

for $t\in[0,1]$. Then with $\omega:=\omega_1+\omega_2+\|X\|_{p-var}^{p}$ and $\theta:=\min(\theta_1,\theta_2)$,

$$\|\beta\|_{\theta}^{\omega}\leq C_{p,\theta,\omega(0,1)}\|\beta_1\|_{\theta_1}^{\omega_1}\|\beta_2\|_{\theta_2}^{\omega_2}. \tag{18}$$

Remark 3 Let $\beta_2(X_t)(x):=x^1$ for $t\in[0,1]$ and $x\in G(V)$, $x=\sum_{k\geq 0}x^k$, $x^k\in V^{\otimes k}$. Then the β defined above corresponds to the integration of β_1, and integration is a continuous operation on slowly-varying one-forms.

Proof Based on Proposition 2, for $\phi^i \in E_{X_t}^{U_i}$, $i = 1, 2$ and $a \in G(V)$,

$$\left(\left(\phi^1\right)_a \succ \left(\phi^2\right)_a\right)(x) = \left(\phi^1 \succ \phi^2\right)_a (x) - \phi^1(a) \otimes \left(\phi^2\right)_a (x)$$

for $x \in G(V)$.

Denote $X_{s,t} := X_s^{-1} X_t$. For $s \le t$,

$$\begin{aligned}
&\left(\beta(X_t) - \left(\beta(X_s)\right)_{X_{s,t}}\right)(x) \\
&= \int_{r=0}^{s} \beta_1(X_r)\, dX_r \otimes \left(\beta_2(X_t) - \left(\beta_2(X_s)\right)_{X_{s,t}}\right)(x) \\
&\quad + \left(\int_{r=s}^{t} \beta_1(X_r)\, dX_r - \beta_1(X_s)\left(X_{s,t}\right)\right) dX_r \otimes \beta_2(X_t)(x) \\
&\quad + \beta_1(X_s)\left(X_{s,t}\right) \otimes \left(\beta_2(X_t) - \left(\beta_2(X_s)\right)_{X_{s,t}}\right)(x) \\
&\quad + \prod\nolimits_{\le [p]} \left(\left(\beta_1(X_t) \succ \beta_2(X_t)\right) - \left(\left(\beta_1(X_s)\right)_{X_{s,t}} \succ \left(\beta_2(X_s)\right)_{X_{s,t}}\right)\right)(x)
\end{aligned}$$

for $x \in G(V)$. Then the estimate (18) holds based on the definition of the operator norm and Theorem 15.

Acknowledgements The author would like to express sincere gratitude to Prof. Terry Lyons for numerous inspiring discussions that eventually lead to this paper. The author also would like to thank Prof. Martin Hairer, Sina Nejad, Dr. Horatio Boedihardjo, Dr. Xi Geng, Dr. Ilya Chevyrev and Vlad Margarint for discussions and suggestions on (an earlier version of) the paper, and thank Prof. Kurusch Ebrahimi-Fard and anonymous referees for constructive suggestions.

References

1. Aguiar, M.: Pre-Ppoisson algebras. Lett. Math. Phys. **54**(4), 263–277 (2000)
2. Berstel, J., Reutenauer, C.: Noncommutative rational series with applications, vol. 137. Cambridge University Press, Cambridge (2011)
3. Boedihardjo, H., Geng, X., Lyons, T., Yang, D.: The signature of a rough path: uniqueness. Adv. Math. **293**, 720–737 (2016)
4. Chen, K.T.: Iterated integrals and exponential homomorphisms. Proc. Lond. Math. Soc. **3**(1), 502–512 (1954)
5. Chen, K.T.: Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. Ann. Math. **65**, 163–178 (1957)
6. Chen, K.T.: Integration of paths – a faithful representation of paths by noncommutative formal power series. Trans. Am. Math. Soc. **89**(2), 395–407 (1958)
7. Ebrahimi-Fard, K.: Loday-type algebras and the Rota–Baxter relation. Lett. Math. Phys. **61**(2), 139–147 (2002)
8. Eilenberg, S., Lane, S.M.: On the groups H (Π, n), I. Ann. Math. **58**, 55–106 (1953)

9. Feyel, D., de La Pradelle, A., Mokobodzki, G.: A non-commutative sewing lemma. Electron. Commun. Probab. **13**, 24–34 (2008)
10. Friz, P., Gassiat, P., Lyons, T.: Physical Brownian motion in a magnetic field as a rough path. Trans. Am. Math. Soc. **367**(11), 7939–7955 (2015)
11. Gubinelli, M.: Controlling rough paths. J. Funct. Anal. **216**(1), 86–140 (2004)
12. Hambly, B., Lyons, T.: Uniqueness for the signature of a path of bounded variation and the reduced path group. Ann. Math. **171**(1), 109–167 (2010)
13. Kawski, M., Sussmann, H.J.: Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory. In: Operators, Systems and Linear Algebra, pp. 111–128. Springer (1997)
14. Loday, J.L.: Dialgebras. In: Dialgebras and Related Operads, pp. 7–66. Springer, Berlin/ Heidelberg (2001)
15. Lyons, T.: On the non-existence of path integrals. Proc. Roy. Soc. A. **432**, 281–290 (1991)
16. Lyons, T.J.: Differential equations driven by rough signals. Rev. Mat. Iberoam. **14**(2), 215–310 (1998)
17. Lyons, T.J., Yang, D.: The theory of rough paths via one-forms and the extension of an argument of Schwartz to rough differential equations. J. Math. Soc. Jpn. **67**(4), 1681–1703 (2015)
18. Lyons, T.J., Caruana, M., Lévy, T.: Differential Equations Driven by Rough Paths. Springer, Berlin/Heidelberg (2007)
19. Lyons, T., Ni, H., Oberhauser, H.: A feature set for streams and an application to high-frequency financial tick data. In: Proceedings of the 2014 International Conference on Big Data Science and Computing. ACM (2014)
20. Malvenuto, C.: Produits et coproduits des fonctions quasi-symétriques et de l'algèbre des descentes. Université du Québec à Montréal, Montréal (1994)
21. Malvenuto, C., Reutenauer, C.: Duality between quasi-symmetrical functions and the solomon descent algebra. J. Algebra **177**(3), 967–982 (1995)
22. Poirier, S., Reutenauer, C.: Algèbres de Hopf de tableaux. Ann. Sci. Math. Quebec **19**(1), 79–90 (1995)
23. Ree, R.: Lie elements and an algebra associated with shuffles. Ann. Math. **68**, 210–220 (1958)
24. Reutenauer, C.: Free Lie algebras. Clarendon Press, Oxford (1993)
25. Ronco, M.: Primitive elements in a free dendriform algebra. Contemp. Math. **267**, 245–264 (2000)
26. Schützenberger, M.P.: Sur une propriété combinatoire des algebres de Lie libres pouvant être utilisée dans un probleme de mathématiques appliquées. Séminaire Dubreil. Algèbre et théorie des nombres **12**(1), 1958–1959 (1958)
27. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions, vol. 2. Princeton University Press, Princeton (1970)
28. Wiener, N.: The quadratic variation of a function and its fourier coefficients. J. Math. Phys. **3**(2), 72–94 (1924)
29. Xie, Z., Sun, Z., Jin, L., Ni, H., Lyons, T.: Learning spatial-semantic context with fully convolutional recurrent network for online handwritten Chinese text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(8), 1903–1917 (2017)
30. Young, L.C.: An inequality of the Hölder type, connected with Stieltjes integration. Acta Math. **67**(1), 251–282 (1936)