Rough paths and SPDE

Ismaël Bailleul, Carlo Bellingeri, Yvain Bruned, Adeline Fermanian, Nicolas Marie

ESAIM: PROCEEDINGS AND SURVEYS, November 2023, Vol. 74, p. 169-184
Editors: Marguerite Zani, Clément Marteau - Journées MAS 2020, Fédération Siméon Poisson (Orléans-Tours),

ROUGH PATHS AND SPDE *

Ismael Bailleul[1], Carlo Bellingeri[2], Yvain Bruned[3], Adeline Fermanian[4] and Nicolas Marie[5]

Abstract. After a brief survey on rough paths theory, from the seminal paper of T. Lyons to its recent developments, this proceeding provides details on C. Bellingeri, Y. Bruned and A. Fermanian's talks during the Journées MAS 2020: a new formulation and a generalization of the signature operator, an extension of branched rough paths called Volterra branched rough paths, and recurrent neural networks investigated as a kernel method thanks to the signature operator and neural ODEs.

Résumé. Après un bref résumé de la théorie des trajectoires rugueuses, de l'article fondateur de T. Lyons à ses récents développements, cet acte fournit quelques détails sur le contenu des exposés de C. Bellingeri, Y. Bruned et A. Fermanian durant les Journées MAS 2020 : une reformulation et une généralisation de l'opérateur signature, une extension de la notion de trajectoire rugueuse branchée appelée trajectoire rugueuse branchée de Volterra, et une étude des réseaux de neurones récurrents vus comme une méthode à noyau grace à l'opérateur signature et aux EDO neuronales.

Introduction: a brief survey on rough paths theory – I. Bailleul and N. Marie

In the seminal paper [75], published in 1936, L.C. Young gave a sense to the integral

$$\int_0^T y_u \, dx_u \tag{1}$$

of a β-Hölder continuous map $y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$ with respect to an α-Hölder continuous function $x : [0,T] \to \mathbb{R}^d$, where $d, e \in \mathbb{N}^*$, $T > 0$, and $\alpha, \beta \in (0,1]$ satisfy $1/\alpha + 1/\beta > 1$. Precisely, L.C. Young proved that the Riemann sums

$$\sum_{k=0}^{n-1} y_{t_k}(x_{t_{k+1}} - x_{t_k})$$

associated with dissections $(t_0, \dots, t_n)$ of the interval $[0,T]$ converge when $n \to \infty$ and that the limit doesn't depend on the choice of $(t_0, \dots, t_n)$. This limit is called the Young integral of y with respect to x on $[0,T]$.

* *The authors thank the GdR TRAG (CNRS) for its support.*

[1] Univ. Rennes, IRMAR - UMR CNRS 6625, 35000 Rennes, France
[2] TU Berlin, Strasse des 17. Juni 135, 10623 Berlin, Germany
[3] Univ. Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, Scotland
[4] Mines ParisTech - CBIO, Institut Curie, INSERM - U900, Sorbonne Université - LPSM, Paris, France
[5] Univ. Paris Nanterre, MODAL'X - UMR CNRS 9023, 92001 Nanterre, France

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202374169

Then, in 1994, for $\alpha \in (1/2, 1]$, T. Lyons [64] used Young's integral in order to define and to solve the differential equation

$$y_t = y_0 + \int_0^t f(y_s) dx_s \tag{2}$$

where $y_0 \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to L(\mathbb{R}^d, \mathbb{R}^e)$ is *smooth enough*. This was done via a fixed point argument. In particular, if B is a stochastic process with Hölder continuous paths of exponent strictly larger than $1/2$, then the stochastic differential equation $dY_t = f(Y_t)dB_t$ can be defined as a collection of equations of type (2) indexed in the canonical probability space where B takes its values. For instance, this works for a fractional Brownian motion of Hurst index strictly larger than $1/2$.

However, between 1936 and the end of the 20th century, non-pathwise approaches had to be developed to study stochastic differential equations driven by the standard Brownian motion. Indeed, by Kolmogorov's continuity theorem, the Brownian motion has α-Hölder continuous paths but for $\alpha < 1/2$ only. Between 1940 and 1950, K. Itô solved the problem via a probabilistic approach, using the martingale property of the Brownian motion – see Itô [57,58]. The current stochastic integral with respect to semi-martingales has resulted from the developments of J.L. Doob, H. Kunita, J. Watanabe, and of the french probability community between 1960 and 1980 – see *e.g.* Jacod [59]. To be able to consider only semi-martingales as driving signal is a limitation of the stochastic calculus. Indeed, for instance, it is not possible to consider the fractional Brownian motion of Hurst index different from $1/2$.

In the seminal paper [68], published in 1998, T. Lyons introduced the theory of rough paths, inspired by K.T. Chen's works [26, 29, 30]. For $\alpha \in (0, 1/2]$, a rough path over a signal x is a α-Hölder continuous map $\mathbb{X}$ from $[0, T]$ into the truncated tensor algebra $T^N(\mathbb{R}^d)$, with $N = [1/\alpha]$, satisfying $\mathbb{X}^1 = x$ and Chen's identity $\mathbb{X}_{s,t} = \mathbb{X}_{s,u} \otimes \mathbb{X}_{u,t}$, for every $0 \leq s \leq u \leq t \leq T$. For instance, if it exists, a *geometric rough path* over an $\mathbb{R}^d$-valued path x is the limit when n goes to infinity, for an appropriate notion of α-Hölder distance for paths taking their values in $T^N(\mathbb{R}^d)$, of the step-N signature

$$S_N(x^n)(t) := \left(1, x_t^n, \int_{0<s_1<s_2<t} dx_{s_1}^n \otimes dx_{s_2}^n, \ldots, \int_{0<s_1<\cdots<s_N<t} dx_{s_1}^n \otimes \cdots \otimes dx_{s_N}^n\right) \tag{3}$$

of a Lipschitz continuous approximation x^n of an α-Hölder control x. Thanks to such *enhancement* of x, T. Lyons bypassed the lack of regularity of the signal x and established a *universal limit theorem* giving sense to Equation (2) and proving its well-posed character. Lyons' first proof of his result involved the formulation of Equation (2) as a fixed point equation in the space of rough paths over $\mathbb{R}^e$. He used a Picard iteration to solve it uniquely.

If B is a fractional Brownian motion of Hurst index $H \in (1/4, 1/2]$, and $\alpha \in (1/4, H)$, and if B^n is the linear approximation of B along the dissection of $[0, T]$ into n equal subintervals, then the step-N signature $S_N(B^n)$ of B^n converges almost surely in $C^\alpha([0, T], T^N(\mathbb{R}^d))$ to a limit $\mathbb{B}$ called the enhanced fractional Brownian motion. In the special case $H = 1/2$ the first component of the rough integral with respect to $\mathbb{B}$ coincides with the so-called second order iterated Stratonovich integral. This connects the rough integration to Itô's calculus mentioned above.

Other approaches of the theory of rough differential equations have been investigated after Lyons' seminal work. On the one hand, in [35], A.M. Davie constructed the solution to Equation (2) directly as the uniform limit of numerical schemes in which the rough paths increments play the role of the Taylor expansion coefficients. This approach was reshaped by P.K. Friz and N. Victoir into a slightly different form, where the rough dynamics driven by a rough path $\mathbf{X}$ is obtained as the limit of dynamics driven by smooth controls whose canonical lift (3) into a rough path converges as a rough path to $\mathbf{X}$. No notion of rough integral is needed in this dynamical picture that has however the drawback of working essentially only with geometric rough paths. Gubinelli developped in [48] another approach. As in T. Lyons theory, the notion of rough integral plays a central role and Equation (2) is formulated as a fixed point problem for an integral equation whose existence and uniqueness is obtained

by a fixed point argument. The cornerstone of Gubinelli's approach is the notion of *controlled path*. A α-Hölder continuous map $y : [0,T] \to L(\mathbb{R}^d, \mathbb{R}^e)$ is controlled by x if and only if there exists a α-Hölder continuous map $y' : [0,T] \to L(\mathbb{R}^d, L(\mathbb{R}^d, \mathbb{R}^e))$ and a 2α-Hölder continuous map $R_y : [0,T]^2 \to L(\mathbb{R}^d, \mathbb{R}^e)$ such that

$$y_t - y_s = y'_s(x_t - x_s) + R_y(s,t)$$

for every $s, t \in [0,T]$ satisfying $s < t$. Let us take $\alpha \in (1/3, 1/2]$ for the sake of simplicity. In [48], M. Gubinelli proved that for every dissection $(t_0, \dots, t_n)$ of $[0,T]$, the compensated Riemann sums

$$\sum_{k=0}^{n-1} (y_{t_k}(x_{t_{k+1}} - x_{t_k}) + y'_{t_k}\mathbb{X}^2(t_k, t_{k+1}))$$

converge when $n \to \infty$ and that the limit does not depend on the choice of $(t_0, \dots, t_n)$. This limit is called the rough integral of y with respect to $\mathbb{X}$ on $[0,T]$. The reader will find in Friz and Hairer's book [43] a very nice account of Gubinelli's approach of rough paths theory.

Rough paths theory has been used and enriched in a number of directions. On the deterministic side the most important development was the introduction by Gubinelli in [49] of the notion of branched rough paths. It involves algebraic structures that are different from the algebraic structure involved in the study of geometric rough paths and turns out to be a good setting for the incorporation of Itô's theory into the rough paths world – as opposed to Stratonovich picture of stochastic dynamics. This interaction between the deterministic setting of rough paths theory and the tools from probability theory that can be used when working with random controls has been one of the guiding force in a number of projects.

- One of these early successes was the exploration of the interaction between Gaussian and Malliavin calculus and rough paths, motivated by the will to prove the existence of densities for fixed times of solutions of rough differential equations driven by Gaussian rough paths. Three landmark papers on this subject were Cass, Litterer and Lyons work [21], in which they introduced the notion of local variation of a rough path and proved integrability estimates for the derivative flow of a rough differential equation, Cass and Friz's work [23], where they gave the first density result for a class of Gaussian processes, and the final word on the subject by Cass, Hairer, Litterer and Tindel [24]. The road was streamlined along the way by the use of Friz and Oberhauser's generalised Fernique's Gaussian estimate [44]. See Friz and Hairer's book for an up-to-date account.
- The use of rough paths methods on questions about homogenization of fast/slow systems was pioneered by Kelly and Melbourne in [62, 63], after an interesting work by Friz, Gassiat and Lyons [42] on the physical Brownian motion in a magnetic field. It was followed by works of Bailleul et al. [3, 4] and was the object of intense recent activity by Melbourne et al. [32] and X.M. Li et al. [47, 53], amongst others.
- The first interaction of rough paths theory with mean field theory was due to Cass and Lyons [22], after which Bailleul et al. [5, 8, 9] pushed the machinery far beyond. The field is rapidly developing.
- Gubinelli' sewing lemma [40, 48] is the workhorse of his approach to rough dynamics. Lê's recent stochastic version of the sewing lemma [65] offers a very interesting refinement that takes profit both of the deterministic mechanics and the stochastic cancellations encoded in a condition on a conditional expectation.
- Since 2015, some papers also deal with *constrained* rough differential equations and rough differential inclusions. As for ordinary differential equations, there are at least two ways to *constrain* the solution to Equation (2) to live in a convex-compact subset C of $\mathbb{R}^e$: to assume that the map f fulfills an invariance condition or to consider a Skorokhod reflection problem associated to Equation (2). On an invariance theorem for rough differential equations, see Coutin and Marie [34]. On the existence of solutions to rough Skorokhod reflection problems with C not depending on time, see Aida [1], and with C depending on time, see Castaing, Marie and Raynaud de Fitte [25] and Allan, Liu and Prömel [2]. To establish the uniqueness of the solution to such problems is a difficult challenge and there exists at least one

example of rough Skorokhod reflection problem having multiple solutions – see Gassiat [46]. About rough differential inclusions, in [7], Bailleul, Brault and Coutin established the existence of solutions to $dy_t \in F(y_t)d\mathbb{X}_t$, where F is a regular enough multifunction.

- One of the talks of the session organised for the 2021 Journées MAS dealt with the use of the signature of data streams for learning purposes. This field has literally exploded after initial works by T. Lyons et al. [66, 67] motivated by diverse applications ranging from hand writing recognition devices to analysis on mental health and financial markets! H. Oberhauser and his co-authors also recently contribute to this field with, for instance, Oberhauser and Király [61].

The two other talks of the 2021 Journées MAS dealt with the deterministic side of rough paths theory. C. Bellingeri's talk dealt with the active field of extending the notion of rough path to richer algebraic settings that the tensor algebra while Y. Bruned's talk introduced a powerful rough paths like setting adapted to the study of Volterra type rough differential equations. Both talks can be understood in the light of the algebraic setting that was developed by Hairer for his theory of regularity structures.

1. The space of iterated integrals: generalizations and new formulations – C. Bellingeri

From the first seminal article on the subject [68] to its many applications and developments today, a key-result in the foundation of rough paths theory is the continuity property of the so-called "Itô-Lyons map". Given in a nutshell, this property ensures that the solution $Y\colon [0,T] \to W$ of a controlled differential equation[1]

$$\dot{Y}_t = f(Y_t)\dot{X}_t$$

is a continuous function (in a weaker norm than the usual one on C^1) of a smooth signal $X\colon [0,T] \to V$ and an additional extra information provided by a finite number of iterated integrals

$$(s,t) \mapsto \int_{\Delta^n_{s,t}} dX_{t_1} \otimes \cdots \otimes dX_{t_n}, \quad \Delta^n_{s,t} = \{s < t_1 < \cdots < t_n < t\}, \quad n \leq N.$$

The presence of a larger input than X itself makes it reasonable to consider the path and the iterated integral as new input of the controlled equation. Due to the non-linear behavior of the higher-order iterated integral, this source can be naturally described as a path with values in a specific group $G(V)$. In this note, we will recall some classic and new ways to look at the group $G(V)$ together with some possible generalisations. These new formulations constitute the algebraic framework to extend the Ito-Lyons map to new classes of controlled dynamics, an approach that has been fruitful in several contexts, see the recent monograph [43].

1.1. Iterated integrals and Lie groups

The classical way to introduce $G(V)$ in [68] is directly taken from the seminal work of Kuo-Tsai Chen [27]. Starting from a finite-dimensional vector space V (infinite-dimensional extensions will be mentioned afterwards), we introduce the *N-truncated tensor algebra*

$$T^N(V) = \bigoplus_{n=0}^{N} V^{\otimes n}, \quad V^{\otimes 0} \approx \mathbb{R},$$

where $\otimes$ is the algebraic tensor product. Elements $\mathbf{x} \in T^N(V)$ will be denoted by $(N+1)$-tuples $\mathbf{x} = (\mathbf{x}^0, \ldots, \mathbf{x}^N)$ with $\mathbf{x}^k \in V^{\otimes k}$ for $k = 0, \ldots, N$, and are called *truncated tensors*. In addition to the vector space structure

[1] We use the standard notations W, V to denote two real vector spaces and $f\colon W \to L(V,W)$ to denote a sufficiently smooth vector field with values in the linear maps from V to W.

inherited from V, it is also possible to define over $T^N(V)$ a unital algebra structure using the *truncated tensor product* $\otimes_N$, a proper quotient of $\otimes$ defined for any couple $\mathbf{x}, \mathbf{y} \in T^N(V)$ by

$$(\mathbf{x} \otimes_N \mathbf{y})^k = \sum_{\substack{l,m \geq 0 \\ l+m=k}} \mathbf{x}^l \otimes \mathbf{y}^m, \quad k = 0, \dots, N.$$

The unity element of $T^N(V)$ is given by $\mathbf{1} = (1, 0, \dots, 0)$.

The usual way to encode iterated integrals in $T^N(V)$ is via the map $\mathbf{X}\colon [0,T]^2 \to T^N(V)$ given by

$$\mathbf{X}_{s,t} = \left(1, X_t - X_s, \dots, \int_{\Delta^N_{s,t}} dX_{t_1} \otimes \cdots \otimes dX_{t_N}\right) \tag{4}$$

A first important property that shows the emergence of a group structure to describe $\mathbf{X}$ is *Chen's relation*, [27]: for any $s, u, t \in [0,T]$, one has

$$\mathbf{X}_{s,t} = \mathbf{X}_{s,u} \otimes_N \mathbf{X}_{u,t} \tag{5}$$

This formula implies that the two parameters of $\mathbf{X}$ behave multiplicatively with respect to $\otimes_N$. Furthermore, because of the identity $\mathbf{X}_{t,t} = \mathbf{1}$, Equation (5) allows also to write $\mathbf{X}$ in terms of the path $t \mapsto \mathbf{X}_{0,t}$ as long as this path takes values in a group $G(V)$ with operation $\otimes_N$ and unity $\mathbf{1}$.

From the truncated tensor product we introduce the natural Lie bracket associated with it, $[\mathbf{x}, \mathbf{y}] = \mathbf{x} \otimes_N \mathbf{y} - \mathbf{y} \otimes_N \mathbf{x}$, and we define

$$L(V) = V \oplus [V,V] \oplus [V,[V,V]] \oplus \cdots \subset \{\mathbf{x} \in T^N(V)\colon \mathbf{x}^0 = 0\}.$$

The vector space $L(V)$ inherits naturally a Lie algebra structure by restriction of $[.,.]$ over it, which is known in the literature as the *step-N nilpotent free Lie algebra* over V, [72]. Consequently, $G(V)$ can be defined as the unique simply connected Lie group with related Lie algebra $L(V)$. We call it the *step-N nilpotent free group over V*. More explicitly, we set

$$G(V) = \exp(L(V)) \tag{6}$$

where $\exp\colon T^N(V) \to T^N(V)$ is the truncated exponential

$$\exp(\mathbf{x}) = \sum_{n=0}^{N} \frac{\mathbf{x}^{\otimes_N^n}}{n!}.$$

Thanks to the identity $\exp(\mathbf{v})^{-1} = \exp(-\mathbf{v})$ and the Baker-Campbell-Hausdorff formula, $G(V)$ is indeed a Group compatible with $\otimes_N$ and unity $\mathbf{1}$. Moreover, any given smooth signal X induces a map $\mathbf{X}$ with values in $G(V)$ as consequence of a differential equation satisfied by the path $t \mapsto \mathbf{X}_{0,t}$, see [27, 68].

Since this original formulation, the presence of an explicit geometric structure to study $\mathbf{X}$ has motivated several lines of research all along the years, see the monograph [45], where the intrinsic topology of $G(V)$ plays a key-role to deduce properties for rough/stochastic differential equations. We mention also the review [41] for further links with sub-Riemannian geometry.

1.2. **Iterated integrals and Hopf algebras**

Despite the actual success of the first definition (6), this approach has proved inconvenient in some probabilistic contexts. Indeed, if we want to extend map (4) to a stochastic process with values in the manifold $G(V)$, there might be some hidden technicalities. Besides, $G(V)$ is modeled upon iterated integrals of smooth paths $X\colon [0,T] \to V$ satisfying integration by part formulae like

$$\int_{\Delta^2_{s,t}} \mathrm{d}X_{t_1} \otimes \mathrm{d}X_{t_2} + \int_{\Delta^2_{s,t}} \mathrm{d}X_{t_2} \otimes \mathrm{d}X_{t_1} = (X_t - X_s) \otimes (X_t - X_s).$$

Therefore, $G(V)$ cannot model any integration theory for which similar identities do not hold, *e.g.* Itô calculus. To understand how integration by part formulae appear, we can describe $G(V)$ from the shuffle algebra.

We define it by starting with the identification of V with its double dual $(V^*)^*$ and looking at $T^N(V)$ as the dual of

$$T^N(V^*) = \bigoplus_{n=0}^{N} (V^*)^{\otimes n}.$$

Supposing V finite-dimensional and isomorphic to $\mathbb{R}^d$, we can write elements $\mathbf{x} \in T^N(V^*)$ as linear combinations of words in the alphabet $\{1, \dots, d\}$ with length smaller than N, union the empty word $\emptyset$, via the identifications

$$(1, 0, \dots, 0) = \emptyset, \quad i_1 \cdots i_n = e^{i_1} \otimes \cdots \otimes e^{i_n},$$

where $e^1, \dots, e^d$ are the elements of the dual canonical basis of V^*.

This identification allows to introduce two new algebraic operations on $T^N(V^*)$: *the shuffle product* $\sqcup\!\sqcup$, given by the identity $v \sqcup\!\sqcup \emptyset = \emptyset \sqcup\!\sqcup v = v$ and extended recursively[2] with the relation

$$wi \sqcup\!\sqcup vj := (w \sqcup\!\sqcup vj)i + (wi \sqcup\!\sqcup v)j \tag{7}$$

for any couple of words w, v and letters $i, j \in \{1, \dots, d\}$; the *deconcatenation coproduct* $\Delta \colon T^N(V^*) \to T^N(V^*) \otimes T^N(V^*)$, defined on any word $w = i_1 \cdots i_n$ by

$$\Delta w = \emptyset \otimes w + \sum_{k=1}^{n-1} i_1 \cdots i_k \otimes i_{k+1} \cdots i_n + w \otimes \emptyset.$$

Using standard results from Lie polynomials, see [72, Thm.3.2], we can actually describe the operation $\otimes_N$ and the group $G(V)$ via the previous operations and the natural pairing $\langle ., . \rangle$ between $T^N(V^*)$ and $T^N(V)$. In particular, we have the identities

$$\langle f \otimes_N g, w \rangle = \langle f \otimes g, \Delta w \rangle \quad \text{for any } f, g \in T^N(V) \tag{8}$$

$$G(V) = \{ g \in T^N(V) \colon g \neq 0 \ ; \ \langle g, w \sqcup\!\sqcup v \rangle = \langle g, w \rangle \langle g, v \rangle \quad \text{for any } w, v \in T^N(V^*) \} \tag{9}$$

thereby obtaining a new definition for $G(V)$.

The advantages of this equivalence are multiple and make $G(V)$ more treatable for probabilistic applications. Firstly, it encodes the non-linear structure of $G(V)$ in a linear structure with the additional relations coming from $\sqcup\!\sqcup$, which represent integration by part formulae when we write them in terms of $\mathbf{X}$. Furthermore, it represents $G(V)$ as the group of characters of a Hopf algebra. In few words, this is a triple $(\mathcal{H}, m, \delta)$ of a vector space $\mathcal{H}$, a product m and a coproduct δ such that the set

$$G(\mathcal{H}) = \{ g \in \mathcal{H}^* \colon g \neq 0 \ ; \ \langle g, m(w, v) \rangle = \langle g, w \rangle \langle g, v \rangle \quad \text{for any } w, v \in \mathcal{H} \}$$

is a group with the operation

$$\langle f * g, w \rangle = \langle f \otimes g, \delta(w) \rangle \quad \text{for any } f, g \in \mathcal{H}^*.$$

From this point of view, it is then possible to study the group of characters of many Hopf algebras and the properties of paths with values on them. We mention in particular the paper [11], where iterated Itô integrals of a semimartingale are seen as paths with values in the character group of a specific quasi-shuffle algebra, see [56]. We also recall that smooth paths with values in character groups of a specific class of Hopf algebras have several interesting properties, see [13].

[2]The original domain for $\sqcup\!\sqcup$ is the full tensor algebra $T(V^*) = \bigoplus_{n=0}^{\infty} (V^*)^{\otimes n}$. We can restrict it to $T^N(V^*)$ by setting $w \sqcup\!\sqcup v = 0$ whenever the sum of the lengths of w and v is strictly bigger than N.

1.3. **Iterated integrals and representations**

We conclude this note with the existence of an explicit representation of $G(V)$ on $T^N(V^*)$. This property will be linked with a weaker description of $G(V)$ obtained in [14].

At the basis of this representation we can use the coproduct Δ to define, for any $g \in T^N(V)$, a linear map $\Gamma_g \colon T^N(V^*) \to T^N(V^*)$ given by

$$\Gamma_g w = (\mathrm{id} \otimes g)\Delta w,$$

where we look g as a linear map $g \colon T^N(V^*) \to \mathbb{R}$ via the natural pairing. This application $g \mapsto \Gamma_g$ is not simply an injective linear map, but we can actually define a representation when we restrict ourselves to $G(V)$. Thanks to the Identities (8) and (9), the elements of $G(V)$ satisfy also the additional properties

$$\Gamma_g \Gamma_h = \Gamma_{g \otimes_N h}, \quad \Gamma_g(w \sqcup\!\sqcup v) = \Gamma_g w \sqcup\!\sqcup \Gamma_g v, \quad \Gamma_g \emptyset = \emptyset,$$

for any $g, h \in G(V)$ and any $w, v \in T^N(V^*)$. Therefore, we obtain a representation of $G(V)$ with values in the smaller group $\mathrm{End}_{\sqcup\!\sqcup}(T^N(V^*))$ of invertible $\sqcup\!\sqcup$-algebra endomorphisms. Recalling the definition of Δ, we can prove the strict inclusion

$$G(V) \subset \{\Gamma \in \mathrm{End}_{\sqcup\!\sqcup}(T^N(V^*)) : (\Gamma - \mathrm{id})((V^*)^{\otimes n}) \subset T^{n-1}(V^*) \quad \text{for any } 1 \leq n \leq N\} \tag{10}$$

thereby identifying $G(V)$ with a proper subset of triangular matrices.

Even if property (10) is not a characterization, the group on the right-hand side of (10) might constitute a possible replacement of $G(V)$ when there is no proper Hopf algebra. This phenomenon happens in a recent result [14], where we study smooth drivers X with values in an infinite-dimensional $C^\star$-algebra $\mathcal{A}$ with product $\cdot$ and involution $\star$. Inspired by the properties of the product Levy area [36], we enrich X with a different family of non-linear functionals of the form

$$(s,t) \mapsto X^\sigma_{s,t}(A_0, \dots, A_n) = \int_{\Delta^n_{s,t}} A_0 \cdot \mathrm{d}X_{t_{\sigma(1)}} \cdot A_1 \cdots A_{n-1} \cdot \mathrm{d}X_{t_{\sigma(n)}} \cdot A_n \tag{11}$$

for any $1 \leq n \leq N$, any permutation $\sigma \in \mathcal{S}_n$, and $A_0, \dots, A_n \in \mathcal{A}$. The multi-linear operators in (11) are deduced from iterated integrals but they belong naturally to the more sophisticated operadic structures over $\mathcal{A}$. In this context, it seems there is group of characters containing them but it is still possible to define a graded algebra $\mathrm{LT}^\sharp(\mathrm{FC})$, a proper group $G(\mathrm{FC})$ of endomorphisms over it and a map $\mathcal{X} \colon [0,T]^2 \to G(\mathrm{FC})$ encoding the elements in (11), see [14, Thm.4.18].

A group like $G(\mathrm{FC})$ might constitute a new type of algebraic structure to extend the "rough path philosophy" to new classes of controlled differential equations. In particular, $G(\mathrm{FC})$ was tailored to study rough differential equations driven by non-commutative processes.

2. Ramification of Volterra-type rough paths – Y. Bruned

Volterra equations comprise a thoroughly studied class of differential equations capable of adequately capturing the behavior of a wide range of natural models: viscoelastic material, spread of epidemics and volatility models in mathematical finance. We are interested in stochastic Volterra equations of the form:

$$u(t) = u_0 + \sum_{i=0}^{d} \int_0^t k(t,r) f_i(u_r) dq^i_r, \quad u_0 \in \mathbb{R}^e \tag{12}$$

where k is a kernel allowed to be singular in the diagonal $t = r$, the f_i's are sufficiently regular vector fields on $\mathbb{R}^e$ and $q \in C^\alpha([0,T]; \mathbb{R}^{d+1})$ a.s. is a stochastic process on $\mathbb{R}^{d+1}$ with $q^0_r = r$ that is Hölder regular of some degree $\alpha \in (0,1]$ almost surely. We are interested in pathwise techniques such as rough paths [48, 49, 68] for giving a meaning to these equations for low regularity α. The first works in this direction are coming from

Deya and Tindel that treat non-singular Volterra equations with rough paths theory in [37] and [38]. A more recent approach is given by paracontrolled calculus [6,50] in [71] by Prömel and Trabs, where the authors treat first order case namely $\alpha \geq \frac{1}{3}$. With regularity structures [51], one can reach this threshold in [10,12], with the expectation that methods therein are amenable to generalization in the case of arbitrarily low exponent. One should then be able to obtain existence and uniqueness results that are, however, only local.

In this work, we are interested in extending the approach introduced lately in [54] and [55], which is to generalize the ideas of rough path theory in order to treat the case of stochastic Volterra equations. The idea is to keep track of iterated integral convolutions with the Volterra kernel. They consider a generalization of rough paths with a new Chen's type relation via a convolution product. They have only treated the cases of Hölder regularity that is higher than $1/3$ and $1/4$ respectively. In this work, we clarify the Hopf-algebraic framework necessary for the description of the iterated integrals at play. Our idea is based on a plugging coproduct used in [18] for recovering the algebraic structures of [17]. Note that such a structure is also used in numerical analysis [19]. It can be understood as a Butcher-Connes-Kreimer type coproduct [20,33], where one keeps along the edges that would normally be lost when performing an admissible cut. Pinpointing certain properties of convolutional integrals in a sufficiently regular setting as well as drawing from the ideas of branched rough paths together with the more suitable framework provided by this algebraic structure, we are able to formulate the theory on any order and to extend the main results obtained in [55].

We first introduce a natural space of decorated rooted trees. Let $\widehat{\mathcal{T}}$ (resp. $\mathcal{T}$) be the set of rooted trees with nodes decorated by $\{0,\dots,d\}$ (resp. decorated by $\{0,\dots,d\}$ except for the root which carries no decoration). We grade elements $\tau \in \widehat{\mathcal{T}}$ (resp. $\mathcal{T}$) by the number $|\tau|$ of their nodes having a decoration and we set

$$\mathcal{T}_n := \{\tau \in \mathcal{T} : |\tau| \leq n\}, \quad n \in \mathbb{N}$$

(and resp. for $\widehat{\mathcal{T}}$). We denote by $\widehat{\mathcal{F}}$ (resp. $\mathcal{F}$) the set of forests, *i.e.* set consisting of trees in $\widehat{\mathcal{T}}$ (resp. $\mathcal{T}$). For any $h \in \mathcal{F}$ or $h \in \widehat{\mathcal{F}}$ we shall denote by E_h the set of its edges and by N_h the set of its nodes. Let $T > 0$, we will denote by $\Delta_n = \Delta_n([0,T])$ the subset $\{0 < t_1 < \cdots < t_n < T\}$ of $\mathbb{R}^n$.

Definition 2.1. Let q be a path in $C^1([0,T];\mathbb{R}^{d+1})$ and k a Volterra kernel. Let $h \in \mathcal{T}_{N+1}$ be a rooted tree with $n+1$ vertices. Let $h^\star \in \widehat{\mathcal{F}}$ denote the forest one obtains after removing the root of h and its adjacent edges. Then, using the convention that $r_\varrho = \tau$, where ϱ is the root of h, we define the h-th *iterated Volterra integral* as a mapping $\mathbf{z}^h : \Delta_3 \to \mathbb{R}$ given by

$$\mathbf{z}^{h,\tau}_{ts} = \int_{A^h_{ts} \subseteq \mathbb{R}^n} \prod_{(i,j)\in E_h} k(r_i, r_j) \prod_{\ell \in N_h} dq^{i_\ell}_{r_\ell},$$

where i_ℓ is the decoration attached to the node ℓ and the domain of integration is the set

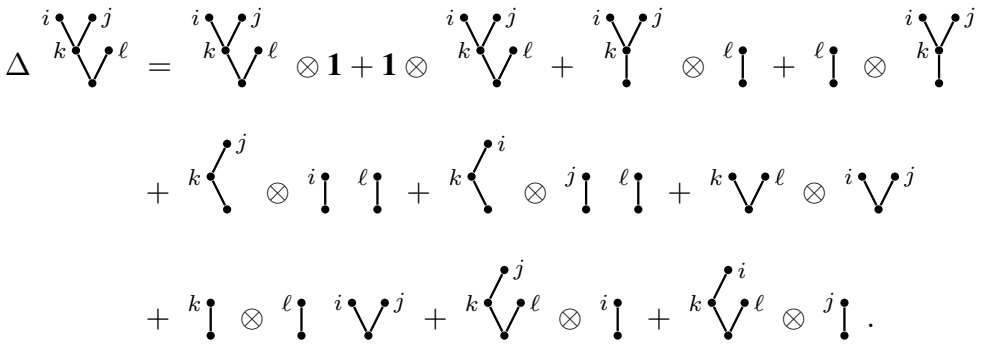
$$A^h_{ts} = \bigcap_{(i,j)\in E_{h^\star}} \{t > r_i > r_j > s\},$$

i.e. the order relations defining the variable ranges are directly given by the partial ordering induced by the forest $h^\star \in \widehat{\mathcal{F}}$. Let $V \subset N_h$ be of cardinality m. Then, we also define:

$$\overline{\mathbf{z}}^{h,\tau}_{ts}((r_\ell)_{\ell\in V}) = \prod_{w\in V} \dot{q}^{i_w}_{r_w} \int_{A^h_{ts}((r_\ell)_{\ell\in V}) \subseteq \mathbb{R}^{n-m}} \prod_{(i,j)\in E_h} k(r_i, r_j) \prod_{\ell \in N_h \setminus V} dq^{i_\ell}_{r_\ell}$$

where $A^h_{ts}((r_\ell)_{\ell\in V})$ corresponds to A^h_{ts} when one fixes the values of $(r_\ell)_{\ell\in V}$.

In the sequel, we will use a Butcher-Connes-Kreimer type coproduct $\Delta : \mathcal{F} \to \mathcal{F} \otimes \mathcal{F}$ that we illustrate on an example:



Using Sweedler's notation for the coproduct of a forest h, we will write:

$$\Delta h = \sum_{(h)} h^{(1)} \otimes h^{(2)}.$$

In the sequel, we use an extension of the map Δ by identifying the nodes. We will consider the nodes of $h^{(1)}$ and $h^{(2)}$ as a subset of those of h. Therefore, the roots of the trees in $h^{(2)}$ will be identified with some nodes in $h^{(1)}$. This property is crucial in order to define a convolution operation on tree-indexed iterated integrals. We also define the reduced coproduct $\widetilde{\Delta}$ as follows:

$$\widetilde{\Delta} h = \Delta h - h \otimes \mathbf{1} - \mathbf{1} \otimes h = \sum_{(h)'} h^{(1)} \otimes h^{(2)}.$$

We extend Definition 2.1 to objects with forest indices. Let $h = h_1 \cdot \cdots \cdot h_n$ be in $\mathcal{F}$. We set:

$$\mathbf{z}_{ts}^{h,\tau_1,\ldots,\tau_n} = \prod_{i=1}^{n} \mathbf{z}_{ts}^{h_i,\tau_i}$$

where $\tau_1, \ldots, \tau_n \in [s,t]$. Then, one has the following convolution identity: let h be a tree in $\mathcal{T}$ and $(s,u,t,\tau) \in \Delta_4$, we have

$$\mathbf{z}_{ts}^{h,\tau} = \sum_{(h)} \mathbf{z}_{tu}^{h^{(1)},\tau} \star \mathbf{z}_{us}^{h^{(2)},\cdot} \tag{13}$$

where the convolution product $\star$ is defined as follows

$$\mathbf{z}_{tu}^{h^{(1)},\tau} \star \mathbf{z}_{us}^{h^{(2)},\cdot} := \int_{\mathbb{R}^m} \overline{\mathbf{z}}_{tu}^{h^{(1)},\tau}((r_i)_{i\in V}) \prod_{i\in V} \mathbf{z}_{us}^{h_i^{(2)},r_i} dr_i.$$

Here, $V \subset N_h$ is of cardinality m and is such that every $i \in V$ considered as a node of h appears also as a root of a tree $h_i^{(2)}$ in $h^{(2)}$ and $h^{(2)} = \prod_{i\in V} h_i^{(2)}$. One can extend the definition of the convolution recursively in the rough case and provide a definition of a Volterra branched rough path. It generalizes the classical definition of branched rough paths given in [49, 52].

Definition 2.2 (Volterra branched rough path)**.** Fix $\alpha, \gamma \in (0,1)$ such that $\alpha - \gamma > 0$. Let $(z_i)_{i\in\{0,\dots,d\}}$ be such that $z_i \in \mathcal{V}^{(\alpha,\gamma)}(\Delta_2, \mathbb{R})$. For n with $(n+1)\varrho + \gamma > 1$, we suppose given a tree-indexed family of iterated integrals $(\mathbf{z}_{ts}^{\tau,h})_{|h|\leq n}$ indexed by the trees of $\mathcal{T}$ such that

$$\mathbf{z}^{e_i} = z_i, \quad e_i = \ {}^{i}\!\bullet, \quad \delta_u \mathbf{z}_{ts}^{\tau,h} = \sum_{(h)'} \mathbf{z}_{tu}^{h^{(1)}} \star \mathbf{z}_{us}^{h^{(2)}}$$

where the Sweedler's notation corresponds to the reduced coproduct $\widetilde{\Delta}$. Let h be a tree with $m \leq n$ nodes. We suppose that for every $y \in \mathcal{V}_m^{(\alpha,\gamma)}$, one has

$$\delta_u \mathbf{z}_{ts}^{h,\tau} \star y_s = \sum_{(h)'} \left(\mathbf{z}_{tu}^{h^{(1)},\tau} \star \left(\mathbf{z}_{us}^{h^{(2)}} \star y_s \right) \right).$$

We also assume that $\mathbf{z}^h \in \mathcal{V}^{(|h|\varrho+\gamma,\gamma)}$.

We have omitted the precise definitions of the spaces $\mathcal{V}^{(\alpha,\gamma)}$ and $\mathcal{V}_m^{(\alpha,\gamma)}$. These spaces reflect the regularity assumed on both z_i and y. It depends on the γ-singularity of the kernel k considered and the α-Hölder regularity of the drivers of (12). Below, we define a new space of controlled Volterra branched rough paths that is suitable for running a fixed point argument for the Equation (12).

Definition 2.3 (Controlled Volterra branched rough path)**.** Let $\mathbf{z}$ be an α-Hölder Volterra branched rough path and let $n = \lfloor 1/\alpha \rfloor$. A Volterra branched rough path controlled by $\mathbf{z}$ is a function $y = (y^h)_{h\in\mathcal{T}_{n-1}}$ such that, for every $h \in \mathcal{T}_{n-1}$, we have $y^h \in \mathcal{V}_{|h|+1}^{(\alpha,\gamma)}$ and the remainder terms, for every $\tau \in [s,t]$,

$$R_{ts}^h = y_{ts}^h - \sum_{\varrho\in\mathcal{F}_{n-1}} \sum_{\sigma\in\mathcal{T}_{n-1}} c(\sigma, h, \varrho) \mathbf{z}_{ts}^{\varrho} \star y_s^{\sigma} \tag{14}$$

satisfy $R^h \in \mathcal{V}_{|h|+1}^{((n-|h|)\alpha,(n-|h|)\gamma)}$. Here $c(\sigma, h, \varrho)$ is the counting function for the number of appearances of the term $h \otimes \varrho$ in the expansion of the reduced coproduct $\widetilde{\Delta}\sigma$. The space of such functions is called the space of controlled Volterra branched rough paths.

3. Framing RNN as a kernel method: A neural ODE approach – A. Fermanian

Recurrent neural networks (RNN) are among the most successful methods for modeling sequential data. They have achieved state-of-the-art results in difficult problems such as natural language processing or speech recognition. This class of neural networks has a natural interpretation in terms of discretization of ordinary differential equations (ODE), which casts them in the field of neural ODE (Chen et al. [32]). This paradigm allows us to show that RNN are, in the continuous-time limit, linear predictors over a specific space associated with the signature of the input sequence (Levin et al. [66]). This frames RNN as a kernel method, which allows to obtain statistical guarantees. The results presented here have been published in Fermanian et al. [39]. Many assumptions and technical details are omitted for clarity, we refer to Fermanian et al. [39] for precise statements and proofs.

3.0.1. *Mathematical context*

We place ourselves in a binary classification setting: the data is a set of n i.i.d. pairs

$$\mathcal{D}_n = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\},$$

where $\mathbf{x}^{(i)}$ is a vector-valued sequence:

$$\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_T^{(i)}) \in (\mathbb{R}^d)^T, \quad T \geq 1,$$

and $\mathbf{y}^{(i)} \in \{-1, 1\}$. For example $\mathbf{x}^{(i)}$ may correspond to the recording of d physiological variables of a patient over a period of time. The label $\mathbf{y}^{(i)}$ may then correspond to saying whether the patient should be sent to an intensive care unit or not. Even if we only observe discrete sequences, each $\mathbf{x}^{(i)}$ is mathematically considered as a regular discretization of a continuous-time process $X^{(i)} : [0, 1] \to \mathbb{R}^d$ – we have that

$$x_j^{(i)} = X^{(i)}_{\frac{j}{T}}, \quad 1 \leq j \leq T.$$

To compute its signature, the main assumption on a process $X : [0, 1] \to \mathbb{R}^d$ is that it is of finite variation, that is, of finite length. From now on, we denote by $BV^c([0, 1], \mathbb{R}^d)$ the space of continuous functions from $[0, 1]$ to $\mathbb{R}^d$ of finite total variation, and assume that for any $1 \leq i \leq n$, $X^{(i)} \in \mathcal{X}$, where

$$\mathcal{X} = \{X \in BV^c([0, 1], \mathbb{R}^d) \mid X_0 = 0 \quad \text{and} \quad \|X\|_{TV;[0,1]} \leq L < 1\}.$$

3.1. **The signature**

The signature, first defined by Chen [28] and central in rough path theory, summarizes sequential inputs by a graded feature set of their iterated integrals. It has several good properties, summarized in Proposition 3.3, that make it a relevant tool in machine learning (Levin et al. [66]). Combined with deep neural networks, it has achieved state-of-the-art performance for several applications (see, *e.g.*, Yang et al. [74]; Morrill et al. [70]).

3.1.1. *Tensor Hilbert spaces*

The natural environment of signatures is a tensor space that can be endowed with a Reproducing Kernel Hilbert Space (RKHS) structure (Király and Oberhauser [61]). We denote by $(\mathbb{R}^d)^{\otimes k}$ the k-th tensor power of $\mathbb{R}^d$ with itself, which is a Hilbert space of dimension d^k. Our space of interest consists of infinite square-summable sequences of tensors of increasing order:

$$\mathcal{T} = \{a = (a_0, \dots, a_k, \dots) \mid a_k \in (\mathbb{R}^d)^{\otimes k}, \ \sum_{k=0}^{\infty} \|a_k\|^2_{(\mathbb{R}^d)^{\otimes k}} < \infty\}.$$

Endowed with the scalar product $\langle a, b\rangle_{\mathcal{T}} := \sum_{k=0}^{\infty} \langle a_k, b_k\rangle_{(\mathbb{R}^d)^{\otimes k}}$, $\mathcal{T}$ is a Hilbert space.

Definition 3.1. Let $X \in BV^c([0, 1], \mathbb{R}^d)$. For any $t \in [0, 1]$, the signature of X on $[0, t]$ is defined by $S_{[0,t]}(X) = (1, \mathbb{X}^1_{[0,t]}, \dots, \mathbb{X}^k_{[0,t]}, \dots)$ where, for each $k \geq 1$,

$$\mathbb{X}^k_{[0,t]} = k! \int\limits_{0 \leq u_1 < \cdots < u_k \leq t} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_k} \in (\mathbb{R}^d)^{\otimes k} \tag{15}$$

The expert reader is warned that this definition differs from the usual one by the normalization of $\mathbb{X}^k_{[0,t]}$ by $k!$, which is more adapted to our context. When the signature is taken on the whole interval $[0, 1]$, we simply write $S(X)$ and $\mathbb{X}^k$. The integrals in (15) should be understood a Riemann-Stieljes integrals. Although this definition is technical, the signature should simply be thought of as a feature map that embeds a bounded variation process into an infinite-dimensional tensor space. We refer to Lyons et al. [69], Chapter 2, for more details.

Example 3.2. Let X be the d-dimensional linear path defined by $X_t = (a_1 + b_1 t, \dots, a_d + b_d t)^\top$, $a_i, b_i \in \mathbb{R}$. Then $\mathbb{X}^k = b^{\otimes k}$.

Proposition 3.3. *Let $X \in BV^c([0,1], \mathbb{R}^d)$. We have the following properties:*

(*i*) *The signature is invariant by translation and reparametrization. More precisely, for any $x_0 \in \mathbb{R}^d$, if $\widetilde{X}$ is the path defined by $\widetilde{X} = x_0 + X$, then $S(\widetilde{X}) = S(X)$. Similarly, for any continuous increasing bijection $\psi : [0,1] \to [0,1]$, if $\widetilde{X}_t = X_{\psi(t)}$, then $S(\widetilde{X}) = S(X)$.*

(*ii*) *If X has at least one monotone coordinate, then $S(X)$ determines X uniquely, up to translation and reparametrization.*

(*iii*) *For any $k \geq 0$, if $\|X\|_{TV}$ denotes the total variation of X, then $\|\mathbb{X}^k\|_{(\mathbb{R}^d)^{\otimes k}} \leq \|X\|_{TV}^k$.*

3.1.2. *The signature kernel*

Generally, the idea of kernel methods in machine learning is to map data to a high dimensional space so that linear methods may be used in this space, while being highly nonlinear in the original data (see, *e.g.*, Schölkopf and Smola [73]). By taking advantage of the structure of Hilbert space of $\mathcal{T}$, it is natural to introduce the following kernel:

$$K : \left\{ \begin{array}{ccc} \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathbb{R} \\ (X, Y) & \longmapsto & \langle S(X), S(Y) \rangle_{\mathcal{T}} \end{array} \right. \tag{16}$$

which is well defined according to Proposition 3.3 (indeed, if $X \in \mathcal{X}$, we then have $\|S(X)\|_{\mathcal{T}} \leq 2(1-L)^{-1} < \infty$). The RKHS associated with K is the space of functions

$$\mathcal{H} = \{\xi_\alpha : \mathcal{X} \to \mathbb{R} \mid \xi_\alpha(X) = \langle \alpha, S(X) \rangle_{\mathcal{T}},\, \alpha \in \mathcal{T}\},$$

with scalar product $\langle \xi_\alpha, \xi_\beta \rangle_{\mathcal{H}} = \langle \alpha, \beta \rangle_{\mathcal{T}}$.

3.2. **Framing RNN as a kernel method**

Our main result (Theorem 3.4) shows that RNN may be rewritten as a kernel method associated to the signature kernel K defined by (16). The proof is based on the neural ODE paradigm of Chen et al. [32], who observed that infinite-depth residual neural networks exactly correspond to a specific type of ordinary differential equations (ODE), called neural ODE. In other words, any residual neural networks may be seen as an Euler discretization of a neural ODE. This connection with differential equations is very rich and may be used to develop new algorithms, for example for irregularly sampled time series (Kidger et al. [60]). Here, we use it to leverage the theory of ODE and signatures to obtain theoretical guarantees for RNN.

3.2.1. *Recurrent Neural Networks*

We consider a (residual) RNN defined by a sequence of hidden states $h_1, \ldots, h_T \in \mathbb{R}^e$, where for $\mathbf{x} = (x_1, \ldots, x_T)$ a generic data sample, for any $0 \leq j \leq T-1$,

$$h_{j+1} = h_j + \frac{1}{T} f_\theta(h_j, x_{j+1}) \tag{17}$$

At each time step j, the output of the network is $z_j = \psi(h_j)$, where ψ is a linear function. The simplest choice for the function f_θ is the feedforward model, defined by $f_\theta(h, x) = \sigma(Uh + Vx + b)$, where σ is an activation function, $U \in \mathbb{R}^{e \times e}$ and $V \in \mathbb{R}^{e \times e}$ are weight matrices, and $b \in \mathbb{R}^e$ is the bias (θ is then a vector containing all the coefficients in U, V and b).

3.2.2. *Neural ODE*

Following the neural ODE paradigm of Chen et al. [32], the recursive equation (17) may be seen as an Euler discretization of an ordinary differential equation (ODE) of the form

$$dH_t = f_\theta(H_t, X_t)dt, \quad H_0 = h_0, \quad t \in [0,1] \tag{18}$$

where $H : [0,1] \to \mathbb{R}^e$ is a continuous-time hidden state. Classical results on Euler discretizations show that the difference between h_j and $H_{j/T}$ is a $\mathcal{O}(1/T)$: when the depth of the RNN increases, the approximation of (17) by (18) becomes better.

3.2.3. *Main result*

By increasing the dimension of the hidden state, Kidger et al. [60] show how, with a simple algebraic trick, the ODE (18) can be rewritten as a controlled differential equation (CDE) of the form

$$d\overline{H}_t = \mathbf{F}_\theta(\overline{H}_t)d\overline{X}_t \tag{19}$$

where $\overline{H} : [0,1] \to \mathbb{R}^{\overline{e}}$, $\overline{e} = e + d$, is a new hidden state, $\overline{X} : [0,1] \to \mathbb{R}^{\overline{d}}$, $\overline{d} = d + 1$, is the time-augmented process $\overline{X}_t = (X_t^\top, \frac{1-L}{2}t)^\top$, and $\mathbf{F}_\theta : \mathbb{R}^{\overline{e}} \to \mathbb{R}^{\overline{e}\times\overline{d}}$ is a tensor field such that the right hand side of (19) is a matrix-vector multiplication. Finally, using a Taylor expansion of $\overline{H}$, under some assumptions on the regularity of $\mathbf{F}_\theta$, it is possible to show the following result.

Theorem 3.4. *Let z_T be the output of a RNN defined by* (17)*. There exists $\alpha(\theta) \in \mathcal{T}$ such that*

$$|z_T - \langle \alpha(\theta), S(\overline{X})\rangle_{\mathcal{T}}| \leq \frac{C_1}{T},$$

where $\alpha(\theta)$ depends only on the parameters of the RNN and C_1 is a constant.

In other words, if g_θ is the function that maps an input sequence $\mathbf{x}$ to the final output of the RNN (that is, $g_\theta(\mathbf{x}) = z_T$), then $g_\theta \in \mathcal{H}$ at a $\mathcal{O}(1/T)$ error term. This result allows to reinterpret the action of the recurrent network (RNN) as a scalar product in an (infinite-dimensional) Hilbert space, thereby framing the RNN as a kernel method.

3.2.4. *Generalization bounds*

A first consequence of Theorem 3.4 is that it gives natural generalization bounds under mild assumptions. For example, going back to the binary classification problem, where the predicted class is $2 \cdot \mathbf{1}_{g_\theta(\mathbf{x})>0} - 1$. The parameters $\theta \in \Theta$ of the RNN are fitted by empirical risk minimization using a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, for example the logistic loss. The training loss is defined by

$$\widehat{\mathcal{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{y}^{(i)}, g_\theta(\mathbf{x}^{(i)})),$$

and we let $\widehat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta}\widehat{\mathcal{R}}_n(\theta)$. Using the kernel approach described above, Fermanian et al. [39], Theorem 2, obtain the following informal result. Assume that f_θ is the feedfoward model, such that for any $\theta \in \Theta$, $\|U\|_F + \|V\|_F \leq \frac{1-L}{16d}$. Then, there exists a constant $B > 0$ such that for any $\theta \in \Theta$, $\|\alpha(\theta)\|_{\mathcal{T}} \leq B$, and with probability at least $1 - \delta$,

$$\mathbb{P}(\mathbf{y}g_{\widehat{\theta}_n}(\mathbf{x}) \leq 0|\mathcal{D}_n) \leq \widehat{\mathcal{R}}_n(\widehat{\theta}_n) + \frac{C_2}{T} + \frac{8BK_\ell}{(1-L)\sqrt{n}} + \frac{2BK_\ell}{1-L}\sqrt{\frac{\log(1/\delta)}{2n}}.$$

The term in $\mathcal{O}(1/T)$ comes from the continuous-time approximation, whereas the speed in $\mathcal{O}(1/\sqrt{n})$ is classical. Note that similar bounds can be obtained for the multiclass and sequence-to-sequence settings.

3.2.5. *Stability*

In addition to providing a sound theoretical framework, framing deep learning in an RKHS provides a natural norm, which can be used for regularization. Indeed, for two inputs $\mathbf{x}$ and $\mathbf{y}$, we can bound the difference between the RNN outputs $g_\theta(\mathbf{x})$ and $g_\theta(\mathbf{y})$ by using Theorem 3.4 and the Cauchy-Schwartz inequality:

$$|g_\theta(\mathbf{x}) - g_\theta(\mathbf{y})| \leq \frac{C_1}{T} + |\langle \alpha(\theta), S(\overline{X}) - S(\overline{Y})\rangle_{\mathcal{T}}| \leq \frac{C_1}{T} + \|S(\overline{X}) - S(\overline{Y})\|_{\mathcal{T}}\|\alpha(\theta)\|_{\mathcal{T}}.$$

If $\mathbf{x}$ and $\mathbf{y}$ are close, so are their embeddings $\overline{X}$ and $\overline{Y}$, and the term $\|S(\overline{X}) - S(\overline{Y})\|_{\mathcal{T}}$ is therefore small. When T is large, we see that the magnitude of $\|\alpha(\theta)\|_{\mathcal{T}}$ determines how close the predictions are. A natural training strategy to ensure stable predictions is then to penalize the problem by minimizing the loss $\widehat{\mathcal{R}}_n(\theta) + \lambda\|\alpha(\theta)\|_{\mathcal{T}}$.

3.3. **Conclusion**

To conclude, the action of a RNN can be interpreted as an element of a RKHS $\mathcal{H}$, and signatures are the key element to obtain this RKHS. By leveraging the theory of kernel methods, this point of view facilitates the analysis of generalization for a large class of RNN and gives new regularization strategies.

References

[1] S. Aida. *Reflected Rough Differential Equations.* Stoch. Proc. Appl. 125, 9, 3570–3595, 2015.
[2] A.L. Allan, C. Liu and D.J. Prömel. *Càdlàg Rough Differential Equations with Reflecting Barriers.* Stoch. Proc. Appl. 142, 79–104, 2021.
[3] J. Angst, I. Bailleul and C. Tardif. *Kinetic Brownian Motion.* Elec. J. Probab. 20, 1–40, 2015.
[4] J. Angst, I. Bailleul and P. Perruchaud. *Kinetic Brownian Motion on the Diffeomorphism Group of a Closed Riemannian Manifold.* arXiv:1905.04103, 2019.
[5] I. Bailleul. *Flows Driven by Rough Paths.* Revista Mat. Iberoamericana 31, 3, 901–934, 2015.
[6] I. Bailleul and F. Bernicot. *High Order Paracontrolled Calculus.* Forum Math. Sigma 7, e44, 1–94, 2019.
[7] I. Bailleul, A. Brault and L. Coutin. *Young and Rough Differential Inclusions.* Revista Mat. Iberoamericana 37, 4, 1489–1512, 2021.
[8] I. Bailleul, R. Catellier and F. Delarue. *Solving Mean Field Rough Differential Equations.* Elec. J. Probab. 25, 21, 1–51, 2020.
[9] I. Bailleul, R. Catellier and F. Delarue. *Propagation of Chaos for Mean Field Rough Differential Equations.* Ann. Probab. 49, 2, 944–996, 2021.
[10] C. Bayer, P. Friz, P. Gassiat, J. Martin and B. Stemper. *A Regularity Structure for Rough Volatility.* Math. Finance 30, 3, 782–832, 2020.
[11] C. Bellingeri. *Quasi-Geometric Rough Paths and Rough Change of Variable Formula.* arXiv:2009.00903, 2020.
[12] C. Bellingeri, P. Friz and M. Gerencsér. *Singular Paths Spaces and Applications.* Stochastic Analysis and Applications (accepted), 2021.
[13] C. Bellingeri, P. Friz, S. Paycha and R. Preiß. *Smooth Rough Paths, their Geometry and Algebraic Renormalization.* arXiv:2111.15539, 2021.
[14] C. Bellingeri and N. Gilliers. *The Non-Commutative Signature of a Path.* arXiv:2102.11816, 2021.
[15] P. Bonnier and H. Oberhauser. *Signature Cumulants, Ordered Partitions, and Independence of Stochastic Processes.* arXiv:1908.06496, 2019.
[16] P. Bonnier and H. Oberhauser. *Proper Scoring Rules, Gradients, Divergences, and Entropies for Paths and Time Series.* arXiv:2111.06314, 2021.
[17] Y. Bruned, M. Hairer and L. Zambotti. *Algebraic Renormalisation of Regularity Structures.* Invent. Math. 215, 3, 1039–1156, 2019.
[18] Y. Bruned and D. Manchon. *Algebraic Deformation for (S)PDEs.* arXiv:2011.05907, 2020.
[19] Y. Bruned and K. Schratz. *Resonance Based Schemes for Dispersive Equations via Decorated Trees.* Forum Math. Pi 10, e2, 1–76, 2022.
[20] J.C. Butcher. *An Algebraic Theory of Integration Methods.* Math. Comp. 26, 79–106, 1972.
[21] T. Cass, C. Litterer and T. Lyons. *Integrability and Tail Estimates for Gaussian Rough Differential Equations.* Ann. Probab. 41, 4, 3026–3050, 2013.
[22] T. Cass and T. Lyons. *Evolving Communities with Individual Preferences.* Proc. London Math. Soc. 110, 1, 83–107, 2015.
[23] T. Cass and P. Friz. *Densities for Rough Differential Equations under Hoermander's Condition.* Annals of Math. 171, 2115–2141, 2010.
[24] T. Cass, M. Hairer, C. Litterer and S. Tindel. *Smoothness of the Density for Solutions to Gaussian Rough Differential Equations.* Ann. Probab. 43, 1, 188–239, 2015.
[25] C. Castaing, N. Marie and P. Raynaud de Fitte. *Sweeping Processes Perturbed by Rough Signals.* Séminaire de Probabilités LI, Lecture Notes in Mathematics, Springer, 2022.
[26] K.T. Chen. *Iterated Integrals and Exponential Homomorphisms.* Proc. London Math. Soc. 3, 4, 502–512, 1954.
[27] K.T. Chen. *Integration of Paths, Geometric Invariants and a Generalized Baker-Hausdorff Formula.* Annals of Math. 2, 65, 163–178, 1957.
[28] K.T. Chen. *Integration of Paths – a Faithful Representation of Paths by Non-Commutative Formal Power Series.* Transactions of the American Mathematical Society 89, 395–407, 1958.
[29] K.T. Chen. *Iterated Path Integrals and Generalized Paths.* Bull. Amer. Math. Soc. 73, 935–938, 1967.

[30] K.T. Chen. *Algebraization of Iterated Integration Along Paths.* Bull. Amer. Math. Soc. 73, 975–978, 1967.
[31] R.T.Q. Chen, Y. Rubanova, J. Bettencourt, and D.K. Duvenaud. *Neural Ordinary Differential Equations.* In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 6572–6583. Curran Associates, Inc., 2018.
[32] I. Chevyrev, P. Friz, A. Korepanov, I. Melbourne and H. Zhang. *Multiscale Systems, Homogenization, and Rough Paths.* arXiv:1712.01343, 2017.
[33] A. Connes and D. Kreimer. *Hopf Algebras, Renormalization and Noncommutative Geometry.* Comm. Math. Phys. 199, 1, 203–242, 1998.
[34] L. Coutin and N. Marie. *Invariance for Rough Differential Equations.* Stoch. Proc. Appl. 127, 7, 2373–2395, 2017.
[35] A.M. Davie. *Differential Equations Driven by Rough Paths: An Approach via Discrete Approximation.* Appl. Math. Res. Express 2, 9–40, 2007.
[36] A. Deya and R. Schott. *On the Rough Paths Approach to Non- Commutative Stochastic Calculus.* J. Funct. Anal. 265, 4, 594–628, 2013.
[37] A. Deya and S. Tindel. *Rough Volterra Equations 1: The Algebraic Integration Setting.* Stochastics and Dynamics 9, 3, 437–477, 2009.
[38] A. Deya and S. Tindel. *Rough Volterra Equations 2: Convolutional Generalized Integrals.* Stoch. Proc. Appl. 121, 8, 1864–1899, 2011.
[39] A. Fermanian, P. Marion, J-P. Vert and G. Biau. *Framing RNN as a Kernel Method: A Neural ODE Approach.* In A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
[40] D. Feyel and A. de la Pradelle. *Curvilinear Integrals Along Enriched Paths.* Elec. J. Probab. 34, 860–892, 2006.
[41] P. Friz and P. Gassiat. Geometric Foundations of Rough Paths. In *Geometry, Analysis and Dynamics on sub-Riemannian Manifolds*, pp. 171–210, European Mathematical Society Publishing House, 2016.
[42] P. Friz, P. Gassiat and T. Lyons. *Physical Brownian Motion in Magnetic Field as Rough Path.* Trans. AMS 367, 11, 7939–7955, 2015.
[43] P. Friz and M. Hairer. *A Course on Rough Paths with an Introduction to Regularity Structures.* Universitext, Springer, Cham, 2014.
[44] P. Friz and H. Oberhauser. *A Generalized Fernique Theorem and Applications.* Proc. AMS 138, 10, 3679–3688, 2010.
[45] P. Friz and N. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications.* Cambridge Studies in Applied Mathematics 120, Cambridge University Press, Cambridge, 2010.
[46] P. Gassiat. *Non-Uniqueness for Reflected Rough Differential Equations.* Ann. Inst. H. Poincaré, Probab. Stat. 57, 3, 1369–1387, 2021.
[47] J. Gehringer and X.M. Li. *Slow-Fast Systems with Fractional Environment and Dynamics.* arXiv:2011.00075, 2020.
[48] M. Gubinelli. *Controlling Rough Paths.* J. Funct. Anal. 216, 86–140, 2004.
[49] M. Gubinelli. *Ramification of Rough Paths.* J. Diff. Eq. 248, 693–721, 2010.
[50] M. Gubinelli, P. Imkeller and N. Perkowski. *Paracontrolled Distributions and Singular PDEs.* Forum Math. Pi 3, e6, 1–75, 2015.
[51] M. Hairer. *A Theory of Regularity Structures.* Invent. Math. 198, 2, 269–504, 2014.
[52] M. Hairer and D. Kelly. *Geometric Versus Non-Geometric Rough Paths.* Ann. Inst. H. Poincaré, Probab. Stat. 51, 1, 207–251, 2015.
[53] M. Hairer and X.M. Li. *Generating Diffusions with Fractional Brownian Motion.* arXiv:2109.06948, 2021.
[54] F.A. Harang and S. Tindel. *Volterra Equations Driven by Rough Signals.* Stoch. Proc. Appl. 142, 34–78, 2021.
[55] F. A. Harang, S. Tindel and X. Wang. *Volterra Equations Driven by Rough Signals 2: Higher Order Expansions.* arXiv:2102.10119, 2021.
[56] M. Hoffman. *Quasi-Shuffle Products.* J. Algebraic Combin. 11, 1, 49–68, 2000.
[57] K. Itô. *Stochastic Integral.* Proc. Imp. Acad. Tokyo 20, 519–524, 1944.
[58] K. Itô. *Stochastic Differential Equations.* Memoirs AMS 4, 1951.
[59] J. Jacod. *Calcul stochastique et problèmes de martingales.* Lecture Notes in Mathematics 714, Springer, 1979.
[60] P. Kidger, J. Morrill, J. Foster and T. Lyons. *Neural Controlled Differential Equations for Irregular Time Series.* In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707, Curran Associates, Inc., 2020.
[61] J. Király and H. Oberhauser. *Kernels for Sequentially Ordered Data.* Journal of Machine Learning Research 20, 1–45, 2019.
[62] D. Kelly and I. Melbourne. *Smooth Approximation of Stochastic Differential Equations.* Ann. Probab. 44, 1, 479–520, 2016.
[63] D. Kelly and I. Melbourne. *Deterministic Homogenization for Fast-Slow Systems with Chaotic Noise.* arXiv:1409.5748, 2014.
[64] T. Lyons. *Differential Equations Driven by Rough Signals (I): An Extension of an Inequality of L.C. Young.* Mathematical Research Letters 1, 451–464, 1994.
[65] K. Lê. *A Stochastic Sewing Lemma and Applications.* Elec. J. Probab. 25, 1–55, 2020.
[66] D. Levin, T. Lyons and H. Ni. *Learning from the Past, Predicting the Statistics for the Future, Learning an Evolving System.* arXiv:1309.0260, 2013.

[67] S. Liao, T. Lyons, W. Yang and H. Ni. *Learning Stochastic Differential Equations Using RNN with Log Signature Features.* arXiv:1908.08286, 2019.
[68] T. Lyons. *Differential Equations Driven by Rough Signals.* Rev. Mat. Iberoamericana 14, 2, 215–310, 1998.
[69] T. Lyons, M. Caruana and T. Lévy. *Differential Equations Driven by Rough Paths.* Volume 1908 of Lecture Notes in Mathematics. Springer, Berlin, 2007.
[70] J. H. Morrill, A. Kormilitzin, A. J. Nevado-Holgado, S. Swaminathan, S. D. Howison and T. J. Lyons. *Utilization of the Signature Method to Identify the Early Onset of Sepsis from Multivariate Physiological Time Series in Critical Care Monitoring.* Critical Care Medicine 48, e976–e981, 2020.
[71] D. Prömel and M. Trabs. *Paracontrolled Distribution Approach to Stochastic Volterra Equations.* J. Diff. Eq. 302, 222–272, 2021.
[72] C. Reutenauer. *Free Lie Algebras.* LMS Monographs. Clarendon Press, 1993.
[73] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, Massachusetts, 2002.
[74] W. Yang, L. Jin and M. Liu. *DeepWriterID: An End-to-End Online Text-Independent Writer Identification System.* IEEE Intelligent Systems 31, 45–53, 2016.
[75] L.C. Young. *An Inequality of Hölder Type Connected with Stieljès Integration.* Acta Math. 67, 251–282, 1936.