

Rough paths, Signatures and the modelling of functions on streams

Terry Lyons*

Abstract. Rough path theory is focused on capturing and making precise the interactions between highly oscillatory and non-linear systems. The techniques draw particularly on the analysis of LC Young and the geometric algebra of KT Chen. The concepts and theorems, and the uniform estimates, have found widespread application; the first applications gave simplified proofs of basic questions from the large deviation theory and substantially extending Ito's theory of SDEs; the recent applications contribute to (Graham) automated recognition of Chinese handwriting and (Hairer) formulation of appropriate SPDEs to model randomly evolving interfaces. At the heart of the mathematics is the challenge of describing a smooth but potentially highly oscillatory and vector valued path x_t parsimoniously so as to effectively predict the response of a nonlinear system such as $dy_t = f(y_t)dx_t$, $y_0 = a$. The Signature is a homomorphism from the monoid of paths into the grouplike elements of a closed tensor algebra. It provides a graduated summary of the path x . Hambly and Lyons have shown that this non-commutative transform is faithful for paths of bounded variation up to appropriate null modifications. Among paths of bounded variation with given Signature there is always a unique shortest representative. These graduated summaries or features of a path are at the heart of the definition of a rough path; locally they remove the need to look at the fine structure of the path. Taylor's theorem explains how any smooth function can, locally, be expressed as a linear combination of certain special functions (monomials based at that point). Coordinate iterated integrals form a more subtle algebra of features that can describe a stream or path in an analogous way; they allow a definition of rough path and a natural linear "basis" for functions on streams that can be used for machine learning.

Mathematics Subject Classification (2010). Primary 00A05; Secondary 00B10.

Keywords. Rough paths, Regularity Structures, Machine Learning, Functional Regression, Numerical Approximation of Parabolic PDE, Shuffle Product, Tensor Algebra

*Acknowledges the support of the Oxford-Man Institute, the support provided by ERC advanced grant ESig (agreement no. 291244), and particularly the contributions of his colleagues and his students without whom none of this would have happened and in addition to Kelly Wyatt, Justin Sharp, Horatio Boedihardjo, Hao Ni and Danyu Yang for helping the author finalise this mss. The data analysis is reproduced from the cited paper with Gyurko et al., Gyurko did the analysis, The raw data for that paper is available on Reuters.

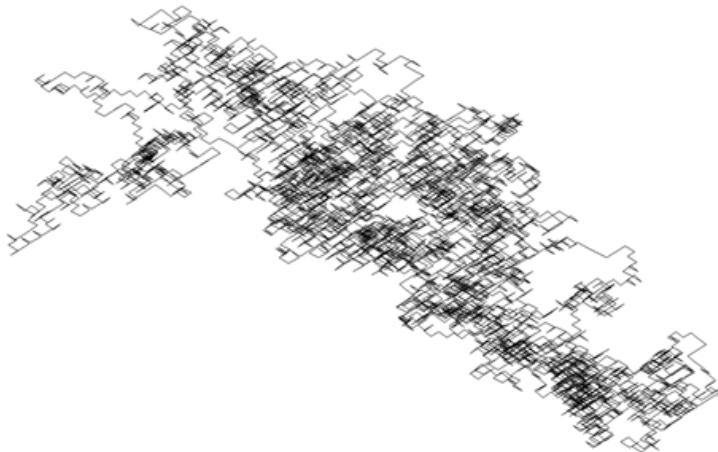
Contents

1 A path or a text?	3
2 Financial Data or a Semimartingale	4
3 Paths - Simply Everywhere - Evolving systems	5
4 A simple model for an interacting system	5
5 Remarkable Estimates (for $p > 1$)	8
6 The Log Signature	10
7 The ODE method	11
8 Going to Rough Paths	12
9 Coordinate Iterated Integrals	14
10 Expected Signature	15
11 Computing expected Signatures	15
12 Characteristic Functions of Signatures	16
13 Moments are complicated	17
14 Regression onto a feature set	17
15 The obvious feature set for streams	19
16 Machine learning, an amateur's first attempt	19
17 Linear regression onto a law on paths	22

1. A path or a text?

The mathematical concept of a path embraces the notion of an evolving or time ordered sequence of events, parameterised by a continuous variable. Our mathematical study of these objects does not encourage us to think broadly about the truly enormous range of "paths" that occur. This talk will take an analyst's perspective, we do not expect to study a particular path but rather to find broad brush tools that allow us to study a wide variety of paths - ranging from very "pure" mathematical objects that capture holonomy to very concrete paths that describe financial data. Our goal will be to explain the progress we have made in the last 50 years or so in describing such paths effectively, and some of the consequences of these developments.

Let us start by noting that although most mathematicians would agree on a definition of a path, most have a rather stereotyped and limited imagination about the variety of paths that are "in the wild". One key observation is that in most cases we are interested in paths because they represent some evolution that interacts with and influences some wider system. Another is that in most paths, in standard presentations, the content and influence are locked into complex multidimensional oscillations.



The path in the figure is a piece of text. Each character in the text is encoded using ascii as a byte of 8 bits, each byte is represented as four letters of two bits, each two bit letter is represented by a line from the centre to one of the four corners of a square (for visual reasons the centre of this square is displaced slightly to create a loop). The text can easily be represented in other ways, perhaps in different font or with each character as a bitmap. Each stream has broadly the same effect on a coarse scale although the detailed texture is perhaps a bit different.

2. Financial Data or a Semimartingale

One important source of sequential data comes from financial markets. An intrinsic feature of financial markets is that they are high dimensional but there is a strong notion of sequencing of events. Buying with future knowledge is forbidden. Much of the information relates to prices, and one of the radical successes of applied mathematics over the last 20-30 years came out of the approximation of price processes by simple stochastic differential equations and semimartingales and the use of Itô's calculus. However, modern markets are not represented by simple price processes. Most orders happen on exchanges, where there are numerous bids, offers, and less commonly, trades. Much activity in markets is concerned with market making and the provision of liquidity; decisions to post to the market are based closely on expectation of patterns of behaviour, and most decisions are somewhat distant from any view about fundamental value. If one is interested in alerting the trader who has a bug in his code, or understanding how to trade a large order without excessive charges then the semi-martingale model has a misplaced focus.

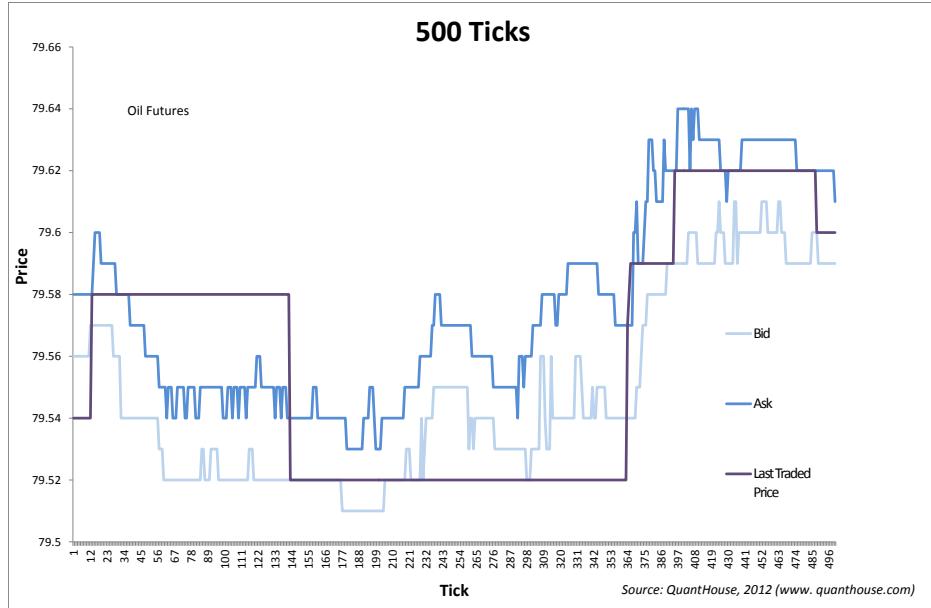


Figure 1. A snapshot of level one order book data

The data in the figure 1 is a snapshot of the level one order book showing activity on a market for oil futures over 500 changes (roughly a 15 minute period). One can see the bid and offer prices changing, although trades happen (and so the last executed price changes) much less frequently. It is questionable whether a semi-martingale model for prices can capture this rich structure effectively.

3. Paths - Simply Everywhere - Evolving systems

Informally, a stream is a map γ from a totally ordered set I to some state space, where we are interested in the effect (or transformation of state) this stream achieves. As we have noted the same stream of information can admit different representations with different fidelity. When the totally ordered set I is an interval and there are reasonable path properties (e.g. such as right continuity) we will call the stream a path. Nonetheless, many interesting streams are finite and discrete. There are canonical and informative ways to convert them [10] to continuous paths.

It is worth noting that, even at this abstract level, there are natural mathematical operations and invariances that are applied to a stream. One can reparameterise the speed at which one examines the stream and simultaneously the speed at which one looks at the effects. One can split a stream into two or more segments (a coproduct). One can sub-sample a stream. In general we will focus on those streams which are presented in a way where such sub-sampling degrades the information in the stream gradually. One can also merge or interleave discrete streams according to their time stamps if the totally ordered sets I, I' can be interleaved. All of these properties are inherited for the properties of totally ordered sets. If the target "effect" or state space is linear there is also the opportunity to translate and so concatenate streams or paths [15] and so get richer algebraic structures. One of the most interesting and economically important questions one can ask about a stream is how to summarise (throw away irrelevant information) so as to succinctly capture its effects. We give a few examples in Table 1.

text	schoolchild	precis
sound	audio engineer	faithful perception
web page	search provider	interest for reader
web click history	advertiser	effective ad placement
Brownian path	numerical analysis	effective simulation
rough paths	analyst	RDEs

Table 1. Examples of contexts where streams are summarised while retaining their essence.

What is actually quite surprising is that there is a certain amount of useful work one can do on this problem that does not depend on the nature of the stream or path.

4. A simple model for an interacting system

We now focus on a very specific framework where the streams are maps from a real interval, that we will intuitively refer to as the time domain, into an a Banach space that we will refer to as the state space. We will work with continuous paths in continuous time but, as we mentioned, there are canonical ways to embed discrete

tick style data into this framework using the Hoff process and in financial contexts this is important. There is also a more general theory dealing with paths with jumps [Williams, Simon].

4.1. Controlled Differential Equations. A path is a map γ from an interval $J = [J_-, J_+]$ into a Banach space E . The dimension of E may well be finite, but we allow for the possibility that it is not. It has bounded (p -)variation if

$$\begin{aligned} \sup_{\dots u_i < u_{i+1} \dots \in [J_-, J_+]} \sum_i \|\gamma_{u_{i+1}} - \gamma_u\| &< \infty \\ \sup_{\dots u_i < u_{i+1} \dots \in [J_-, J_+]} \sum_i \|\gamma_{u_{i+1}} - \gamma_u\|^p &< \infty \end{aligned}$$

where $p \geq 1$. In our context the path γ is controlling the system, and we are interested in its effect as measured by y and the interactions between γ and y . It would be possible to use the theory of rough paths to deal with the internal interactions of autonomous and "rough" systems, one specific example of deterministic McKean Vlasov type is [4].

Separately there needs to be a space F that carries the state of the system and a family of different ways to evolve. We represent the dynamics on F through the space $\Omega(F)$ of vector fields on F . Each vector field provides a different way for the state to evolve. We connect this potential to evolve the state in F to the control γ via a linear map

$$V : E \xrightarrow{\text{linear}} \Omega(F).$$

Immediately we can see the controlled differential equation

$$\begin{aligned} dy_t &= V(y_t) d\gamma_t, \quad y_{J_-} = a \\ \pi_J(y_{J_-}) &: = y_{J_+} \end{aligned}$$

provides a precise framework allowing for the system y to respond to γ according to the dynamics V . We call such a system a controlled differential equation.

The model of a controlled differential equation is a good one. Many different types of object can be positioned to fit the definition. Apart from the more obvious applied examples, one can view a finite automata (in computer science sense) and the geometric concept of lifting a path along a connection as producing examples.

There are certain apparently trivial properties that controlled differential equations and the paths that control them have; none the less they are structurally essential so we mention them now.

Lemma 4.1 (Reparameterisation). If $\tau : I \rightarrow J$ is an increasing homeomorphism, and if

$$dy_t = V(y_t) d\gamma_t, \quad y_{J_-} = a,$$

then the reparameterised control produces the reparameterised effect:

$$dy_{\tau(t)} = V(y_{\tau(t)}) d\gamma_{\tau(t)}, \quad y_{\tau(I_-)} = a.$$

Lemma 4.2 (Splitting). Let π_J be the diffeomorphism capturing the transformational effect of $\gamma|_J$. Let $t \in J$. Then π_J can be recovered by composing the diffeomorphisms $\pi_{[J_-, t]}$, $\pi_{[t, J_+]}$ associated with splitting the interval J at t and considering the composing the effect of $\gamma|_{[J_-, t]}$ and $\gamma|_{[t, J_+]}$ separately:

$$\pi_{[t, J_+]} \pi_{[J_-, t]} = \pi_J.$$

In this way we see that, assuming the vector fields were smooth enough to solve the differential equations uniquely and for all time, a controlled differential equation is a homomorphism from the monoid of paths with concatenation into the diffeomorphisms/transformations of the state space. By letting π act as an operator on functions we see that every choice of V defines a representation of the monoid of paths in E

Remark 4.3 (Subsampling). Although there is a good behaviour with respect to sub-sampling, which in effect captures and quantifies the numerical analysis of these equations, it is more subtle and we do not make it explicit here.

Remark 4.4. Fixing V , restricting γ to smooth paths on $[0, 1]$ and considering the solutions y with $y_0 = a$, generically the closure of the set of pairs (γ, y) in the uniform topology is NOT the graph of a map; $\gamma \rightarrow y$ is not closable and so is not well defined as a (even an unbounded and discontinuous) function in the space of continuous paths. Different approximations lead to different views as to what the solution should be.

4.2. Linear Controlled Differential Equations. Where the control γ is fixed and smooth, the state space is linear, and all the vector fields are linear, then the space of responses y , as one varies the starting location a , is a linear space and $\pi_{[S, T]} : a = y_S \rightarrow y_T$ is a linear automorphism. This case is essentially Cartan's development of a path in a Lie Algebra into a path in the Lie Group starting at the identity. From our point of view it is a very important special case of our controlled differential equations; it reveals one of the key objects we want to discuss in this paper.

Suppose F is a Banach space, and A is a linear map $E \rightarrow \text{Hom}_{\mathbb{R}}(F, F)$ and that γ_t is a path in E . Consider the linear differential equation

$$dy_t = A y_t d\gamma_t.$$

By iterating using Picard iteration one obtains

$$y_{J_+} = \left(\sum_{n=0}^{\infty} A^n \int_{J_- \leq u_1 \leq \dots \leq u_n \leq J_+} \dots \int d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} \right) y_0$$

The Signature of γ over the interval $J = [J_-, J_+]$

Definition 4.5. The Signature S of a bounded variation path (or more generally a weakly geometric p -rough path) γ over the interval $J = [J_-, J_+]$ is the tensor

sequence

$$S(\gamma|_J) := \sum_{n=0}^{\infty} \int \cdots \int_{u_1 \leq \dots \leq u_n \in J^n} d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} \in \bigoplus_{n=0}^{\infty} E^{\otimes n}$$

It is sometimes written $S(\gamma)_J$ or $S(\gamma)_{J_-, J_+}$.

Lemma 4.6. The path $t \rightarrow S(\gamma)_{0,t}$ solves a linear differential equation controlled by γ .

Proof. The equation is the universal non-commutative exponential:

$$\begin{aligned} dS_{0,t} &= S_{0,t} \otimes d\gamma_t \\ S_{0,0} &= 1 \end{aligned}$$

□

The solution to any linear equation is easily expressed in terms of the Signature

$$\begin{aligned} dy_t &= Ay_t d\gamma_t \\ y_{J_+} &= \left(\sum_0^{\infty} A^n S_J^n \right) y_{J_-} \\ \pi_J &= \sum_0^{\infty} A^n S_J^n \end{aligned} \tag{1}$$

and we will see in the next sections that this series converges very well and even the first few terms in S are effective in describing the response y_T leading to the view that $\gamma|_J \rightarrow S(\gamma|_J)$ is a transform with some value. The use of S to describe solutions to linear controlled differential equations goes back at least to Chen, and Feynman. The *magic* is that one can estimate the errors in convergence of the series (1) without detailed understanding of γ or A .

5. Remarkable Estimates (for $p > 1$)

It seems strange, and even counter intuitive, that one should be able to identify and abstract a finite sequence of features or coefficients describing γ adequately so that its effect on a broad range of different systems could be accurately predicted without detailed knowledge of the system A or the path γ - beyond those few coefficients. But that is the truth of it, there are easy uniform estimates capturing the convergence of the series (1) based entirely on the length (or more generally p -rough path variation) of the control and the norm of A as a map from E to the linear vector fields on F .

Lemma 5.1. If γ is a path of finite variation on J with length $|\gamma_J| < \infty$, then

$$\begin{aligned} S_J^n & : = \int_{u_1 \leq \dots \leq u_n \in J^n} \cdots \int d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} \\ & \leq \frac{|\gamma_J|^n}{n!} \end{aligned}$$

giving uniform error control

$$\left\| y_{J_+} - \sum_0^{N-1} A^n \int_{J_- \leq u_1 \leq \dots \leq u_n \leq J_+} \cdots \int d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} y_0 \right\| \leq \left(\sum_{n=N}^{\infty} \frac{\|A\|^n |\gamma_J|^n}{n!} \right) \|y_0\|.$$

Proof. Because the Signature of the path always solves the characteristic differential equation it follows that one can reparameterise the path γ without changing the Signature of γ . Reparameterise γ so that it is defined on an interval J of length $|\gamma|$ and runs at unit speed. Now there are $n!$ disjoint simplexes inside a cube obtained by different permuted rankings of the coordinates and thus

$$\begin{aligned} \|S_J^n\| & : = \left\| \int_{u_1 \leq \dots \leq u_n \in J^n} \cdots \int d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} \right\| \\ & = \left\| \int_{u_1 \leq \dots \leq u_n \in J^n} \dot{\gamma}_{u_1} \otimes \dots \otimes \dot{\gamma}_{u_n} du_1 \dots du_n \right\| \\ & = \int_{u_1 \leq \dots \leq u_n \in J^n} \|\dot{\gamma}_{u_1} \otimes \dots \otimes \dot{\gamma}_{u_n}\| du_1 \dots du_n \\ & = \int_{u_1 \leq \dots \leq u_n \in J^n} du_1 \dots du_n \\ & = \frac{|\gamma_J|^n}{n!}. \end{aligned}$$

from which the second estimate is clear. \square

The Poisson approximation of a normal distribution one learns at high school ensures that the estimates on the right become very sharply estimated in terms of $\lambda \rightarrow \infty$ and pretty effective as soon as $N \geq \|A\| |\gamma_J| + \lambda \sqrt{\|A\| |\gamma_J|}$.

Remark 5.2. The uniform convergence of the series

$$\sum_{n=0}^{N-1} A^n \int_{J_- \leq u_1 \leq \dots \leq u_n \leq J_+} \cdots \int d\gamma_{u_1} \otimes \dots \otimes d\gamma_{u_n} y_0$$

and the obvious continuity of the terms of the series in the inputs (A, γ, y_0) guarantees that the response y_T is jointly continuous (uniform limits of continuous

functions are continuous) in (A, γ, y_0) where γ is given the topology of 1-variation (or any of the rough path metrics). It is already the case that

$$\gamma \rightarrow \int_{J_-}^{\gamma} \cdots \int_{J_+}^{\gamma} d\gamma_{u_1} \otimes d\gamma_{u_2}$$

fails the closed graph property in the uniform metric.

6. The Log Signature

It is easy to see that the Signature of a path segment actually takes its values in a very special curved subspace of the tensor algebra. Indeed, Chen noted that the map S is a homomorphism of path segments with concatenation into the algebra, and reversing the path segment produces the inverse tensor. As a result one sees that the range of the map is closed under multiplication and has inverses so it is a group (inside the grouplike elements) in the tensor series. It is helpful to think of the range of this Signature map as a curved space in the tensor series. As a result there is a lot of valuable structure. One important map is the logarithm; it is one to one on the group and provides a flat parameterisation of the group in terms of elements of the free Lie series.

Definition 6.1. If $\gamma_t \in E$ is a path segment and S is its Signature then

$$\begin{aligned} S &= 1 + S^1 + S^2 + \dots \quad \forall i, \quad S^i \in E^{\otimes i} \\ \log(1+x) &= x - x^2/2 + \dots \\ \log S &= (S^1 + S^2 + \dots) - (S^1 + S^2 + \dots)^2/2 + \dots \end{aligned}$$

The series $\log S = (S^1 + S^2 + \dots) - (S^1 + S^2 + \dots)^2/2 + \dots$ which is well defined, is referred to as the log Signature of γ .

Because the space of tensor series $T((E)) := \bigoplus_0^\infty E^{\otimes n}$ is a unital associative algebra under $\otimes, +$ it is also a Lie algebra, and with $[A, B] := A \otimes B - B \otimes A$.

Definition 6.2. There are several canonical Lie algebras associated to $T((E))$; we use the notation $\mathcal{L}(E)$ for the algebra generated by E (the space of Lie polynomials), $\mathcal{L}^{(n)}(E)$ the projection of this into $T^{(n)}(E) = T((E))/\bigoplus_{n+1}^\infty E^{\otimes m}$ (the Lie algebra of the free nilpotent group G^n of n steps) and $\mathcal{L}((E))$ the projective limit of the $\mathcal{L}^{(n)}(E)$ (the Lie Series).

Because we are working in characteristic zero, we may take the exponential, and this recovers the Signature, so no information is lost. A key observation of Chen [6] was that if γ is a path segment then $\log S(\gamma) \in \mathcal{L}((E))$. The map from paths [23, 8] to $\mathcal{L}^{(n)}(E)$ via the projection $\pi_n : T((E)) \rightarrow T^{(n)}(E)$ is onto. Up to equivalence under a generalised notion of reparameterisation of paths known as treelike equivalence, the map from paths γ of finite length in E to their Signatures $S(\gamma) \in T((E))$ or log-Signatures $\log S \in \mathcal{L}((E))$ is injective [15]. Treelike

equivalence is an equivalence relation on paths of finite variation, each class has a unique shortest element, and these tree reduced paths form a group. However the range of the log-Signature map in $\mathcal{L}((E))$, although well behaved under integer multiplication is not closed under integer division [21] and so the Lie algebra of the group of tree reduced paths is well defined but not a linear space; it is altogether a more subtle object.

Implicit in the definition of a controlled differential equation

$$dy_t = f(y_t) d\gamma_t, \quad y_0 = a$$

is the map f . This object takes an element $e \in E$ and an element $y \in F$ and produces a second vector in F , representing the infinitesimal change to the state y of the system that will occur if γ is changed infinitesimally in the direction e . This author is clear that the best way to think about f is as a linear map from the space E into the vector fields on F . In this way one can see that the integral of f along γ in its simplest form is a path in the Lie algebra and that in solving the differential equation we are developing that path into the group. Now, at least formally, the vector fields are a Lie algebra (for the diffeomorphisms of F) and subject to the smoothness assumptions we can take Lie brackets to get new vector fields. Because $\mathcal{L}((E))$ is the free Lie algebra over E (Chapter II, [2]) any linear map f of E into a Lie algebra \mathfrak{g} induces a unique Lie map extension f_* to a Lie map from $\mathcal{L}((E))$ to \mathfrak{g} . This map can be readily implemented and is well defined because of the abstract theory

$$\begin{aligned} e &\rightarrow f(e) \quad \text{a vector field} \\ e_1 e_2 - e_2 e_1 &\rightarrow f(e_1) f(e_2) - f(e_2) f(e_1) \quad \text{a vector field} \\ \tilde{f} &: \mathcal{L}^{(n)}(E) \rightarrow \text{vector fields}. \end{aligned}$$

although in practice one does not take the map to the full projective limit.

7. The ODE method

The linkage between truncations of the log-Signature in $\mathcal{L}((E))$ and vector fields on Y is a practical one for modelling and understanding controlled differential equations. It goes well beyond theory and underpins some of the most effective and stable numerical approaches (and control mechanisms) for translating the information in the control γ into information about the response.

If $dy_t = f(y_t) d\gamma_t$, and $y_{J_-} = a$ then how can we use the first few terms of the (log-)Signature of γ to provide a good approximation to y_{J_+} ? We could use picard iteration, or better an euler method based on a Taylor series in terms of the Signatures. Picard iteration for $\exp z$ already illustrates one issue. Picard interation yields a power series as approximation - fine if $z = 100$,but awful if $x = -100$. However, there is a more subtle problem to do with stability that almost all methods based on Taylor series have - stability - they can easily produce approximations that are not feasible. These are aggravated in the controlled case

because of the time varying nature of the systems. It can easily happen that the solutions to the vector fields are hamiltonian etc. The ODE method uses the first few terms of the Signature to construct a time invariant ODE (vector field) that if one solves it for unit time, it provides an approximation to the desired solution. It pushes the numerics back onto state of the art ODE solvers. Providing the ODE solver is accurate and stable then the approximation to y will also be. One can use symplectic solvers etc. At the level of rough paths, the approximation is obtained by replacing the path γ with a new rough path $\hat{\gamma}$ (a geodesic in the nilpotent group G^n) with the same first few terms in the Signature; this guarantees the feasibility of the approximations. Today, rough path theory can be used to estimate the difference between the solution and the approximation in terms of the distance between γ and $\hat{\gamma}$ even in infinite dimensions.[5][3]

Remark 7.1. A practical numerical scheme can be built as follows.

1. Describe γ over a short interval J in terms of first few terms of $\log S(\gamma_{[J_-, J_+]})$ expressed as a linear combination of terms of a fixed hall basis:

$$\begin{aligned}\log S_J &= l^1 + l^2 + \dots \in \mathcal{L}((E)) \\ l^{(n)} &= \pi_n(\log S_J) = l^1 + \dots + l^n \in \mathcal{L}^{(n)}(E) \\ l^1 &= \sum_i \lambda_i e_i \\ l^2 &= \sum_{i < j} \lambda_{ij} [e_i, e_j], \\ &\dots\end{aligned}$$

and use this information to produce a path dependent vector field $V = \tilde{f}(l^{(n)})$.

2. Use an appropriate ODE solver to solve the ODE $\dot{x}_t = V(x_t)$, where $x_0 = y_{J_-}$. A stable high order approximation to y_{J_+} is given by x_{J_+} .
3. Repeat over small enough time steps for the high order approximations to be effective.
4. The method is high order, stable, and corresponding to replacing γ with a piecewise geodesic path on successively finer scales.

8. Going to Rough Paths

As this is a survey, we have deliberately let the words rough path enter the text before they are introduced more formally. Rough path theory answers the following question. Suppose that γ is a smooth path but still on normal scales, a highly rough and oscillatory path. Suppose that we have some smooth system f . Give a simple metric on paths γ and a continuity estimate that ensures that if two paths that are close in this metric then their responses are quantifiably close as well. The

estimate should only depend on f through its smoothness. There is such a theory [20], and a family of rough path metrics which make the function $\gamma \rightarrow y$ uniformly continuous. The completion of the smooth paths γ under these metrics are the rough paths we speak about. The theory extends to an infinite dimensional one and the estimates are uniform in a way that does not depend on dimension.

There are many sources for this information on rough paths for different kinds of audience and we do not repeat that material. We have mentioned that two smooth paths have quantifiable close responses to a smooth f over a fixed time interval if the first terms in the Signature agree over this time interval. We can build this into a metric:

$$d_p(\gamma|_J, \hat{\gamma}|_J) = \sup_{J_- \leq u_1 \leq \dots \leq u_n \leq J_+} \sum_i \max_{m \leq [p]} \|S^m(\gamma|_{[u_i, u_{i+1}]}) - S^m(\hat{\gamma}|_{[u_i, u_{i+1}]})\|^{p/m}$$

and providing the system is $Lip(p + \varepsilon)$ the response will behave uniformly with the control. The completion of the piecewise smooth paths under d_p are p -variation paths. They do not have smoothness but they do have a "top down" description and can be viewed as living in a $[p]$ -step nilpotent group over E .

It is worth distinguishing the Kolmogorov and the rough path view on paths. In the former, one considers fixed times t_i , open sets O_i , and considers the probability that for all i , $x_{t_i} \in O_i$. In other words the emphasis is on where the path is at given times. This gated description will never capture the rough path; parameterisation is irrelevant but increments over small intervals $[u_i, u_{i+1}]$, are critical. More accurately one describes a path through an examination of the effect of it's path segment into a simple nonlinear system (the lift onto a nilpotent group). Knowing this information in an analytically adequate way is all one needs to know to predict the effect of the path on a general system.

The whole rough path theory is very substantial and we cannot survey it adequately here. The range is wide, and is related to any situation where one has a family of non-commuting operators and one wants to do analysis on apparently divergent products and for example it is interesting to understand the paths one gets as partial integrals of complex Fourier transform as the nonlinear Fourier transform is a differential equation driven by this path. Some results have been obtained in this direction [22] while the generalisations to spatial contexts are so huge that they are spoken about elsewhere at this congress. Many books are now written on the subject [11].and new lecture notes by Friz are to appear soon with recent developments. So in what is left of this paper we will focus on one topic the Signature of a path and the expected Signature of the path with a view to partially explaining how it is really an extension of Taylor's theorem to various infinite dimensional groups, and how we can get practical traction from this perspective. One key point we will not mention is that using Taylor's theorem twice works! This is actually a key point that the whole rough path story depends on and which validates its use. One needs to read the proofs to understand this adequately and, except for this sentence, suppress it completely here.

9. Coordinate Iterated Integrals

In this short paper we have to have a focus, and as a result we cannot explore the analysis and algebra needed to fully describe rough paths or to discuss the spatial generalisations directly even though they are having great impact[14][13]. Nonetheless much of what we say can be thought of as useful foundations for this work. We are going to focus on the Signature as a tool for understanding paths and as a new tool to help with machine learning.

The essential remark may seem a bit daunting to an analyst, but will be standard to others. *The dual of the enveloping algebra of a group(like) object has a natural abelian product structure and linearises polynomial functions on a group.* This fact allows one to use linear techniques on the linear spaces to approximate generic smooth (and nonlinear) functions on the group. Here the group is the "group" of paths.

Monomials are special functions on \mathbb{R}^n , and polynomials are linear combinations of these monomials. Because monomials span an algebra, the polynomials are able to approximate any continuous function on a compact set. Coordinate iterated integrals are linear functionals on the tensor algebra and at the same time they are the monomials or the features on path space.

Definition 9.1. Let $e = e_1 \otimes \dots \otimes e_n \in (E^*)^{\otimes n} \subset T(E^*)$, and $\phi_e(\gamma) := \langle e, S(\gamma) \rangle$ then we call $\phi_e(\gamma)$ a coordinate iterated integral.

Remark 9.2. Note that $S(\gamma) \in T((E)) = \bigoplus_0^\infty E^{\otimes n}$ and

$$\begin{aligned}\phi_e(\gamma) &= \langle e, S(\gamma) \rangle \\ &= \int_{u_1 \leq \dots \leq u_n \in J^n} \dots \int \langle e_1, d\gamma_{u_1} \rangle \dots \langle e_n, d\gamma_{u_n} \rangle\end{aligned}$$

justifying the name. ϕ_e is a real valued function on Signatures of paths.

Lemma 9.3. The shuffle product Π on $T(E^*)$ makes $T(E^*)$ a commutative algebra and corresponds to point-wise product of coordinate integrals

$$\phi_e(\gamma) \phi_f(\gamma) = \phi_{e\Pi f}(\gamma)$$

This last identity, which goes back to Ree, is important because it says that if we consider two linear functions on $T((E))$ and multiply them together then their product - which is quadratic actually agrees with a linear functional on the group like elements. The shuffle product identifies the linear functional that does the job.

Lemma 9.4. Coordinate iterated integrals, as features of paths, span an algebra that separates Signatures and contains the constants.

This lemma is as important for understanding smooth functions on path spaces as monomials are for understanding smooth functions on \mathbb{R}^n . There are only finitely many of each degree if E is finite dimensional (although the dimension of the spaces grow exponentially) [20]. We will see later that this property is important

for machine learning and nonlinear regression applications but first we want to explain how the same remark allows one to understand measures on paths and formulate the notion of Fourier and Laplace transform.

10. Expected Signature

The study of the expected Signature was initiated by Fawcett in his thesis [9]. He proved

Proposition 10.1. Let μ be a compactly supported probability measure on paths γ with Signatures in a compact set K . Then $\hat{S} = \mathbb{E}_\mu(S(\gamma))$ uniquely determines the law of $S(\gamma)$.

Proof. Consider $\mathbb{E}_\mu(\phi_e(\gamma))$.

$$\begin{aligned}\mathbb{E}_\mu(\phi_e(\gamma)) &= \mathbb{E}_\mu(\langle e, S(\gamma) \rangle) \\ &= \langle e, \mathbb{E}_\mu(S(\gamma)) \rangle \\ &= \langle e, \hat{S} \rangle\end{aligned}$$

Since the e with the shuffle product form an algebra and separate points of K the Stone-Weierstrass Theorem implies they form a dense subspace in $C(K)$ and so determine the law of the Signature of γ . \square

Given this lemma it immediately becomes interesting to ask how does one compute $\mathbb{E}_\mu(S)$. Also, $\mathbb{E}_\mu(S)$ is like a Laplace transform and will fail to exist for reasons of tail behaviour of the random variables. Is there a characteristic function? Can we identify the general case where the expected Signature determines the law in the non-compact case. All of these are fascinating and important questions. Partial answers and strong applications are emerging. One of the earliest was the realisation that one could approximate effectively to a complex measure such as Wiener measure by a measure on finitely many paths that has the same expected Signature on $T^{(n)}(E)$ [19, 17].

11. Computing expected Signatures

Computing Laplace and Fourier transforms can often be a challenging problem for undergraduates. In this case suppose that X a Brownian motion with Lévy area on a bounded C^1 domain $\Omega \subset \mathbb{R}^d$, stopped on first exit. The following result explains how one may construct the expected Signature as a recurrence relation in PDEs[18].

Theorem 11.1. *Let*

$$\begin{aligned} F(z) & : = \mathbb{E}_z(S(X|_{[0,T_\Omega]})) \\ F & \in S((\mathbb{R}^d)) \\ F & = (f_0, f_1, \dots,) \end{aligned}$$

Then F satisfies and is determined by a PDE finite difference operator

$$\begin{aligned} \Delta f_{n+2} & = - \sum_{i=1}^d e_i \otimes e_i \otimes f_n - 2 \sum_{i=1}^d e_i \otimes \frac{\partial}{\partial z_i} f_{n+1} \\ f_0 & \equiv 1, \quad f_1 \equiv 0, \quad \text{and } f_j|_{\partial\Omega} \equiv 0, \quad j > 0 \end{aligned}$$

Combining this result with Sobolev and regularity estimates from PDE theory allow one to extract much nontrivial information about the underlying measure although it is still open whether in this case the expected Signature determines the measure. This question is difficult even for Brownian motion on $\min(T_\tau, t)$ although (unpublished) it looks as if the question can be resolved.

Other interesting questions about expected Signatures can be found for example in [1].

12. Characteristic Functions of Signatures

It is possible to build a characteristic function out of the expected Signature by looking at the linear differential equations corresponding to development of the paths into finite dimensional unitary groups. These linear images of the Signature are always bounded and so expectations always make sense.

Consider $SU(d) \subset M(d)$ and realise $su(d)$ as the space of traceless Hermitian matrices and consider

$$\begin{aligned} \psi & : E \rightarrow su(d) \\ d\Psi_t & = \psi(\Psi_t) d\gamma_t. \end{aligned}$$

Essential features of the co-ordinate interated integrals included that they were linear functions on the tensor algebra, that they were real valued functions that separated signatures, and that they spanned an algebra.

It is core to rough path theory that any representation of paths via a linear controlled equation can also be regarded as a linear function and that products can also be represented as sums. If one can show that products associated to the finite dimensional unitary groups can be expressed as sums of finite linear combinations of finite dimensional unitary representations, and add an appropriate topology on grouplike elements, one can repeat the ideas outlined above but now with expectations that always exist and obtain the analogue of characteristic function.

Theorem 12.1. Ψ_t is a linear functional on the tensor algebra restricted to the Signatures $S(\gamma|_{[0,t]})$ and is given by a convergent series. It is bounded and so its expectation as γ varies randomly always makes sense. The function $\psi \rightarrow \mathbb{E}(\Psi_{J_+}(S))$ is an extended characteristic function.

Proposition 12.2. $\psi \rightarrow \Psi(S)$ (polynomial identities of Gambruni and Valentini) span an algebra and separate Signatures as ψ and d vary.

Corollary 12.3. The laws of measures on Signatures are completely determined by $\psi \rightarrow \mathbb{E}(\Psi(S))$

Proof. Introduce a polish topology on the grouplike elements. \square

These results can be found in [7], the paper also gives a sufficient condition for the expected Signature to determine the law of the underlying measure on Signatures.

13. Moments are complicated

The question of determining the Signature from its moments seems quite hard at the moment.

Example 13.1. Observe that if X is $N(0, 1)$ then although X^3 is not determined by its moments, if $Y = X^3$ then (X, Y) is. The moment information implies $\mathbb{E}((Y - X^3)^2) = 0$.

We repeat our previous question. Does the expected Signature determine the law of the Signature for say stopped Brownian motion. The problem seems to capture the challenge.

Lemma 13.2 ([7]). If the radius of convergence of $\sum z^n \mathbb{E} \|S^n\|$ is infinite then the expected Signature determines the law.

Lemma 13.3 ([18]). If X a Brownian motion with Lévy area on a bounded C^1 domain $\Omega \subset \mathbb{R}^d$ then $\sum z^n \mathbb{E} \|S^n\|$ has at the least a strictly positive lower bound on the radius of curvature.

The gap in understanding between the previous two results is, for the author, a fascinating and surprising one that should be closed!

14. Regression onto a feature set

Learning how to regress or learn a function from examples is a basic problem in many different contexts. In what remains of this paper, we will outline recent work that explains how the Signature engages very naturally with this problem and why it is this engagement that makes it valuable in rough path theory too.

We should emphasise that the discussion and examples we give here is at a very primitive level of fitting curves. We are not trying to do statistics, or model and make inference about uncertainty. Rather we are trying to solve the most basic problems about extracting relationships from data that would exist even if one had perfect knowledge. We will demonstrate that this approach can be easy to implement and effective in reducing dimension and doing effective regression. We would expect Bayesian statistics to be an added layer added to the process where uncertainty exists in the data that can be modelled reasonably.

A core idea in many successful attempts to learn functions from a collection of known (point, value) pairs revolves around the identification of basic functions or features that are readily evaluated at each point and then try to express the observed function a *linear* combination of these basic functions. For example one might evaluate a smooth function ρ at a generic collection $\{x_i \in [0, 1]\}$ of points producing pairs $\{(y_i = \rho(x_i), x_i)\}$. Now consider as feature functions $\{\phi_n : x \rightarrow x^n, n = 0, \dots, N\}$. These are certainly easy to compute for each x_i . We try to express

$$\rho \simeq \sum_{n=0}^N \lambda_n \phi_n$$

and we see that if we can do this (that is to say ρ is well approximated by a polynomial) then the λ_n are given by the linear equation

$$y_j = \sum_{n=0}^N \lambda_n \phi_n(x_j).$$

In general one should expect, and it is even desirable, that the equations are significantly degenerate. The purpose of learning is presumably to be able to use the function $\sum_{n=0}^N \lambda_n \phi_n$ to predict ρ on new and unseen values of x and to at least be able to replicate the observed values of y .

There are powerful numerical techniques for identifying robust solutions to these equations. Most are based around least squares and singular value decomposition, along with L^1 constraints and Lasso.

However, this approach fundamentally depends on the assumption that the ϕ_n span the class of functions that are interesting. It works well for monomials because they span an algebra and so every $C^n(K)$ function can be approximated in $C^n(K)$ by a multivariate real polynomial. It relies on a priori knowledge of smoothness or Lasso style techniques to address over-fitting.

I hope the reader can now see the significance of the coordinate iterated integrals. If we are interested in functions (such as controlled differential equations) that are effects of paths or streams, then we know from the general theory of rough paths that the functions are indeed well approximated locally by linear combinations of coordinate iterated integrals. Coordinate iterated integrals are a natural feature set for capturing the aspects of the data that predicting the effects of the path on a controlled system.

The shuffle product ensures that linear combinations of coordinate iterated integrals are an algebra which ensures they span adequately rich classes of func-

tions. We can use the classical techniques of non-linear interpolation with these new feature functions to learn and model the behaviour of systems.

In many ways the machine learning perspective explains the whole theory of rough paths. If I want to model the effect of a path segment, I can do a good job by studying a few set features of my path locally. On smaller scales the approximations improve since the functionals the path interacts with become smoother. If the approximation error is small compared with the volume, and consistent on different scales, then knowing these features, and only these features, on all scales describes the path or function adequately enough to allow a limit and integration of the path or function against a Lipchitz function.

15. The obvious feature set for streams

The feature set that is the coordinate iterated integrals is able (with uniform error - even in infinite dimension) via linear combinations whose coefficients are derivatives of f , to approximate solutions to controlled differential equations [3]. In other words, any stream of finite length is characterised up to reparameterisation by its log Signature (see [15]) and the Poincare-Birkhoff-Witt theorem confirms that the coordinate iterated integrals are one way to parameterise the polynomials on this space. Many important nonlinear functions on paths are well approximated by these polynomials...

We have a well defined methodology for linearisation of smooth functions on unparameterised streams as linear functionals of the Signature. As we will explain in the remaining sections, this has potential for practical application even if it comes from the local embedding of a group into its enveloping algebra and identifying the dual with the real polynomials and analytic functions on the group.

16. Machine learning, an amateur's first attempt

Applications do not usually have a simple fix but require several methods in parallel to achieve significance. The best results to date for the use of Signatures have involved the recognition of Chinese characters [24] where Ben Graham put together a set of features based loosely on Signatures and state of the art deep learning techniques to win a worldwide competition organised by the Chinese Academy of Sciences.

We will adopt a different perspective and simply explain a very transparent and naive approach, based on Signatures, can achieve with real data. The work appeared in [12]. The project and the data depended on collaboration with commercial partners acknowledged in the paper and is borrowed from the paper.

16.1. classification of time-buckets from standardised data.

We considered a simple classification learning problem. We considered a moderate data set of 30 minutes intervals of normalised one minute financial market data,

which we will call buckets. The buckets are distinguished by the time of day that the trading is recorded. The buckets are divided into two sets - a learning and a backtesting set. The challenge is simple: learn to distinguish the time of day by looking at the normalised data (if indeed one can - the normalisation is intended to remove the obvious). It is a simple classification problem that can be regarded as learning a function with only two values

$$\begin{aligned} f(\text{time series}) &\rightarrow \text{time slot} \\ f(\text{time series}) = 1 &\quad \text{time slot}=10.30-11.00 \\ f(\text{time series}) = 0 &\quad \text{time slot}=14.00-14.30 \end{aligned}$$

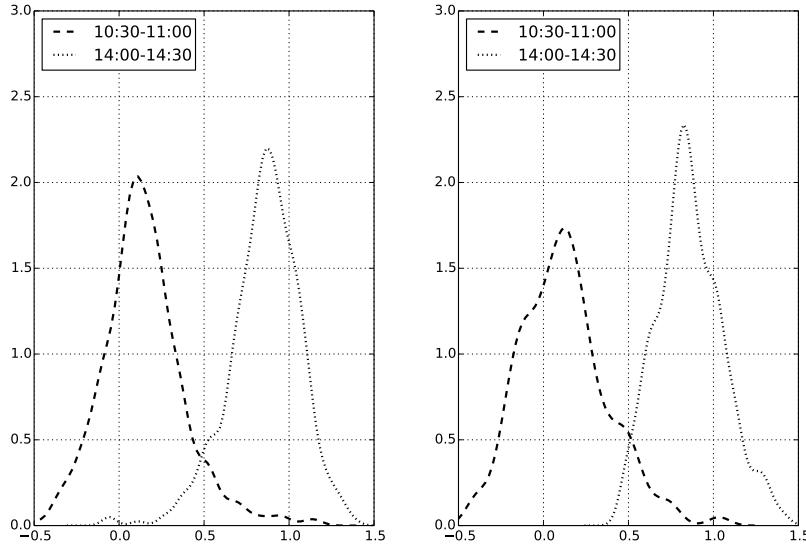
Our methodology has been spelt out. Use the low degree coordinates of the Signature of the normalised financial market data γ as features $\phi_i(\gamma)$, use least squares on the learning set to approximately reproduce f

$$f(\gamma) \approx \sum_i \lambda_i \phi_i(\gamma)$$

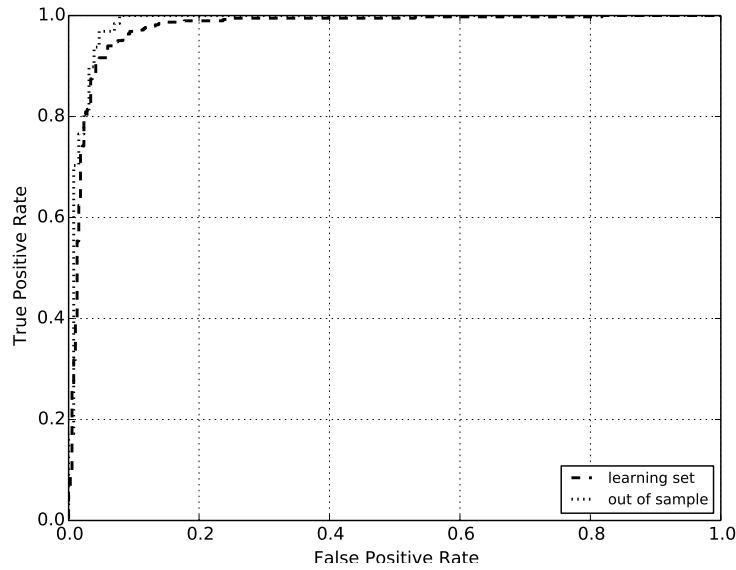
and then test it on the backtesting set. To summarise the methodology:

1. We used futures data normalised to remove volume and volatility information.
2. We used linear regression based pair-wise separation to find the best fit linear function to the learning pairs that assign 0 to one case and 1 to the other. (There are other well known methods that might be better.)
 - (a) We used robust and automated repeated sampling methods of LASSO type (least absolute shrinkage and selection operator) based on constrained L^1 optimisation to achieve shrinkage of the linear functional onto an expression involving only a few terms of the Signatures.
3. and we used simple statistical indicators to indicate the discrimination that the learnt function provided on the learning data and then on the backtesting data. The tests were:
 - (a) Kolmogorov-Smirnov distance of distributions of score values
 - (b) receiver operating characteristic (ROC) curve, area under ROC curve
 - (c) ratio of correct classification.

We did consider the full range of half hour time intervals. The other time intervals were not readily distinguishable from each other but were easily distinguishable from both of these two time intervals using the methodology mapped out here. It seems likely that the differences identified here were due to distinctive features of the market associated with the opening and closing of the open outcry market.



(a) **Learning set:** Estimated densities of the regressed values, K-S distance: 0.8, correct classification: 90%
(b) **Out of sample:** Estimated densities of the regressed values, K-S distance: 0.84, correct classification: 89%



(c) **ROC curve.** Area under ROC – learning set: 0.976, out of sample: 0.986

Figure 2. 14:00-14:30 EST versus 10:30-11:00 EST

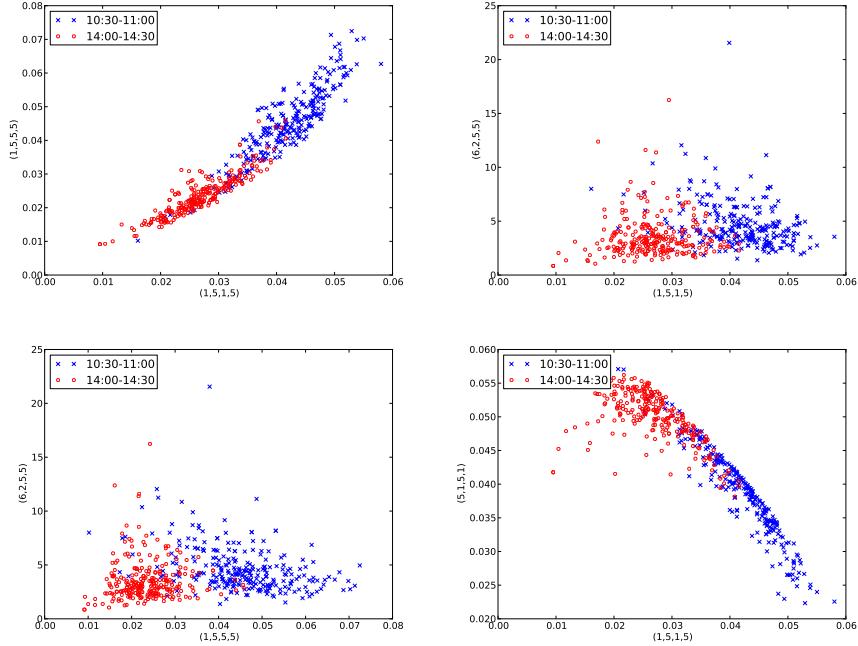


Figure 3. Visualisation: two dimensional projections of the 4th order signature onto coefficients selected as significant by Lasso shrinkage. The selected features allow clear visual separation of the time buckets.

17. Linear regression onto a law on paths

In the previous section we looked at using the linearising nature of the Signature as a practical tool for learning functions. In this final section we want to remain in the world of data and applications but make a more theoretical remark. Classic nonlinear regression is usually stated with a statistical element. One common formulation of linear regression has that a stationary sequence of random data pairs that are modeled by

$$y_i = f(x_i) + \varepsilon_i$$

where ε_i is random and has conditional mean zero. The goal is to determine the linear functional f with measurable confidence.

There are many situations where it is the case that one has a random but stationary sequence (γ, τ) of stream pairs, and one would like to learn, approximately, the law of τ conditional on γ . Suppose that we reformulate this problem in terms of Signatures and expected Signatures (or better: characteristic functions) recalling that expected Signatures etc. characterise laws.

Problem 17.1. Given a random but stationary sequence (γ, τ) of stream pairs

find the function $\Phi : S(\gamma) \rightarrow \mathbb{E}(S(\tau) | S(\gamma))$.

Then putting $Y_i = S(\tau_i)$ and $X_i = S(\gamma_i)$ we see that

$$Y_i = \Phi(X_i) + \varepsilon_i$$

where ε_i is random and has mean zero. If the measure is reasonably localised and smooth then we can well approximate Φ by a polynomial; and using the linearising nature of the tensor algebra to a linear function ϕ of the Signature. In other words the apparently difficult problem of understanding conditional laws of paths becomes (at least locally) a problem of linear regression

$$Y_i = \Phi(X_i) + \varepsilon_i$$

which is infinite dimensional but which has well defined low dimensional approximations [16].

References

- [1] Horatio Boedihardjo, Hao Ni, and Zhongmin Qian, *Uniqueness of signature for simple curves*, ArXiv preprint arXiv:1304.0755 (2013), 1–21.
- [2] Nicolas Bourbaki, *Lie groups and Lie algebras. Chapters 1–3*, Elements of Mathematics (Berlin), Springer-Verlag, Berlin, 1989, Translated from the French, Reprint of the 1975 edition. MR 979493 (89k:17001)
- [3] Youness Boudaïb, Lajos Gergely Gyurkó, Terry Lyons, and Danyu Yang, *Dimension-free euler estimates of rough differential equations*, arXiv:1307.4708 to appear in Rev. Roumaine Math. Pures Appl. (2014), 1–20.
- [4] Thomas Cass and Terry Lyons, *Evolving communities with individual preferences*, 1303.4243 to appear in Proceedings of London Mathematical Society (2014), 1–21.
- [5] Fabienne Castell and Jessica Gaines, *An efficient approximation method for stochastic differential equations by means of the exponential lie series*, Mathematics and computers in simulation **38** (1995), no. 1, 13–19.
- [6] Kuo-Tsai Chen, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Ann. of Math. (2) **65** (1957), 163–178. MR 0085251 (19,12a)
- [7] Ilya Chevyrev, *Unitary representations of geometric rough paths*, arXiv preprint arXiv:1307.3580 (2014).
- [8] Wei-Liang Chow, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann. **117** (1939), 98–105. MR 0001880 (1,313d)
- [9] Thomas Fawcett, *Problems in stochastic analysis: Connections between rough paths and non-commutative harmonic analysis*, Ph.D. thesis, University of Oxford, 2002.
- [10] Guy Flint, Ben Hambly, and Terry Lyons, *Convergence of sampled semimartingale rough paths and recovery of the $it^{\wedge}\{o\}$ integral*, arXiv preprint arXiv:1310.4054v5 (2013), 1–22.
- [11] Peter K Friz and Nicolas B Victoir, *Multidimensional stochastic processes as rough paths: theory and applications*, vol. 120, Cambridge University Press, 2010.

- [12] Lajos Gergely Gyurkó, Terry Lyons, Mark Kontkowski, and Jonathan Field, *Extracting information from the signature of a financial data stream*, arXiv preprint arXiv:1307.7244 (2013).
- [13] Martin Hairer, *A theory of regularity structures*, Invent. Math. (2014).
- [14] Martin Hairer and Natesh S Pillai, *Regularity of laws and ergodicity of hypoelliptic sdes driven by rough paths*, The Annals of Probability **41** (2013), no. 4, 2544–2598.
- [15] Ben Hambly and Terry Lyons, *Uniqueness for the signature of a path of bounded variation and the reduced path group*, Ann. of Math.(2) **171** (2010), no. 1, 109–167.
- [16] Daniel Levin, Terry Lyons, and Hao Ni, *Learning from the past, predicting the statistics for the future, learning an evolving system*, arXiv preprint arXiv:1309.0260 (2013), 1–32.
- [17] Christian Litterer and Terry Lyons, *Cubature on wiener space continued*, Stochastic Processes and Applications to Mathematical Finance (2011), 197–218.
- [18] Terry Lyons and Hao Ni, *Expected signature of two dimensional Brownian Motion up to the first exit time of the domain*, arXiv:1101.5902v4 (2011), 1–27.
- [19] Terry Lyons and Nicolas Victoir, *Cubature on wiener space*, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **460** (2004), no. 2041, 169–198.
- [20] Terry J Lyons, Michael Caruana, and Thierry Lévy, *Differential equations driven by rough paths*, Springer, 2007.
- [21] Terry J. Lyons and Nadia Sidorova, *On the radius of convergence of the logarithmic signature*, Illinois J. Math. **50** (2006), no. 1-4, 763–790 (electronic). MR 2247845 (2007m:60165)
- [22] Terry J Lyons and Danyu Yang, *The partial sum process of orthogonal expansions as geometric rough process with fourier series as an examplean improvement of menshov-rademacher theorem*, Journal of Functional Analysis **265** (2013), no. 12, 3067–3103.
- [23] P. K. Rashevski, *About connecting two points of complete nonholonomic space by admissible curve*, Uch Zapiski ped. inst. Libknekhta **2** (1938), 83–94.
- [24] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu, *Icdar 2013 chinese handwriting recognition competition*, Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, IEEE, 2013, pp. 1464–1470.

Oxford-Man Institute of Quantitative Finance, University of Oxford, England, OX2 6ED

E-mail: terry.lyons@oxford-man.ox.ac.uk