# A Primer on the Signature Method in Machine Learning

**Ilya Chevyrev**[a] **and Andrey Kormilitzin**[a,b]

[a] *Mathematical Institute, University of Oxford, Andrew Wiles Building,Woodstock Road, Oxford, OX2 6CG, UK*

[b] *Oxford-Man Institute, University of Oxford, Eagle House, Walton Well Road, Oxford, OX2 6ED, UK*

*E-mail:* ilya.chevyrev@maths.ox.ac.uk,
andrey.kormilitzin@maths.ox.ac.uk



Abstract: In these notes, we wish to provide an introduction to the signature method, focusing on its basic theoretical properties and recent numerical applications.

The notes are split into two parts. The first part focuses on the definition and fundamental properties of the signature of a path, or the *path signature*. We have aimed for a minimalistic approach, assuming only familiarity with classical real analysis and integration theory, and supplementing theory with straightforward examples. We have chosen to focus in detail on the principle properties of the signature which we believe are fundamental to understanding its role in applications. We also present an informal discussion on some of its deeper properties and briefly mention the role of the signature in rough paths theory, which we hope could serve as a light introduction to rough paths for the interested reader.

The second part of these notes discusses practical applications of the path signature to the area of machine learning. The signature approach represents a non-parametric way for extraction of characteristic features from data. The data are converted into a multi-dimensional path by means of various embedding algorithms and then processed for computation of individual terms of the signature which summarise certain information contained in the data. The signature thus transforms raw data into a set of features which are used in machine learning tasks. We will review current progress in applications of signatures to machine learning problems.

## Contents

---

# 1 Theoretical Foundations

The purpose of the first part of these notes is to introduce the definition of the signature and present some of its fundamental properties. The point of view we take in these notes is that the signature is an object associated with a path which captures many of the path's important analytic and geometric properties.

The signature has recently gained attention in the mathematical community in part due to its connection with Lyons' theory of rough paths. At the end of this first part, we shall briefly highlight the role the signature plays in the theory of rough paths on an informal level. However one of the main points we wish to emphasize is that no knowledge beyond classical integration theory is required to define and study the basic properties of the signature. Indeed, K. T. Chen was one of the first authors to study the signature, and his primary results can be stated completely in terms of piecewise smooth paths, which already provide an elegant and deep mathematical theory.

While we attempt to make all statements mathematically precise, we refrain from going into complete detail. A further in-depth discussion, along with proofs of many of the results covered in the first part, can now be found in several texts, and we recommend the St. Flour lecture notes [20] for the curious reader.

## 1.1 Preliminaries

### 1.1.1 Paths in Euclidean space

Paths form one of the basic elements of this theory. A path $X$ in $\mathbb{R}^d$ is a continuous mapping from some interval $[a, b]$ to $\mathbb{R}^d$, written as $X : [a, b] \mapsto \mathbb{R}^d$. We will use the subscript notation $X_t = X(t)$ to denote dependence on the parameter $t \in [a, b]$.

For our discussion of the signature, unless otherwise stated, we will always assume that paths are piecewise differentiable (more generally, one may assume that the paths are of bounded variation for which exactly the same classical theory holds). By a smooth path, we mean a path which has derivatives of all orders.

Two simple examples of smooth paths in $\mathbb{R}^2$ are presented in Fig.1:

$$\begin{aligned} \text{left panel:} \quad & X_t = \{X_t^1, X_t^2\} = \{t, t^3\} \ , \ t \in [-2, 2] \\ \text{right panel:} \quad & X_t = \{X_t^1, X_t^2\} = \{\cos t, \sin t\} \ , \ t \in [0, 2\pi] \end{aligned} \tag{1.1}$$

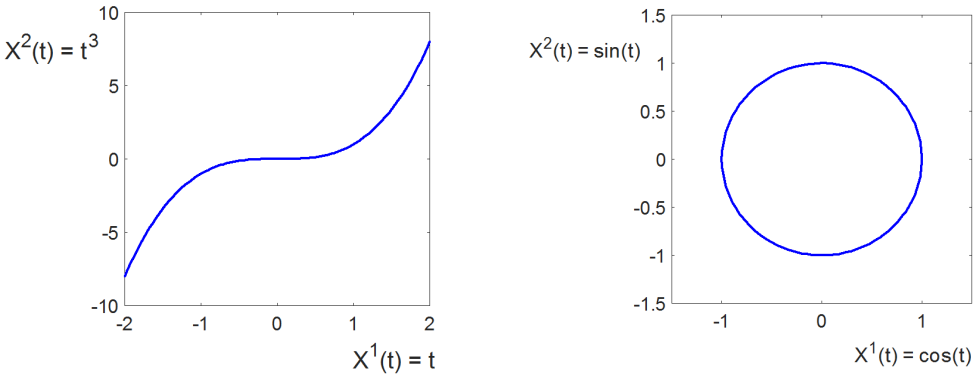



**Figure 1**: Example of two-dimensional smooth paths.

This parametrisation generalizes in $d$-dimensions ($X_t \in \mathbb{R}^d$) as:

$$X : [a, b] \mapsto \mathbb{R}^d \ , \quad X_t = \left\{X_t^1, X_t^2, X_t^3, ..., X_t^d\right\}. \tag{1.2}$$

An example of a piecewise linear path is presented in Fig.2:

$$X_t = \{X_t^1, X_t^2\} = \{t, f(t)\} \ , \ t \in [0, 1], \tag{1.3}$$

where $f$ is a piecewise linear function on the time domain $[0, 1]$. One possible example of the function $f$ is a stock price at time $t$. Such non-smooth paths may represent sequential data or time series, typically consisting of successive measurements made over a time interval.

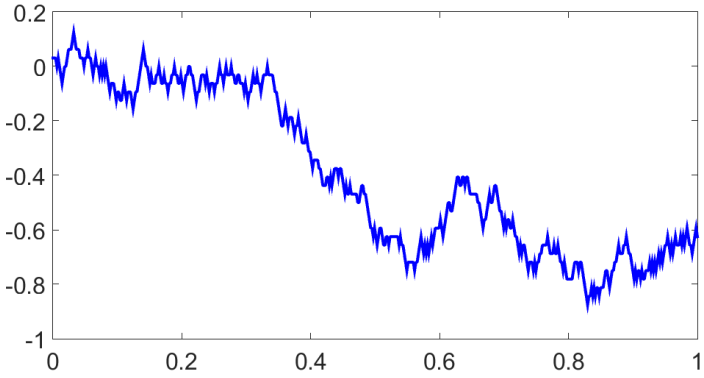



**Figure 2**: Example of non-smooth stochastic path.

### 1.1.2 Path integrals

We now briefly review the path (or line) integral. The reader may already be familiar with the common definition a path integral against a fixed function $f$ (also called a one-form). Namely, for a one-dimensional path $X : [a, b] \mapsto \mathbb{R}$ and a function $f : \mathbb{R} \mapsto \mathbb{R}$, the path integral of $X$ against $f$ is defined by

$$\int_a^b f(X_t)\, dX_t \;=\; \int_a^b f(X_t)\dot{X}_t dt, \tag{1.4}$$

where the last integral is the usual (Riemann) integral of a continuous bounded function and where we use the "upper-dot" notation for differentiation with respect to a single variable: $\dot{X}_t = dX_t/dt$.

In the expression (1.4), note that $f(X_t)$ is itself a real-valued path defined on $[a, b]$. In fact, (1.4) is a special case of the Riemann-Stieltjes integral of one path against another. In general, one can integrate any path $Y : [a, b] \mapsto \mathbb{R}$ against a path $X : [a, b] \mapsto \mathbb{R}$. Namely, for a path $Y : [a, b] \mapsto \mathbb{R}$, we can define the integral

$$\int_a^b Y_t\, dX_t \;=\; \int_a^b Y_t \dot{X}_t dt. \tag{1.5}$$

As previously remarked, we recover the usual path integral upon setting $Y_t = f(X_t)$.

**Example 1.** Consider the constant path $Y_t = 1$ for all $t \in [a, b]$. Then the path integral of $Y$ against any path $X : [a, b] \mapsto \mathbb{R}$ is simply the increment of $X$:

$$\int_a^b dX_t \;=\; \int_a^b \dot{X}_t\, dt \;=\; X_b - X_a. \tag{1.6}$$

**Example 2.** Consider the path $X_t = t$ for all $t \in [a, b]$. It follows that $\dot{X}_t = 1$ for all $t \in [a, b]$, and so the path integral for any $Y : [a, b] \mapsto \mathbb{R}$ is the usual Riemann integral of $Y$:

$$\int_a^b Y_t dX_t \;=\; \int_a^b Y_t\, dt. \tag{1.7}$$

**Example 3.** We present an example involving numerical computations. Consider the two-dimensional path

$$X_t = \{X_t^1, X_t^2\} = \{t^2, t^3\}, \quad t \in [0, 1]. \tag{1.8}$$

Then we can compute the path integral

$$\int_0^1 X_t^1 dX_t^2 \; = \; \int_0^1 t^2 3t^2 dt \; = \; \frac{3}{5}. \tag{1.9}$$

The above example is a special case of an iterated integral, which, as discussed in the following section, is central to the definition of the path signature.

## 1.2 The signature of a path

### 1.2.1 Definition

Having recalled the path integral of one real-valued path against another, we are now ready to define the signature of a path. For a path $X : [a, b] \mapsto \mathbb{R}^d$, recall that we denote the coordinate paths by $(X_t^1, \dots X_t^d)$, where each $X^i : [a, b] \mapsto \mathbb{R}$ is a real-valued path. For any single index $i \in \{1, \dots, d\}$, let us define the quantity

$$S(X)_{a,t}^i = \int_{a<s<t} dX_s^i = X_t^i - X_0^i, \tag{1.10}$$

which is the increment of the $i$-th coordinate of the path at time $t \in [a, b]$. We emphasise that $S(X)_{a,\cdot}^i : [a, b] \mapsto \mathbb{R}$ is itself a real-valued path. Note that $a$ in the subscript of $S(X)_{a,t}^i$ is only used to denote the starting point of the interval $[a, b]$.

Now for any pair $i, j \in \{1, \dots, d\}$, let us define the *double-iterated* integral

$$S(X)_{a,t}^{i,j} = \int_{a<s<t} S(X)_{a,s}^i dX_s^j \; = \; \int_{a<r<s<t} dX_r^i \, dX_s^j, \tag{1.11}$$

where $S(X)_{a,s}^i$ is given by (1.10) and the integration limits are simply:

$$a < r < s < t = \begin{cases} a < r < s \\ a < s < t. \end{cases} \tag{1.12}$$

The integration limits (1.12) also correspond to integration over a triangle (or, more generally, over a simplex in higher dimension). We emphasise again that $S(X)_{a,s}^i$ and $X_s^j$ are simply real-valued paths, so the expression (1.11) is a special case of the path integral, and that $S(X)_{a,\cdot}^{i,j} : [a, b] \mapsto \mathbb{R}$ is itself a real-valued path.

Likewise for any triple $i, j, k \in \{1, \dots, d\}$ we define the *triple-iterated* integral

$$S(X)_{a,t}^{i,j,k} = \int_{a<s<t} S(X)_{a,s}^{i,j} dX_s^k \; = \; \int_{a<q<r<s<t} dX_q^i dX_r^j \, dX_s^k. \tag{1.13}$$

Again, since $S(X)^{i,j}_{a,s}$ and $X^k_s$ are real-valued paths, the above is just a special case of the path integral, and $S(X)^{i,j,k}_{a,\cdot} : [a,b] \mapsto \mathbb{R}$ is itself a real-valued path.

We can continue recursively, and for any integer $k \geq 1$ and collection of indexes $i_1, \ldots, i_k \in \{1, \ldots, d\}$, we define

$$S(X)^{i_1,\ldots,i_k}_{a,t} = \int_{a<s<t} S(X)^{i_1,\ldots,i_{k-1}}_{a,s} dX^{i_k}_s. \tag{1.14}$$

As before, since $S(X)^{i_1,\ldots,i_{k-1}}_{a,s}$ and $X^{i_k}_s$ are real-valued paths, the above is defined as a path integral, and $S(X)^{i_1,\ldots,i_k}_{a,\cdot} : [a,b] \mapsto \mathbb{R}$ is itself a real-valued path. Observe that we may equivalently write

$$S(X)^{i_1,\ldots,i_k}_{a,t} = \int_{a<t_k<t} \cdots \int_{a<t_1<t_2} dX^{i_1}_{t_1} \ldots dX^{i_k}_{t_k}. \tag{1.15}$$

The real number $S(X)^{i_1,\ldots,i_k}_{a,b}$ is called the *k-fold iterated integral* of $X$ along the indexes $i_1, \ldots, i_k$.

**Definition 1** (Signature)**.** The *signature* of a path $X : [a,b] \mapsto \mathbb{R}^d$, denoted by $S(X)_{a,b}$, is the collection (infinite series) of all the iterated integrals of $X$. Formally, $S(X)_{a,b}$ is the sequence of real numbers

$$S(X)_{a,b} \;=\; (1, S(X)^1_{a,b}, \ldots, S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \ldots) \tag{1.16}$$

where the "zeroth" term, by convention, is equal to 1, and the superscripts run along the set of all *multi-indexes*

$$W = \{(i_1, \ldots, i_k) \mid k \geq 1, i_1, \ldots, i_k \in \{1, \ldots, d\}\}. \tag{1.17}$$

The set $W$ above is also frequently called the set of *words* on the *alphabet* $A = \{1, \ldots, d\}$ consisting of $d$ letters.

**Example 4.** Consider an alphabet consisting of three letters only: $\{1, 2, 3\}$. There is infinite number of words which could be composed from this alphabet, namely:

$$\{1, 2, 3\} \to (1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, 121, \ldots). \tag{1.18}$$

An important property of the signature which we immediately note is that the iterated integrals of a path $X$ are independent of the starting point of $X$. That is, if for some $x \in \mathbb{R}^d$, we define the path $\widetilde{X}_t = X_t + x$, then $S(\widetilde{X})^{i_1,\ldots,i_k}_{a,b} = S(X)^{i_1,\ldots,i_k}_{a,b}$.

We shall often consider the *k-th level* of the signature, defined as the finite collection of all terms $S(X)^{i_1,\ldots,i_k}_{a,b}$ where the multi-index is of length $k$. For example, the first level of the signature is the collection of $d$ real numbers $S(X)^1_{a,b}, \ldots, S(X)^d_{a,b}$, and the second level is the collection of $d^2$ real numbers

$$S(X)^{1,1}_{a,b}, \ldots, S(X)^{1,d}_{a,b}, S(X)^{2,1}_{a,b}, \ldots, S(X)^{d,d}_{a,b}. \tag{1.19}$$

### 1.2.2 Examples

**Example 5.** The simplest example of a signature which one should keep in mind is that of a one-dimensional path. In this case our set of indexes (or alphabet) is of size one, $A = \{1\}$, and the set of multi-indexes (or words) is $W = \{(1, \dots, 1) \mid k \geq 1\}$, where 1 appears $k$ times in $(1, \dots, 1)$.

Consider the path $X : [a, b] \mapsto \mathbb{R}$, $X_t = t$. One can immediately verify that the signature of $X$ is given by

$$\begin{aligned} S(X)^1_{a,b} &= X_b - X_a, \\ S(X)^{1,1}_{a,b} &= \frac{(X_b - X_a)^2}{2!}, \\ S(X)^{1,1,1}_{a,b} &= \frac{(X_b - X_a)^3}{3!}, \\ &\vdots . \end{aligned} \tag{1.20}$$

One can in fact show that the above expression of the signature remains true for any path $X : [a, b] \mapsto \mathbb{R}$. Hence, for one-dimensional paths, the signature depends only on the increment $X_b - X_a$.

**Example 6.** We present now a more involved example of the signature for a two-dimensional path. Our set of indexes is now $A = \{1, 2\}$, and the set of multi-indexes is

$$W = \{(i_1, \dots, i_k) \mid k \geq 1, i_1, \dots, i_k \in \{1, 2\}\}, \tag{1.21}$$

the collection of all finite sequences of 1's and 2's. Consider a parabolic path in $\mathbb{R}^2$ as depicted in Fig.3.

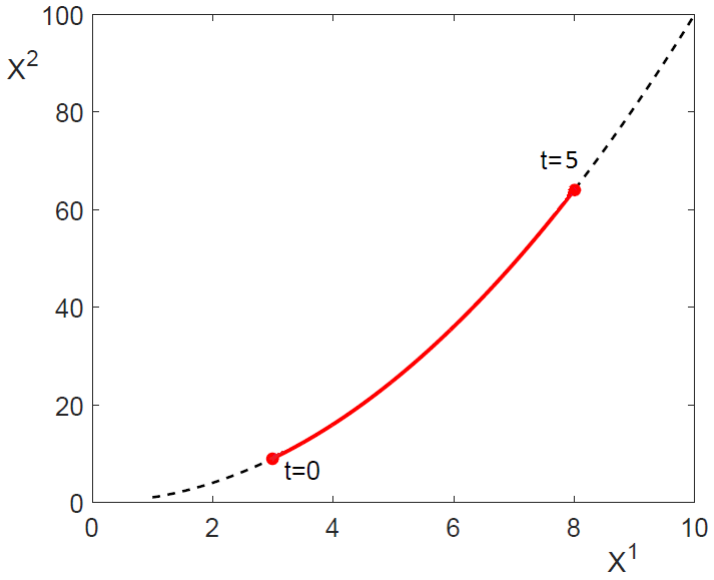



**Figure 3**: Example of a two dimensional path parametrized in (1.22).

Explicitly:

$$\begin{aligned} X_t &= \{X_t^1, X_t^2\} = \{3+t, (3+t)^2\} \quad t \in [0,5]\ ,\ (a=0, b=5), \\ dX_t &= \{dX_t^1, dX_t^2\} = \{dt, 2(3+t)dt\}. \end{aligned} \tag{1.22}$$

A straightforward computation gives:

$$\begin{aligned} S(X)_{0,5}^1 &= \int_{0<t<5} dX_t^1 = \int_0^5 dt = X_5^1 - X_0^1 = 5, \\ S(X)_{0,5}^2 &= \int_{0<t<5} dX_t^2 = \int_0^5 2\,(3+t)\,dt = X_5^2 - X_0^2 = 55, \\ S(X)_{0,5}^{1,1} &= \iint_{0<t_1<t_2<5} dX_{t_1}^1 dX_{t_2}^1 = \int_0^5 \left[\int_0^{t_2} dt_1\right] dt_2 = \frac{25}{2}, \\ S(X)_{0,5}^{1,2} &= \iint_{0<t_1<t_2<5} dX_{t_1}^1 dX_{t_2}^2 = \int_0^5 \left[\int_0^{t_2} dt_1\right] 2\,(3+t_2)\,dt_2 = \frac{475}{3}, \\ S(X)_{0,5}^{2,1} &= \iint_{0<t_1<t_2<5} dX_{t_1}^2 dX_{t_2}^1 = \int_0^5 \left[\int_0^{t_2} 2\,(3+t_1)\,dt_1\right] dt_2 = \frac{350}{3}, \\ S(X)_{0,5}^{2,2} &= \iint_{0<t_1<t_2<5} dX_{t_1}^2 dX_{t_2}^2 = \int_0^5 \left[\int_0^{t_2} 2\,(3+t_1)\,dt_1\right] 2\,(3+t_2)\,dt_2 = \frac{3025}{2}, \\ S(X)_{0,5}^{1,1,1} &= \iiint_{0<t_1<t_2<t_3<5} dX_{t_1}^1 dX_{t_2}^1 dX_{t_3}^1 = \int_0^1 \left[\int_0^{t_3} \left[\int_0^{t_2} dt_1\right] dt_2\right] dt_3 = \frac{125}{6}, \\ &\vdots\ . \end{aligned} \tag{1.23}$$

Continuing this way, one can compute every term $S(X)_{0,5}^{i_1,\dots,i_k}$ of the signature for every multi-index $(i_1,\dots,i_k)$, $i_1,\dots,i_k \in \{1,2\}$.

### 1.2.3 Picard iterations: motivation for the signature

Before reviewing the basic properties of the signature, we take a moment to show how the signature arises naturally in the classical theory of ordinary differential equations (ODEs). In a sense, this provides one of the first reasons for our interest in the signature. We do not provide all the details, but hope that the reader finds the general idea clear.

It is instructive to start the discussion with an intuitive and simplistic example of Picard's method. Let us consider the first order ODE

$$\frac{dy}{dx} = f(x,y), \qquad y(x_0) = y_0, \tag{1.24}$$

where $y(x)$ is a real valued function of a scalar variable $x$. Picard's method allows us to construct an approximated solution to (1.24) in the form of an iterative series. The integral form of (1.24) is given by

$$y(x) = y(x_0) + \int_{x_0}^{x} f(t, y(t))\,dt. \tag{1.25}$$

We now define a sequence of functions $y_k(x)$, $k = 0, 1, \ldots$, where the first term is the constant function $y_0(x) = y(x_0)$, and for $k \geq 1$, we define inductively

$$y_k(x) \;=\; y(x_0) + \int_{x_0}^{x} f(t, y_{k-1}(t))\, dt. \tag{1.26}$$

The classical Picard-Lindelöf theorem states that, under suitable conditions, the solution to (1.24) is given by $y(x) = \lim_{k\to\infty} y_k(x)$.

**Example 7.** Consider the ODE:

$$\frac{dy}{dx} = y(x), \quad y(0) = 1. \tag{1.27}$$

The first $k$ terms of the Picard iterations are given by:

$$\begin{aligned}
y_0(x) &= 1 \\
y_1(x) &= 1 + \int_0^x y_0(t)dt = 1 + x \\
y_2(x) &= 1 + \int_0^x y_1(t)dt = 1 + x + \frac{1}{2}x^2 \\
y_3(x) &= 1 + \int_0^x y_2(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 \\
y_4(x) &= 1 + \int_0^x y_3(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 \\
&\vdots \\
y_k(x) &= \sum_{n=0}^{k} \frac{1}{n!}x^n,
\end{aligned} \tag{1.28}$$

which converges to $y(x) = e^x$ as $k \to \infty$, which is indeed the solution to (1.27). These approximations are plotted in Fig. 4.

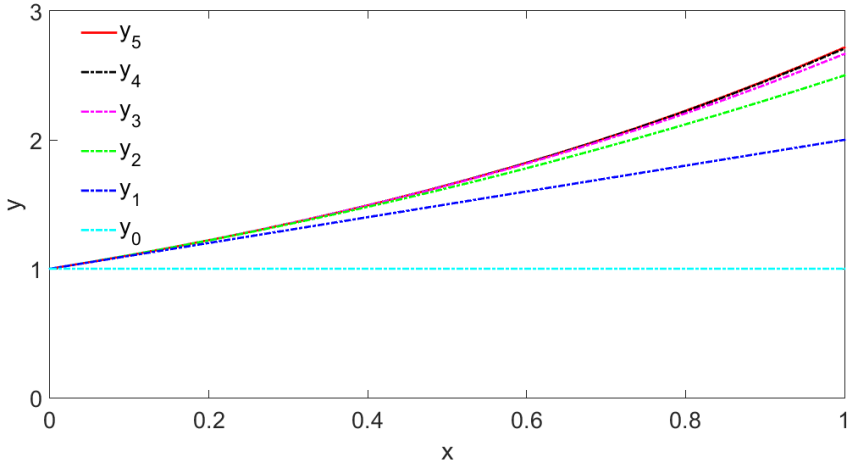



**Figure 4**: Example of sequential Picard approximation to the true solution.

We are now ready to consider a controlled differential equation and the role of the signature in its solution. Consider a path $X : [a, b] \mapsto \mathbb{R}^d$. Let $\mathbf{L}(\mathbb{R}^d, \mathbb{R}^e)$ denote the vector space of linear maps from $\mathbb{R}^d$ to $\mathbb{R}^e$. Equivalently, $\mathbf{L}(\mathbb{R}^d, \mathbb{R}^e)$ can be regarded as the vector space of $d \times e$ real matrices. For a path $Z : [a, b] \mapsto \mathbf{L}(\mathbb{R}^d, \mathbb{R}^e)$, note that we can define the integral

$$\int_a^b Z_t dX_t \tag{1.29}$$

as an element of $\mathbb{R}^e$ in exactly the same way as the usual path integral. For a function $V : \mathbb{R}^e \mapsto \mathbf{L}(\mathbb{R}^d, \mathbb{R}^e)$ and a path $Y : [a, b] \mapsto \mathbb{R}^e$, we say that $Y$ solves the controlled differential equation

$$dY_t = V(Y_t)dX_t, \quad Y_a = y \in \mathbb{R}^e, \tag{1.30}$$

precisely when for all times $t \in [a, b]$

$$Y_t = y + \int_a^t V(Y_s)dX_s. \tag{1.31}$$

The map $V$ in the above expression is often called a collection of *driving vector fields*, the path $X$ is called the *control* or the *driver*, and $Y$ is called the *solution* or the *response*.

A standard procedure to obtain a solution to (1.31) is through Picard iterations. For an arbitrary path $Y : [a, b] \mapsto \mathbb{R}^e$, define a new path $F(Y) : [a, b] \mapsto \mathbb{R}^e$ by

$$F(Y)_t = y + \int_a^t V(Y_s)dX_s. \tag{1.32}$$

Observe that $Y$ a solution to (1.31) if and only if $Y$ is a fixed point of $F$. Consider the sequence of paths $Y_t^n = F(Y^{n-1})_t$ with initial arbitrary path $Y_t^0$ (often taken as the constant path $Y_t^0 = y$). Under suitable assumptions, one can show that $F$ possesses a unique fixed point $Y$ and that $Y_t^n$ converges to $Y$ as $n \to \infty$.

Consider now the case when $V : \mathbb{R}^e \mapsto \mathbf{L}(\mathbb{R}^d, \mathbb{R}^e)$ is a linear map. Note that we may equivalently treat $V$ as a linear map $\mathbb{R}^d \mapsto \mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$, where $\mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$ is the space of all $e \times e$ real matrices. Let us start the Picard iterations with the initial constant path $Y_t^0 = y$ for all $t \in [a, b]$. Denoting by $I_e$ the identity operator (or matrix) in $\mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$, it follows that the iterates of $F$ can be expressed as follows:

$$\begin{aligned}
&Y_t^0 = y, \qquad\qquad (1.33)\\
&Y_t^1 = y + \int_a^t V(Y_s^0)dX_s = \left(\int_a^t dV(X_s) + I_e\right)(y),\\
&Y_t^2 = y + \int_a^t V(Y_s^1)dX_s = \left(\int_a^t \int_a^s dV(X_u)dV(X_s) + \int_a^t dV(X_s) + I_e\right)(y),\\
&\quad\vdots\\
&Y_t^n = y + \int_a^t V(Y_s^{n-1})dX_s = \left(\sum_{k=1}^n \int_{a<t_1<\ldots<t_k<t} dV(X_{t_1})\ldots dV(X_{t_k}) + I_e\right)(y),\\
&\quad\vdots
\end{aligned}$$

Due to the fact that $\mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$ is an algebra of matrices, each quantity

$$\int_{a<t_1<\ldots<t_k<t} dV(X_{t_1}) \ldots dV(X_{t_k}) \tag{1.34}$$

can naturally be defined as an element of $\mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$, which, one can check, is completely determined (in a linear way) by the $k$-th level of the signature $S(X)_{a,t}$ of $X$ at time $t \in [a, b]$.

The conclusion we obtain is that the solution $Y_t$ is completely determined by the signature $S(X)_{a,t}$ for every $t \in [a, b]$. In particular, if the signatures of two controls $X$ and $\widetilde{X}$ coincide at time $t \in [a, b]$, that is, $S(X)_{a,t} = S(\widetilde{X})_{a,t}$, then the corresponding solutions to (1.31) will also agree at time $t$ for any choice of the linear vector fields $V$.

An important, but far less obvious result, is that the same conclusion holds true for non-linear vector fields $V$. This result was first obtained by Chen [5] for a certain class of piecewise smooth paths, and recently extended by Hambly and Lyons [14] to paths of bounded variation, and by Boedihardjo, Geng, Lyons and Yang [2] to a completely non-smooth setting of geometric rough paths for which the signature is still well-defined (see Section 1.4 for further discussion). The latter class of paths is of particular interest from the point of view of stochastic analysis.

### 1.2.4 Geometric intuition of the first two levels

While the signature is defined analytically using path integrals, we briefly discuss here the geometric meaning of the first two levels. As already mentioned, the first level, given by the terms $(S(X)^1_{a,b}, \ldots, S(X)^d_{a,b})$, is simply the increment of the path $X : [a, b] \mapsto \mathbb{R}^d$. For the second level, note that the term $S(X)^{i,i}_{a,b}$ is always equal to $(X^i_b - X^i_a)^2/2$. This relation is a special case of the shuffle product which we shall review in Section 1.3.2. To give meaning to the term $S(X)^{i,j}_{a,b}$ for $i \neq j$, consider the *Lévy area* (illustrated in Fig.5), which is a *signed* area enclosed by the path (solid red line) and the chord (blue straight dashed line) connecting the endpoints. The Lévy area of the two dimensional path $\{X^1_t, X^2_t\}$ is given by:

$$A \;=\; \frac{1}{2}\left(S(X)^{1,2}_{a,b} - S(X)^{2,1}_{a,b}\right). \tag{1.35}$$

The signed areas denoted by $A_-$ and $A_+$ are the negative and positive areas respectively, and $\Delta X^1$ and $\Delta X^2$ represent the increments along each coordinate.

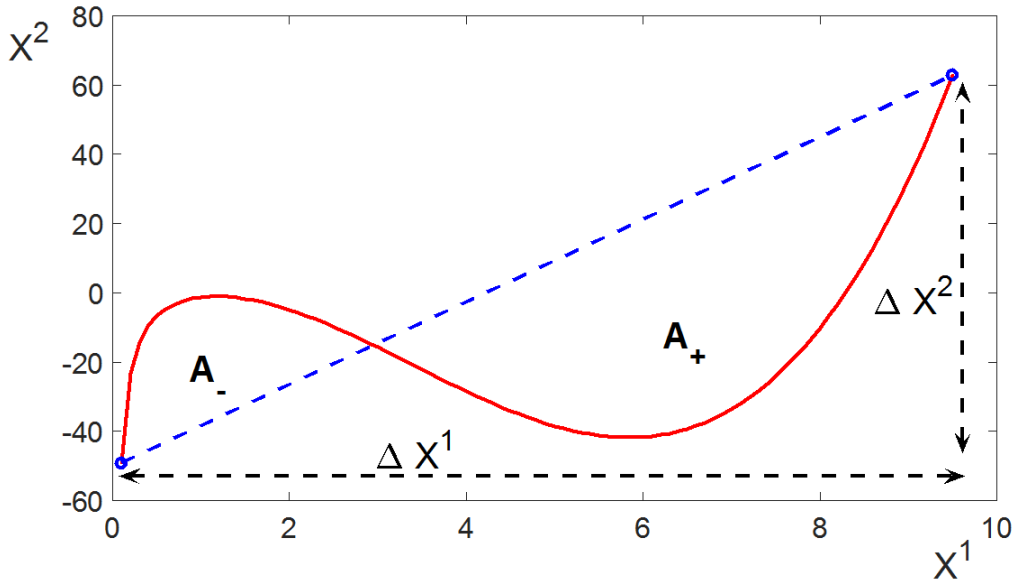



**Figure 5**: Example of signed Lévy area of a curve. Areas above and under the chord connecting two endpoints are negative and positive respectively.

### 1.3 Important properties of signature

We now review several fundamental properties of the signature of paths. We do not provide all the details behind the proofs of these properties, but we emphasize that they are all straightforward consequences of classical integration theory. Several deeper results are discussed in the following Section 1.4, but only on an informal level.

#### 1.3.1 Invariance under time reparametrisations

We call a surjective, continuous, non-decreasing function $\psi : [a, b] \mapsto [a, b]$ a *reparametrization*. For simplicity, we shall only consider smooth reparametrizations, although, just like in the definition of the path integral, this is not strictly necessary.

Let $X, Y : [a, b] \mapsto \mathbb{R}$ be two real-valued paths and $\psi : [a, b] \mapsto [a, b]$ a reparametrization. Define the paths $\widetilde{X}, \widetilde{Y} : [a, b] \mapsto \mathbb{R}$ by $\widetilde{X}_t = X_{\psi(t)}$ and $\widetilde{Y}_t = Y_{\psi(t)}$. Observe that

$$\dot{\widetilde{X}}_t = \dot{X}_{\psi(t)}\dot{\psi}(t), \tag{1.36}$$

from which it follows that

$$\int_a^b \widetilde{Y}_t d\widetilde{X}_t = \int_a^b Y_{\psi(t)}\dot{X}_{\psi(t)}\dot{\psi}(t)dt = \int_a^b Y_u dX_u, \tag{1.37}$$

where the last equality follows by making the substitution $u = \psi(t)$. This shows that path integrals are invariant under a time reparametrization of both paths.

Consider now a multi-dimensional path $X : [a, b] \mapsto \mathbb{R}^d$ and a reparametrization $\psi : [a, b] \mapsto [a, b]$. As before, denote by $\widetilde{X} : [a, b] \mapsto \mathbb{R}^d$ the reparametrized path $\widetilde{X}_t = X_{\psi(t)}$. Since every term of the signature $S(X)_{a,b}^{i_1,\ldots,i_k}$ is defined as an iterated path integral of $X$, it follows from the above that

$$S(\widetilde{X})_{a,b}^{i_1,\ldots,i_k} = S(X)_{a,b}^{i_1,\ldots,i_k}, \ \ \forall k \geq 0, \ i_1, \ldots, i_k \in \{1, \ldots, d\}. \tag{1.38}$$

That is to say, the signature $S(X)_{a,b}$ remains invariant under time reparametrizations of $X$.

### 1.3.2 Shuffle product

One of the fundamental properties of the signature, shown originally by Ree [22], is that the product of two terms $S(X)_{a,b}^{i_1,\ldots,i_k}$ and $S(X)_{a,b}^{j_1,\ldots,j_m}$ can always be expressed as a sum of another collection of terms of $S(X)_{a,b}$ which only depends on the multi-indexes $(i_1, \ldots, i_k)$ and $(j_1, \ldots, j_m)$.

To make this statement precise, we define the shuffle product of two multi-indexes. First, a permutation $\sigma$ of the set $\{1, \ldots, k+m\}$ is called a $(k, m)$-shuffle if $\sigma^{-1}(1) < \ldots < \sigma^{-1}(k)$ and $\sigma^{-1}(k+1) < \ldots < \sigma^{-1}(k+m)$. The list $(\sigma(1), \ldots, \sigma(k+m))$ is also called a shuffle of $(1, \ldots, k)$ and $(k+1, \ldots, k+m)$. Let Shuffles$(k, m)$ denote the collection of all $(k, m)$-shuffles.

**Definition 2** (Shuffle product)**.** Consider two multi-indexes $I = (i_1, \ldots, i_k)$ and $J = (j_1, \ldots, j_m)$ with $i_1, \ldots, i_k, j_1, \ldots, j_m \in \{1, \ldots, d\}$. Define the multi-index

$$(r_1, \ldots, r_k, r_{k+1}, \ldots r_{k+m}) = (i_1, \ldots, i_k, j_1, \ldots, j_m). \tag{1.39}$$

The shuffle product of $I$ and $J$, denoted $I \sqcup\!\sqcup J$, is a finite set of multi-indexes of length $k + m$ defined as follows

$$I \sqcup\!\sqcup J = \{(r_{\sigma(1)}, \ldots, r_{\sigma(k+m)}) \mid \sigma \in \text{Shuffles}(k, m)\}. \tag{1.40}$$

**Theorem 1** (Shuffle product identity)**.** *For a path $X : [a, b] \mapsto \mathbb{R}^d$ and two multi-indexes $I = (i_1, \ldots, i_k)$ and $J = (j_1, \ldots, j_m)$ with $i_1, \ldots, i_k, j_1, \ldots, j_m \in \{1, \ldots, d\}$, it holds that*

$$S(X)_{a,b}^I S(X)_{a,b}^J = \sum_{K \in I \sqcup\!\sqcup J} S(X)_{a,b}^K. \tag{1.41}$$

**Example 8.** To make things more clear, let us consider a simple example of a two-dimensional path $X : [a, b] \mapsto \mathbb{R}^2$. The shuffle product implies that

$$\begin{aligned} S(X)_{a,b}^1 S(X)_{a,b}^2 &= S(X)_{a,b}^{1,2} + S(X)_{a,b}^{2,1}, \\ S(X)_{a,b}^{1,2} S(X)_{a,b}^1 &= 2S(X)_{a,b}^{1,1,2} + S(X)_{a,b}^{1,2,1}. \end{aligned} \tag{1.42}$$

The shuffle product in particular implies that the product of two terms of the signature can be expressed as a linear combination of higher order terms. This fact will be useful for practical applications of the signature to regression analysis which will be further discussed in Chapter 2.

### 1.3.3 Chen's identity

We now describe a property of the signature known as Chen's identity, which provides an algebraic relationship between paths and their signatures. To formulate Chen's identity, we need to introduce the algebra of formal power series, which we have not mentioned thus far, but which should appear natural in light of the definition of the signature.

**Definition 3** (Formal power series)**.** Let $e_1, \dots, e_d$ be $d$ formal indeterminates. The algebra of (non-commuting) *formal power series* in $d$ indeterminates is the vector space of all series of the form

$$\sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} \lambda_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k}, \tag{1.43}$$

where the second summation runs over all multi-indexes $(i_1, \dots, i_k)$, $i_1, \dots, i_k \in \{1, \dots, d\}$, and $\lambda_{i_1,\dots,i_k}$ are real numbers.

A (non-commuting) formal polynomial is a formal power series for which only a finite number of coefficients $\lambda_{i_1,\dots,i_k}$ are non-zero. The terms $e_{i_1} \dots e_{i_k}$ are called monomials. The term corresponding to $k = 0$ is simply just a real number $\lambda_0$. The space of formal power series is often also called the *tensor algebra* of $\mathbb{R}^d$. We stress that the power series we consider are non-commutative; for example, the elements $e_1 e_2$ and $e_2 e_1$ are distinct.

Observe that the space of formal power series may be naturally equipped with a vector space structure by defining addition and scalar multiplication as

$$\left( \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} \lambda_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k} \right) + \left( \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} \mu_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k} \right)$$
$$= \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} (\lambda_{i_1,\dots,i_k} + \mu_{i_1,\dots,i_k}) e_{i_1} \dots e_{i_k} \tag{1.44}$$

and

$$c \left( \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} \lambda_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k} \right) = \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} c\lambda_{i_1,\dots,i_k} e_{i_1} \dots . e_{i_k}. \tag{1.45}$$

Moreover, one may define the product $\otimes$ between monomials by joining together multi-indexes

$$e_{i_1} \dots e_{i_k} \otimes e_{j_1} \dots e_{j_m} = e_{i_1} \dots e_{i_k} e_{j_1} \dots e_{j_m}. \tag{1.46}$$

The product $\otimes$ then extends uniquely and linearly to all power series. We demonstrate the first few terms of the product in the following expression

$$\left(\sum_{k=0}^{\infty}\sum_{i_1,\dots,i_k\in\{1,\dots d\}}\lambda_{i_1,\dots,i_k}e_{i_1}\dots e_{i_k}\right)\otimes\left(\sum_{k=0}^{\infty}\sum_{i_1,\dots,i_k\in\{1,\dots d\}}\mu_{i_1,\dots,i_k}e_{i_1}\dots e_{i_k}\right)$$
$$=\lambda_0\mu_0+\sum_{i=1}^{d}(\lambda_0\mu_i+\lambda_i\mu_0)e_i+\sum_{i,j=1}^{d}\left(\lambda_0\mu_{i,j}+\lambda_i\mu_j+\lambda_{i,j}\mu_0\right)e_ie_j+\dots. \quad (1.47)$$

The space of formal power series becomes an algebra when equipped with this vector space structure and the product $\otimes$.

The reader may have noticed that the indexing set of the monomials $e_{i_1}\dots e_{i_k}$ coincides with the indexing set of the terms of the signature of a path $X:[a,b]\mapsto\mathbb{R}^d$, namely the collection of all multi-indexes $(i_1,\dots,i_k)$, $i_1,\dots,i_k\in\{1,\dots,d\}$. It follows that a convenient way to express the signature of $X$ is by a formal power series where the coefficient of each monomial $e_{i_1}\dots e_{i_k}$ is defined to be $S(X)_{a,b}^{i_1,\dots,i_k}$. We use the same symbol $S(X)_{a,b}$ to denote this representation

$$S(X)_{a,b}=\sum_{k=0}^{\infty}\sum_{i_1,\dots,i_k\in\{1,\dots d\}}S(X)_{a,b}^{i_1,\dots,i_k}e_{i_1}\dots e_{i_k}, \quad (1.48)$$

where, as before, we set the "zero-th" level of the signature $S(X)_{a,b}^0=1$ (corresponding to $k=0$).

To state Chen's identity, it remains to define the concatenations of paths.

**Definition 4** (Concatenation)**.** For two paths $X:[a,b]\mapsto\mathbb{R}^d$ and $Y:[b,c]\mapsto\mathbb{R}^d$, we define their *concatenation* as the path $X*Y:[a,c]\mapsto\mathbb{R}^d$ for which $(X*Y)_t=X_t$ for $t\in[a,b]$ and $(X*Y)_t=X_b+(Y_t-Y_b)$ for $t\in[b,c]$.

Chen's identity informally states that the signature turns the "concatenation product" $*$ into the product $\otimes$. More precisely, we have the following result.

**Theorem 2** (Chen's identity)**.** *Let $X:[a,b]\mapsto\mathbb{R}^d$ and $Y:[b,c]\mapsto\mathbb{R}^d$ be two paths. Then*

$$S(X*Y)_{a,c}=S(X)_{a,b}\otimes S(Y)_{b,c}. \quad (1.49)$$

### 1.3.4 Time-reversal

The time-reversal property informally states that the signature $S(X)_{a,b}$ of a path $X:[a,b]\mapsto\mathbb{R}^d$ is precisely the inverse under the product $\otimes$ of the signature obtained from running $X$ backwards in time. To make this precise, we make the following definition.

**Definition 5** (Time-reversal)**.** For a path $X:[a,b]\mapsto\mathbb{R}^d$, we define its time-reversal as the path $\overleftarrow{X}:[a,b]\mapsto\mathbb{R}^d$ for which $\overleftarrow{X}_t=X_{a+b-t}$ for all $t\in[a,b]$.

**Theorem 3** (Time-reversed signature)**.** *For a path $X : [a, b] \mapsto \mathbb{R}^d$, it holds that*

$$S(X)_{a,b} \otimes S(\overleftarrow{X})_{a,b} = 1. \tag{1.50}$$

The element 1 in the above expression should be understood as the formal power series where $\lambda_0 = 1$ and $\lambda_{i_1,\dots,i_k} = 0$ for all $k \geq 1$ and $i_1, \dots, i_k \in \{1, \dots, d\}$, which is the identity element under the product $\otimes$.

### 1.3.5 Log signature

We now define a transform of the path signature called the log signature. The log signature essentially corresponds to taking the formal logarithm of the signature in the algebra of formal power series.

To this end, for a power series

$$x = \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k \in \{1,\dots d\}} \lambda_{i_1,\dots,i_k} e_{i_1} \dots e_{i_k} \tag{1.51}$$

for which $\lambda_0 > 0$, define its logarithm as the power series given by

$$\log x = \log(\lambda_0) + \sum_{n \geq 1} \frac{(-1)^n}{n} \left(1 - \frac{x}{\lambda_0}\right)^{\otimes n}, \tag{1.52}$$

where $\otimes n$ denotes the $n$-th power with respect to the product $\otimes$.

For example, for a real number $\lambda \in \mathbb{R}$ and the series

$$x = 1 + \sum_{k \geq 1} \frac{\lambda^k}{k!} e_1^{\otimes k}, \tag{1.53}$$

one can readily check that

$$\log x = \lambda e_1. \tag{1.54}$$

Observe that, in general, $\log x$ is a series with an infinite number of terms, however for every multi-index $(i_1, \dots, i_k)$, the coefficient of $e_{i_1} \dots e_{i_k}$ in $\log x$ depends only on the coefficients of $x$ of the form $\lambda_{j_1,\dots,j_m}$ with $m \leq k$, of which there are only finitely many, so that $\log x$ is well-defined without the need to consider convergence of infinite series.

**Definition 6** (Log signature)**.** For a path $X : [a, b] \mapsto \mathbb{R}^d$, the log signature of $X$ is defined as the formal power series $\log S(X)_{a,b}$.

For two formal power series $x$ and $y$, let us define their Lie bracket by

$$[x, y] = x \otimes y - y \otimes x. \tag{1.55}$$

A direct computation shows that the first few terms of the log signature are given by

$$\log S(X)_{a,b} = \sum_{i=1}^{d} S(X)^{i}_{a,b} e_i + \sum_{1 \le i < j \le d} \frac{1}{2} \left( S(X)^{i,j}_{a,b} - S(X)^{j,i}_{a,b} \right) [e_i, e_j] + \dots . \tag{1.56}$$

In particular one can see that the coefficient of the polynomials $[e_i, e_j]$ in the log signature is precisely the Lévy area introduced in Section 1.2.4.

**Example 9.** Consider the two-dimensional path

$$X : [0,2] \mapsto \mathbb{R}^2, \quad X_t = \begin{cases} \{t, 0\} & \text{if } t \in [0,1], \\ \{1, t-1\} & \text{if } t \in [1,2]. \end{cases} \tag{1.57}$$

Note that $X$ is the concatenation of the two linear paths, $Y : [0,1] \mapsto \mathbb{R}^2$, $Y_t = \{t, 0\}$, and $Z : [1,2] \mapsto \mathbb{R}^2$, $Z_t \mapsto \{0, t-1\}$. One can readily check that the signatures of $Y$ and $Z$ (as formal power series) are given by

$$S(Y)_{0,1} = 1 + \sum_{k \ge 1} \frac{1}{k!} e_1^{\otimes k}, \quad S(Z)_{1,2} = 1 + \sum_{k \ge 1} \frac{1}{k!} e_2^{\otimes k}. \tag{1.58}$$

It follows by Chen's identity that

$$S(X)_{0,2} = S(Y)_{0,1} \otimes S(Z)_{1,2} = 1 + e_1 + e_2 + \frac{1}{2!} e_1 + \frac{1}{2!} e_2 + e_1 e_2 + \dots . \tag{1.59}$$

Hence the first few terms of the log signature of $X$ are

$$\log S(X)_{0,2} = e_1 + e_2 + \frac{1}{2}[e_1, e_2] + \dots . \tag{1.60}$$

In fact, one can readily check that coefficients of $\log S(X)_{0,2}$ in the above example are given precisely by the classical Campbell-Baker-Hausdorff formula.

The above example in fact demonstrates the general fact that the log signature can always be expressed as a power series composed entirely of so-called Lie polynomials. This is the content of the following theorem due to Chen [4], which generalises the Campbell-Baker-Hausdorff theorem.

**Theorem 4.** *Let $X : [a,b] \mapsto \mathbb{R}^d$ be a path. Then there exist real numbers $\lambda_{i_1,\dots,i_k}$ such that*

$$\log S(X)_{a,b} = \sum_{k \ge 1} \sum_{i_1,\dots,i_k \in \{1,\dots,d\}} \lambda_{i_1,\dots,i_k} [e_{i_1}, [e_{i_2}, \dots, [e_{i_{k-1}}, e_{i_k}] \dots]]. \tag{1.61}$$

Note that the coefficients $\lambda_{i_1,\dots,i_k}$ are in general not unique since the polynomials of the form $[e_{i_1}, [e_{i_2}, \dots, [e_{i_{k-1}}, e_{i_k}] \dots]]$ are not linearly independent (e.g., $[e_1, e_2] = -[e_2, e_1]$).

## 1.4 Relation with rough paths and path uniqueness

We conclude the first part of these notes with a brief discussion about the role of the signature in the theory of rough paths and the extent to which the signature is able to determine the underlying path. These topics are substantially more involved than the basic properties of the signature discussed above, and so we only offer an informal discussion.

### 1.4.1 Rough paths

Our discussion of the signature has so far been restricted to paths which are piecewise differentiable (or more generally of bounded variation). This restriction was needed to ensure that the iterated integrals of the path existed as Riemann-Stieltjes integrals. More generally, one can define the iterated integrals of a path using the Young integral for any path of finite $p$-variation with $1 \leq p < 2$. The Young integral goes beyond the "classical" definition of the integral and is already able to cover a class of paths substantially more irregular than those of bounded variation.

A problem that arises for paths of infinite $p$-variation for all $p < 2$ (which is a situation of great interest in stochastic analysis due to the fact that the sample paths of Brownian motion have almost surely finite $p$-variation if and only if $p > 2$), is that one can show there is no well-defined notion of an iterated integral for such paths. We stress that this is not due to any technical limitation of the Young integral, but rather due to the fact that there is no unique candidate for the iterated integrals. That is to say, there is not necessarily only one unique way to define the iterated integrals.

One of the key observations of T. Lyons in his introduction of rough paths in [19] was that if one *defines* the first $\lfloor p \rfloor$ iterated integrals of a path $X$ of finite $p$-variation, then there is indeed a unique way to obtain all the other iterated integrals, and hence the signature of $X$. This notion of a path of finite $p$-variation, along with its first $\lfloor p \rfloor$ iterated integrals (which may be defined arbitrarily provided that they satisfy Chen's identity and possess finite $p$-variation), is precisely the definition of a $p$-rough path.

A key feature of the theory of rough paths, known now as the *universal limit theorem*, is that one is able to give meaning to controlled differential equations where the driver belongs to a special class of $p$-rough paths (know as geometric $p$-rough paths), and where the solution depends in a continuous way on the driver provided that the space of $p$-rough paths is equipped with a suitable topology (known as the $p$-variation metric).

For the reader interested, we strongly recommend the St. Flour lecture notes of Lyons, Caruana, and Lévy [20] for an introduction to the theory of rough paths, and the monograph of Friz and Hairer [9] for a more recent treatment of the topic and an introduction to Hairer's emerging theory of regularity structures in the study of stochastic partial differential equations.

### 1.4.2 Path uniqueness

As discussed in Section 1.2.3, the signature of a path $X : [a, b] \mapsto \mathbb{R}^d$ is all that is needed to determine the endpoint of the solution to a linear (and, less trivially, non-linear) differential equation driven by $X$, which was first shown by Chen [5] in the smooth setting, and later extended by Hambly and Lyons [14] and Boedihardjo et al. [2] to less regular paths. The works of these authors in fact shows that the signature captures deep geometric properties of a path, which we briefly discuss here.

A natural question one may ask is the following: is a path completely determined by its signature? In light of the time-reversal property and Chen's identity, as well as the invariance of the signature under time reparametrizations, the answer, in general, is no. For example, one can never recover from the signature the exact speed at which the path is traversed (due to invariance under time reparametrizations), nor can one tell apart the signature of a trivial constant path and that of a path concatenated with its time-reversal.

However, a far less elementary fact is that this is essentially the only information one loses from the signature. For example, for a path $X$ which never crosses itself, the signature is able to completely describe the image and direction of traversal of the path (that is, all the points that $X$ visits and the order in which it visits them). This demonstrates the signature's ability to completely determine the geometric properties of a path which does not possess degeneracies of a certain kind consisting of movements going directly back onto itself (this is made precise using the notion of a *tree-like path* introduced in [14]).

We emphasise however that the question of how one may effectively recover various properties of a path from its signature currently remains a challenging area of research. For recent progress on this topic, see Lyons and Xu [17, 18]) and Geng [11].

## 2 Practical Applications

One of the practical applications of the signature transformation lies in the field of machine learning algorithms. As it has already been discussed, the signature summarizes important information about a path. When the path is composed of a sequential data stream $\{X_i\}$, the terms of the signature $S(X)^{ijk\cdots}$ are good candidates for characteristic *features* of the data stream. The shuffle product property allows us to represent a non-linear function of the signature as a linear combination of iterated integrals. That is similar to basis function expansion and naturally applies to computations of regression models. In the next sections we will demonstrate numerical computations of signatures of paths from various data streams and applications of signatures to machine learning problems, regression and classification. Recently, the method of kernelization of the signature has been proposed [15] and applied to hand movement classification problem [24] from the UCI repository for machine learning.

## 2.1 Elementary operations with signature transformation

In the following sections we introduce the essential ingredients of the signature method and demonstrate how to perform basic computations with the signature.

### 2.1.1 Paths from discrete data

We start with basic computations of the signature applied to synthetic data streams. Consider three one-dimensional sequences of length four:

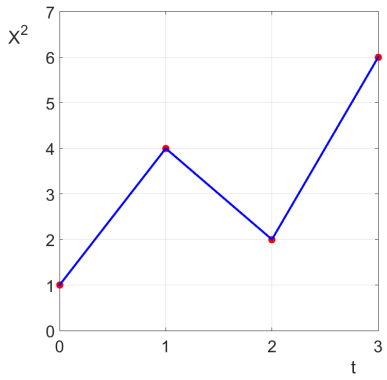
$$\{X_i^1\}_{i=1}^4 = \{1, 3, 5, 8\} \tag{2.1}$$
$$\{X_i^2\}_{i=1}^4 = \{1, 4, 2, 6\} \tag{2.2}$$
$$\{t_i\}_{i=1}^4 = \{0, 1, 2, 3\}, \tag{2.3}$$

where the variable $\{t_i\}$ corresponds to the ordering of the terms in (2.1)-(2.2) and is normally interpreted as *time*. We are interested in transforming this discrete series into a continuous function - *a path*. Among various ways to find this transformation we focus on two main approaches:

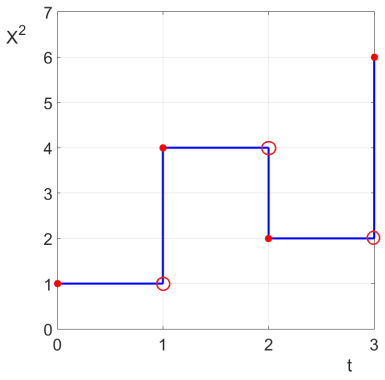
(a) piece-wise linear interpolation,

(b) rectilinear interpolation (i.e. *axis path*).

These two methods are present in Fig. 6 for the 2-dim path comprised of (2.2) and (2.3). The explicit parametrisations in Cartesian coordinates for the two methods read as (2.4)




(a) Piece-wise linear interpolation of $\{t_i, X_i^2\}$.




(b) Rectilinear interpolation of $\{t_i, X_i^2\}$.

**Figure 6**: Examples of two different interpolations.

and (2.5)

$$\{t_i, X_i^2\} = \{(0, 1), (1, 4), (2, 2), (3, 6)\} \tag{2.4}$$
$$\{t_i, X_i^2\} = \{(0, 1), (1, 1), (1, 4), (2, 4), (2, 2), (3, 2), (3, 6)\}, \tag{2.5}$$

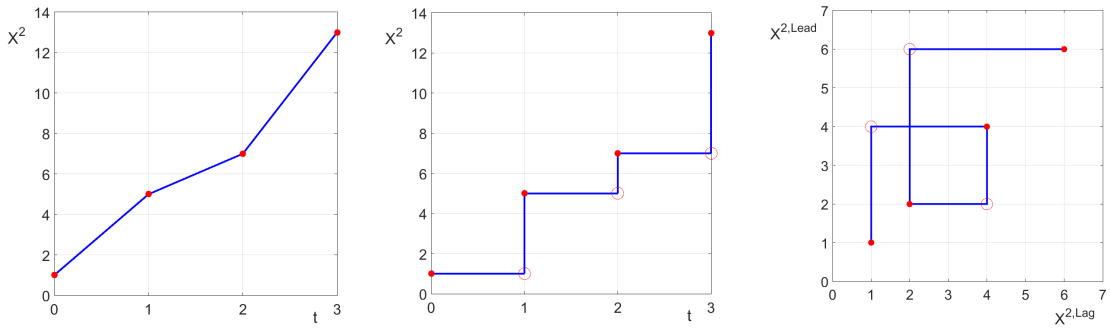
with three auxiliary points $\{(1,1),(2,4),(3,2)\}$ added to construct the rectilinear path in Fig. 6b denoted by empty red circles. For various numerical applications, the original data might be mapped into different forms to exhibit its structure. One of the examples of such a mapping is the *cumulative sum*, or sequence of partial sums. More precisely the cumulative sum is defined as:

$$CS(\{X_i\}_{i=1}^n) = \{X_1, X_1 + X_2, \ldots, S_k, \ldots, S_n\}\,; \quad S_k = \sum_{i=1}^{k} X_i. \tag{2.6}$$

Using again the previous example (2.2), (2.3), a new path is:

$$\begin{aligned} \{\tilde{X^2}\} = CS(X^2) &= \{1, 5, 7, 13\} \\ \{t\} &= \{0, 1, 2, 3\}, \end{aligned} \tag{2.7}$$

and applying the two different interpolations mentioned above, the paths are depicted in Fig. 7, with added auxiliary points denoted by the empty red circles as in previous case.




(a) Piece-wise linear interpolation of $CS(X^2)$.

(b) Rectilinear interpolation of $CS(X^2)$.

(c) Lead-Lag transform of one-dimensional data $\{X_i^2\}$.

**Figure 7**: Examples of two different interpolations of $CS(X^2)$ (a)-(b) and Lead-Lag transform of $\{X_i^2\}$ (c).

### 2.1.2 The lead-lag transformation

Another very important and interesting embedding is the *Lead-Lag* transformation of data, which maps a one-dimensional path into a two-dimensional path. Considering the sequence (2.2), the Lead-Lag mapping is given by:

$$LeadLag\ :\ X^2 = \{1,2,4,6\} \mapsto \begin{cases} X^{2,Lead} &= \{1,4,4,2,2,6,6\} \\ X^{2,Lag} &= \{1,1,4,4,2,2,6\} \end{cases} \tag{2.8}$$

and the resulting embedded path is presented in Fig. 7c with three additional points $\{(1,4),(4,2),(2,6)\}$.

Consider two processes $\{X_t\}$ and $\{Y_t\}$ which follow each other's paths with a small time lag:

$$X_{t+k} \propto Y_t, \quad k > 0, \tag{2.9}$$

giving the name *Lead-Lag process*. Certain properties of the lead-lag process are easily captured by the signature of the path comprised of these processes $\{X_t, Y_t\}$ [8]. This is demonstrated in the following example. Consider a path in $\mathbb{R}^2$ given by $X = \{X^1, X^2\}$. If an increase (resp. decrease) of the component $X^1$ is followed by an increase (resp. decrease) in the component $X^2$, then the area $A$ given by (1.35) is positive. If the relative moves of the components $X^1$ and $X^2$ are in the opposite direction, then the area $A$ is negative. In Fig.8 we can see how the area changes as we increase the endpoint.

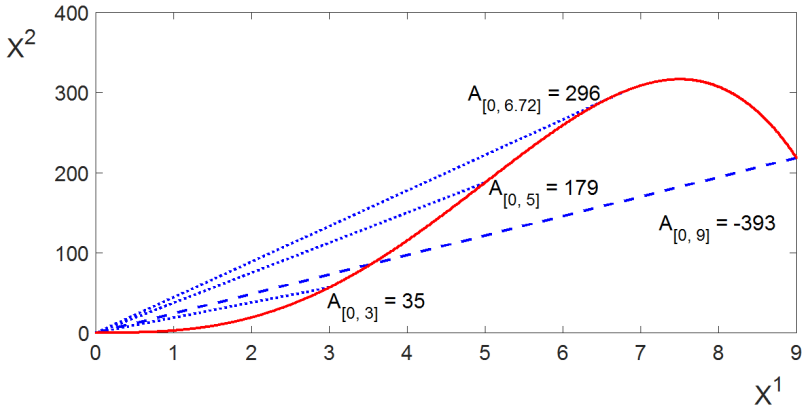



**Figure 8**: Example of the change of the area enclosed by the curve and the chord at different endpoints.

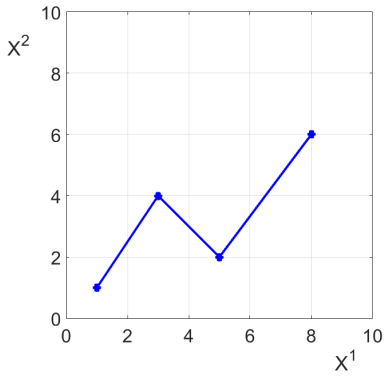
The notation here is: $A_{[X_i^1, X_f^1]}$ area between two points, where $(i, f)$ are *initial* and *final* endpoints along the direction of $X^1$. Moving at first from left to right (increasing the parameter $t$), $X^1$ and $X^2$ both increase, resulting in the increasing area $A_{ij}$: $A_{[0,3]} = 35$, $A_{[0,5]} = 179$. At some point ($X^1 = 6.72$) the area attains its maximum value $A_{[0,6.72]} = 296$ and after this point the area decreases as we increase the value of $X^1$. The total area between the two endpoints is $A_{[0,9]} = -393$.

The advantage of using the signature method is that any multivariate distribution of data could be represented as a path in a high-dimensional space $\mathbb{R}^d$. For example, using the data (2.1)-(2.3) we can construct paths in $\mathbb{R}^2$ and $\mathbb{R}^3$ spaces, presented in Fig. 9a and Fig. 9b respectively.

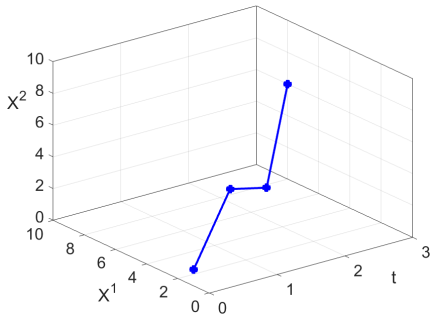
Another example of a multidimensional embedding involves the lead-lag transform. Here we include the time vector $t = \{0, 1, 2, 3, 4\}$ with $X = \{0, 1, 3, 5, 2\}$ and compute the lead-lag transform of $X$. The red dots represent the actual data points of the Cartesian ordered tuple:

$$\{(t^{lead}, X^{lead}, X^{lag})_i\} \;=\; \{(0,0,0), (1,1,1), (2,3,3), (3,5,5), (4,2,2)\}. \tag{2.10}$$

Applying the axis-path interpolation, we get the path presented in Fig. 10.




(a) 2-dim path from two 1-dim paths $\{X_i^1, X_i^2\}$.




(b) 3-dim path from three 1-dim paths $\{t_i, X_i^1, X_i^2\}$.

**Figure 9**: Examples of embedding a collection of one-dimensional paths into a single multi-dimensional path.

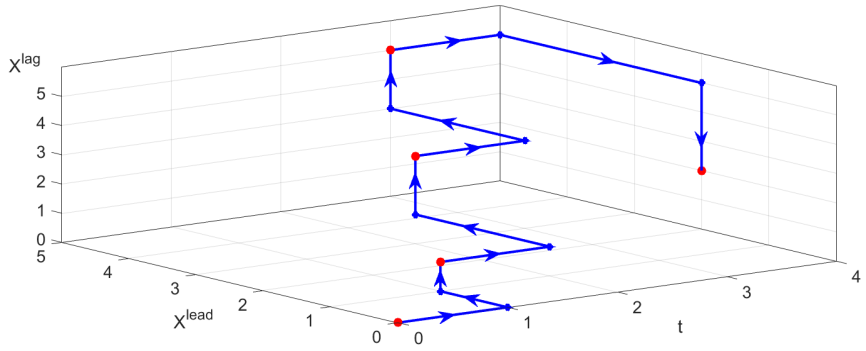



**Figure 10**: Example of a three dimensional lead-lag transform of the data $\{(t^{lead}, X^{lead}, X^{lag})_i\}$. The path is build by means of sequential increments along each Cartesian direction. This path allows to study the variance of $\{X\}$ as well as its time-dependent properties.

### 2.1.3 The signature of paths

For a given path, one can compute its signature according to the rules and definition in the Sec. 1. We will present computations of signatures for various embeddings as presented in Figs. 6-9 and discuss their properties.

We start with the embedded path presented in Fig. 9(a), for which the total increment and the signed area are depicted in Fig. 11.

Computing the signature and the log signature of this path up to level $L = 2$ gives:

$$S(X) = (1, 7, 5, 24.5, 19, 16, 12.5) = (1, S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)}) \quad (2.11)$$

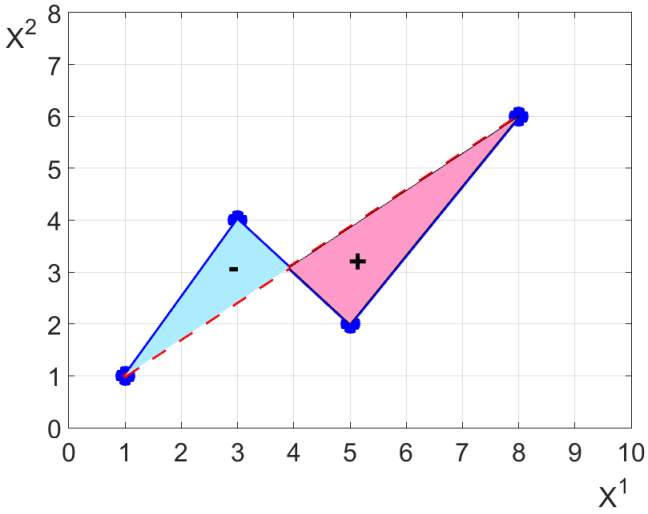
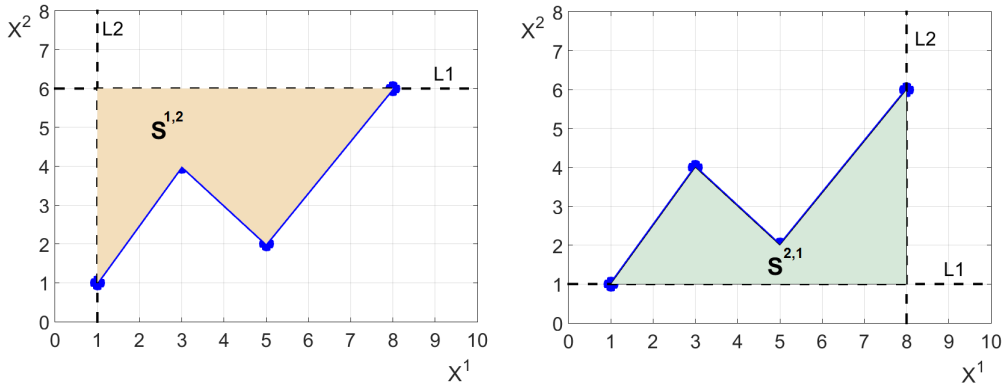



**Figure 11**: Example of signed area enclosed by the piece-wise linear path (blue) and the chord (red dashed line). The light blue area is negative and pink area is positive.

and

$$\log S(X) = (7, 5, 1.5) = (S^{(1)}, S^{(2)}, S^{[1,2]}),$$

where the last term in $\log S(X)$ is given by $\frac{1}{2}(S^{(1,2)} - S^{(2,1)})$ and corresponds to the total area between the endpoints. The fact that it is positive means that the *pink* area is larger than the *light blue* one. The geometric interpretation of the second order terms $S^{(1,2)}$ and $S^{(2,1)}$ are presented in Fig. 12.




(a) Area given by the term $S^{(1,2)} = 19$. (b) Area given by the term $S^{(2,1)} = 16$.

**Figure 12**: The geometric meaning of the terms $S^{(1,2)}$ and $S^{(2,1)}$. The left panel (a) represents the area enclosed by the path and two perpendicular dashed lines passing through the endpoints of the path, while the right panel (b) shows another possibility for the area to be enclosed by the path and two perpendicular dashed lines passing through the endpoints.

Switching the order of integration over the path in the terms $S^{(1,2)}$ and $S^{(2,1)}$ gives rise to

two areas which complete each other and add up to the total area of a rectangular with side lengths $X^1$ and $X^2$. This simple geometrical meaning is nothing but the shuffle product relation:

$$\begin{aligned} S^{(1)} \cdot S^{(2)} &= S^{(1,2)} + S^{(2,1)} \\ 5 \cdot 7 &= 19 + 16. \end{aligned} \tag{2.12}$$

The geometric interpretation of the higher order terms is less intuitive and we omit this discussion.

### 2.1.4 Rules for the sign of the enclosed area

Before we continue, it is important to emphasize the significance of the direction in which we travel along a path and the sign of the area which it encloses. Consider the six possibilities in Fig. 13

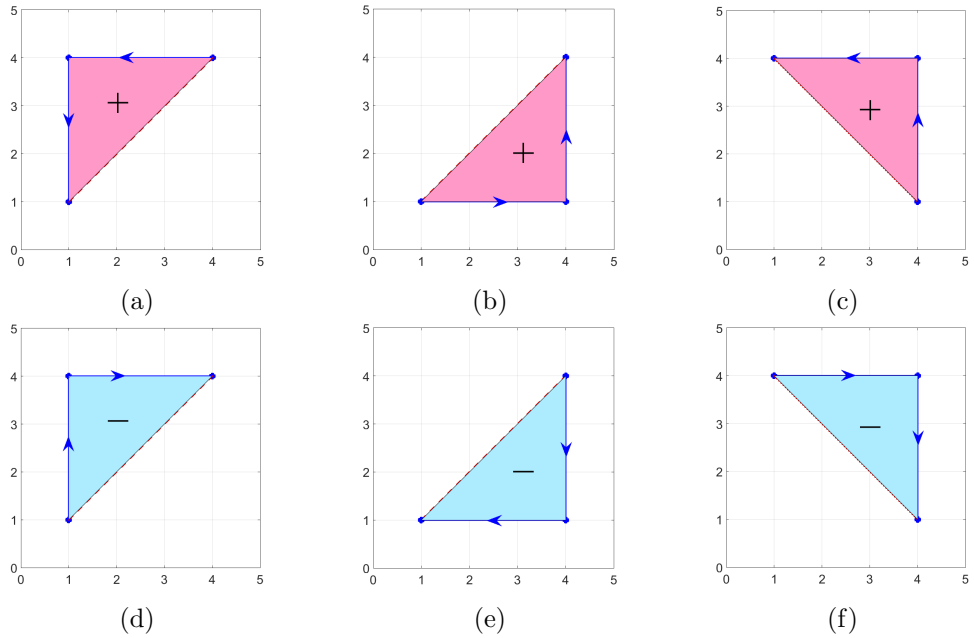



**Figure 13**: Various possibilities of signed area. The first row (a)-(c) corresponds to counter clock-wise movement along the path, and the bottom row (d)-(f) to clock-wise movement.

where we clearly see how direction of movement along the axis paths corresponds to the sign of the enclosed area. This is intimately related to the sign of the winding number of the path [3].

### 2.1.5 Statistical moments from the signature

Having established the basic rules we are ready to explore more important properties of the signature transformation. Combining several data streams $X = \{X_i\}$ into a single one

allows us to compute statistical moments and cross correlations between the individual streams. In the next sections we present a close relation between the statistical moments of a set of data $X$ and the terms of signature $S(X)^I$.

### Relationship between the lead-lag transformation and the variance of data

Consider the lead-lag path Fig. 7c constructed from (2.1). We can decompose this figure into three right-angled isosceles triangles Fig. 14. Note the direction of movement along this path, starting from the point $X_1^2$ and moving towards the end point $X_4^2$ corresponds to a negative sign, thus the total area will be negative. All three triangles have the same direction of movement resulting in the sum of their individual contributions.

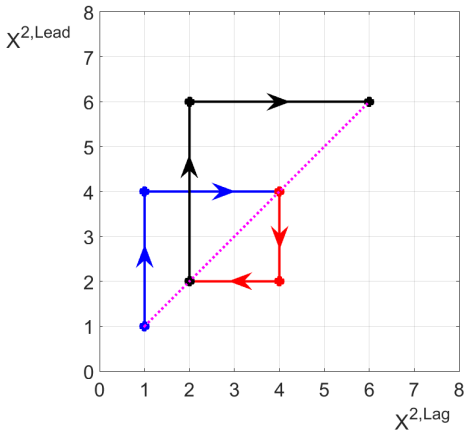



**Figure 14**: Decomposition of the full lead-lag path into individual parts. The direction of movement along the path is shown by arrows.

The absolute value of the total area is then given by:

$$\begin{aligned} |A| &= \frac{1}{2}\left[\left(X_2^2 - X_1^2\right)\left(X_2^2 - X_1^2\right) + \right. \\ &+ \left(X_3^2 - X_2^2\right)\left(X_3^2 - X_2^2\right) + \\ &+ \left.\left(X_4^2 - X_3^2\right)\left(X_4^2 - X_3^2\right)\right] \\ &= \frac{1}{2}\left[(4-1)^2 + (2-4)^2 + (6-2)^2\right]. \end{aligned} \tag{2.13}$$

Let us write:

$$QV(X) = \sum_{i}^{N-1} \left(X_{i+1} - X_i\right)^2, \tag{2.14}$$

which has the simple meaning of the *quadratic variation* of the path constructed from $\{X_i\}_{i=1}^N$ and is related to the *variance* of the path. Thus one can generally write for any sequence $\{X_i\}_{i=1}^N$

$$A_{Lead-Lag} = \frac{1}{2} QV(X). \tag{2.15}$$

The first order terms of the signature correspond to the total increments in each dimension, which are the same and equal to:

$$\Delta X^{2,Lead} = \Delta X^{2,Lag} = X_4^2 - X_1^2 = 6 - 1 = 5. \tag{2.16}$$

Putting all the terms together and omitting all unessential notation, we obtain the truncated log signature of $\{X^2\}$ from (2.2) at level $L = 2$:

$$\log S\left(X^2\right) = \left(\Delta X^2, \Delta X^2, \frac{1}{2} QV(X^2)\right) = (5, 5, -14.5). \tag{2.17}$$

The only reason we are using the log signature above to present our result is because of its compactness and simplicity. One can easily rewrite the above result in terms of the signature with all the terms included:

$$S(X^2) = (1, 5, 5, 12.5, -2, 27, 12.5). \tag{2.18}$$

We are free to work with either of these two representations of the signature, but certain applications will dictate the rationale behind a particular choice.

### The cumulative sum of a sequence

Next we explore certain properties of paths which originate from embedding points using cumulative sums. Consider again the example (2.2). According to (2.6), the cumulative sum of a sequence is:

$$\begin{aligned} \tilde{X^2} &= \{X_1^2, X_1^2 + X_2^2, X_1^2 + X_2^2 + X_3^2, X_1^2 + X_2^2 + X_3^2 + X_4^2\} \\ &= \{1, 5, 7, 13\}. \end{aligned} \tag{2.19}$$

Pairing the above with the time component (2.3), we get a 2-dim path as shown in Fig. 7b. Augmenting the original series $\{X_i\}$ with the zero value and truncating the signature of the new series $\{\tilde{X}\}$ at level $L$ will determine the statistical moments up to level $L$ of the original sequence. Explicitly, for any general sequence $\{X_i\}$:

$$\{X\}_i \to \{0, \{X\}_{i=1}^N\} \to CS(\{X\}_i) = \{\tilde{X}\}_{i=0}^N = \{0, X_1, X_1 + X_2, \dots\}. \tag{2.20}$$

One should pay attention to the indexing of the sequence, $i = 1 \to i = 0$. As we have learnt from (2.13)-(2.17), we can apply the lead-lag transform to the augmented sequence, compute the truncated signature at level $L = 2$, and we will obtain quantities which are proportional to the first two statistical moments of the data points, namely *mean* and *variance*. For an arbitrary series $\{X\}_{i=1}^N$ and its cumulative sum representation $\{\tilde{X}\}_{i=1}^N$:

$$\Delta\tilde{X} = \sum_{i=1}^{N} X_i \tag{2.21}$$

$$QV(\tilde{X}) = \sum_{i=0}^{N-1} \left(\tilde{X}_{i+1} - \tilde{X}_i\right)^2 = \sum_{i=1}^{N} (X_i)^2 \tag{2.22}$$

It is easy to see that these expressions are related to the *mean* and *variance* of $X$ if one makes use of the expected value. Namely, for a collection of data points $\{X_i\}_{i=1}^N$:

$$\begin{aligned} Mean(X) &= E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{\Delta\tilde{X}}{N} \quad \text{(uniformly distributed data)} \\ Var(X) &= E[(X - E[X])^2] = E[X^2] - (E[X])^2 = \frac{1}{N}\left(QV(\tilde{X}) - \frac{1}{N}(\Delta\tilde{X})^2\right) \end{aligned}$$

Recall that the first two levels of the signature (log or full) correspond to the total increment and the Lévy area, which shows how these signature terms determine the first two statistical moments. Similarly, it is straightforward to demonstrate that the higher statistical moments can be obtained from the higher order signature terms.

### Signature terms as a function of data points

To make a closer connection and develop further the intuition behind the first terms of the signature of a data stream, we shall present two different examples of an embedding and the related signature terms. First, consider the cumulative lead-lag embedding of $N$ data points $\{X_i\}_{i=1}^N$ (according to (2.20)) into a continuous path [cf. Fig. 7c], then the resulting truncated signature at level $L = 2$ is simply given by:

$$S(\tilde{X})|_{L=2} = \left(1, S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)}\right) \tag{2.23}$$

with

$$\begin{aligned} S^{(1)} = S^{(2)} &= \sum_{i}^{N} X_i \\ S^{(1,1)} = S^{(2,2)} &= \frac{1}{2}\left(\sum_{i}^{N} X_i\right)^2 \\ S^{(1,2)} &= \frac{1}{2}\left[\left(\sum_{i}^{N} X_i\right)^2 + \sum_{i}^{N} X_i^2\right] \\ S^{(2,1)} &= \frac{1}{2}\left[\left(\sum_{i}^{N} X_i\right)^2 - \sum_{i}^{N} X_i^2\right]. \end{aligned} \tag{2.24}$$

Note that, although we computed the signature of the transformed data $\{\tilde{X}\}$, the final result is given in terms of the original untransformed data $\{X_i\}$.

Next, we consider the untransformed original data $\{X_i\}$ with the same lead-lag embedding which results in the following expression for the signature:

$$S(X)|_{L=2} = \left(1, S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)}\right) \tag{2.25}$$

with

$$
\begin{aligned}
S^{(1)} &= S^{(2)} &&= \sum_{i}^{N-1} (X_{i+1} - X_i) \\
S^{(1,1)} &= S^{(2,2)} &&= \frac{1}{2}\left(\sum_{i}^{N-1} (X_{i+1} - X_i)\right)^2 \\
& & S^{(1,2)} &= \frac{1}{2}\left[\left(\sum_{i}^{N-1} (X_{i+1} - X_i)\right)^2 + \sum_{i}^{N-1} (X_{i+1} - X_i)\right] \\
& & S^{(2,1)} &= \frac{1}{2}\left[\left(\sum_{i}^{N-1} (X_{i+1} - X_i)\right)^2 - \sum_{i}^{N-1} (X_{i+1} - X_i)\right].
\end{aligned}
\tag{2.26}
$$

Comparing (2.24) with (2.26), one can immediately see how the cumulative sum of the terms affects the result. This observation is crucial for further applications of the signature approach, since each individual problem should be treated in the most suitable way. As an illustrative example, one can derive the empirical sample mean and sample variance from the signature terms. Considering data $\{X_i\}_{i=1}^{N}$ with the cumulative sum and the lead-lag embedding (as demonstrated in (2.24)), a bit of algebra brings us to:

$$
\begin{aligned}
Mean(X) &= \frac{1}{N} S^{(1)} \\
Var(X) &= -\frac{N+1}{N^2} S^{(1,2)} + \frac{N-1}{N^2} S^{(2,1)},
\end{aligned}
\tag{2.27}
$$

where $N$ is the total number of data points in the sample. The possible caveat here is the degeneracy in the terms of the signature, causing this representation not to be unique and introducing a problem of colinearity of the signature terms. One of the standard approaches to resolve this problem is the shrinkage techniques, such as the LASSO [25], ridge or their hybrid combination known as the elastic net regularization [27].

## 2.2 The signature in machine learning

The main idea of using the signature transformation for machine learning problems is motivated by its ability to extract characteristic features from data. As already discussed, embedding data into a path and computing its signature provides us with important information about the original data. The workflow is simple and is summarised in the following algorithm:

$$
data \rightarrow path \rightarrow signature\ of\ path \rightarrow features\ of\ data
$$

This algorithm is absolutely general and works for any type of sequential data which can be embedded into a continuous path. The extracted features might be used for various types of machine learning applications, including both supervised and unsupervised learning. For example, one can classify time-series or distinguish clusters of data. One of the advantages of feature extraction with the signature method is that the signature is sensitive to the geometric shape of a path. Sensitivity to the geometric shape of the input

data has lead to a successful application of the signature method to Chinese character recognition problem [12]. One of the most known and natural applications of the signature method is in quantitative finance, namely analysis of time-series data [7, 16]. Time-series data represent ordered sequential data which is an ideal candidate for creating a path from data, followed by computing the signature and applying machine learning algorithms for further analysis. Any type of time-ordered sequential data naturally fits into the signature framework. Moreover, if the input data come from several parallel sources, this will result in a multi-dimensional path. An example for such a type of data is *panel* data (in Econometrics) or *longitudinal* data (in Medicine, Psychology, Biostatistics etc.) which involve repeated observations of the same quantity over periods of time.

Concluding this brief introduction to applications of the path signature method to machine learning problems, we emphasise that this novel method introduces a new concept of dealing with data: thinking of data as geometric paths and using the path signature method in data analysis. In the following sections we will elaborate on these statements and demonstrate practical applications of the signature method.

### 2.2.1 Application of the signature method to data streams

In this section we demonstrate an explicit example of an application of the signature method to classification of time-series. This method is moreover easily applied to analysis of sequential data of any type. In the next section we overview some of the recent applications of the signature method to various problem in machine learning.

We aim at learning a classifier which discriminates between two types of univariate time-series, synthetically simulated using the $ARMA(p,q)$ model. The difference between time-series is encoded in different value of the model parameters. Consider a collection of observations of time-series $\{Y_{i,j}\}$ labeled by two indices $i, j$, where the first index $i$ denotes each distinct time-series and the second index $j$ accounts for the measurement at time $t_j$.

The $ARMA(p,q)$ model is **A**uto**R**egressive **M**oving **A**verage model with parameters $p$ (the order of the autoregressive model) and $q$ (the order of the moving-average model). This is a superposition of the autoregressive $AR(p)$ and the moving-average $MA(q)$ models:

$$\begin{aligned} AR(p) &: Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t_2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \\ MA(q) &: Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \end{aligned} \tag{2.28}$$

where $\mu$ is the mean of the series, $\theta_i$ with $\phi_i$ are the parameters of the models and $\epsilon_{t-i}$ is the white noise error term. We are not diving into the mathematical foundation of the model, but rather focus on its application. To demonstrate how the feature extraction method works in this case, we create a binary classification problem using the ARMA model. We generate 1000 distinct time-series (500 time-series of each class) of length 100. In this particular, the case two classes correspond to the following models:

$$\begin{aligned} \textit{class "0"} &: Y_t - 0.4 Y_{t-1} = 0.5 + \epsilon_t + 0.5 \epsilon_{t-1} \\ \textit{class "1"} &: Y_t - 0.8 Y_{t-1} = 0.5 + \epsilon_t + 0.7 \epsilon_{t-1} \end{aligned} \tag{2.29}$$

where $\epsilon_t$ is normally distributed with zero mean and unit variance. An example of a time-series from each class is depicted in Fig. 15. To give a simple intuitive picture, the resulting

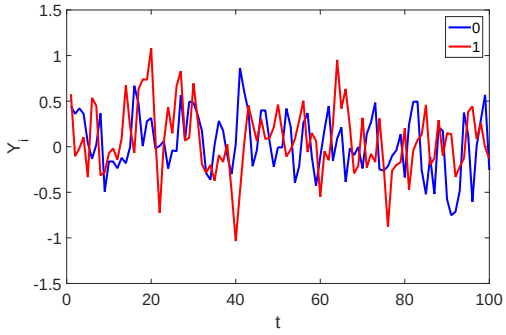



**Figure 15**: Example of time-series $Y_i$ from two distinct classes generated from ARMA(1,1) model specified by (2.29).

two-classes time-series data $\{Y_i\}$ and their class labels are organised in a matrix form:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $\dots$ | $t_{100}$ | Class |
|---|---|---|---|---|---|---|
| $Y_{1,1}$ | $Y_{1,2}$ | $Y_{1,3}$ | $Y_{1,4}$ | $\dots$ | $Y_{1,100}$ | 0 |
| $Y_{2,1}$ | $Y_{2,2}$ | $Y_{2,3}$ | $Y_{2,4}$ | $\dots$ | $Y_{2,100}$ | 0 |
| $Y_{3,1}$ | $Y_{3,2}$ | $Y_{3,3}$ | $Y_{3,4}$ | $\dots$ | $Y_{3,100}$ | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_{500,1}$ | $Y_{500,2}$ | $Y_{500,3}$ | $Y_{500,4}$ | $\dots$ | $Y_{500,100}$ | 0 |
| $Y_{1,1}$ | $Y_{1,2}$ | $Y_{1,3}$ | $Y_{1,4}$ | $\dots$ | $Y_{1,100}$ | 1 |
| $Y_{2,1}$ | $Y_{2,2}$ | $Y_{2,3}$ | $Y_{2,4}$ | $\dots$ | $Y_{2,100}$ | 1 |
| $Y_{3,1}$ | $Y_{3,2}$ | $Y_{3,3}$ | $Y_{3,4}$ | $\dots$ | $Y_{3,100}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Y_{500,1}$ | $Y_{500,2}$ | $Y_{500,3}$ | $Y_{500,4}$ | $\dots$ | $Y_{500,100}$ | 1 |

We outline the essential steps of the procedure.

- create a continuous path $X_i$ from each time-series $\{Y_i\}$ (row-wise)
- if needed, make use of the lead-lag transform to account for the variability in data
- compute the truncated signature $S(X_i)|_L$ of the path $X_i$ up to level $L$
- use the terms of signature $\{S_i^I\}$ as features

Following the aforementioned steps, each observation $\{Y_i\}$ is converted into a two-dimensional path. We also used a cumulative sum transformation (2.20):

$$X_i = \left\{ \left( \tilde{Y}_i^{lead}, \tilde{Y}_i^{lag} \right) \right\}, \tag{2.30}$$

and then the signature terms are computed:

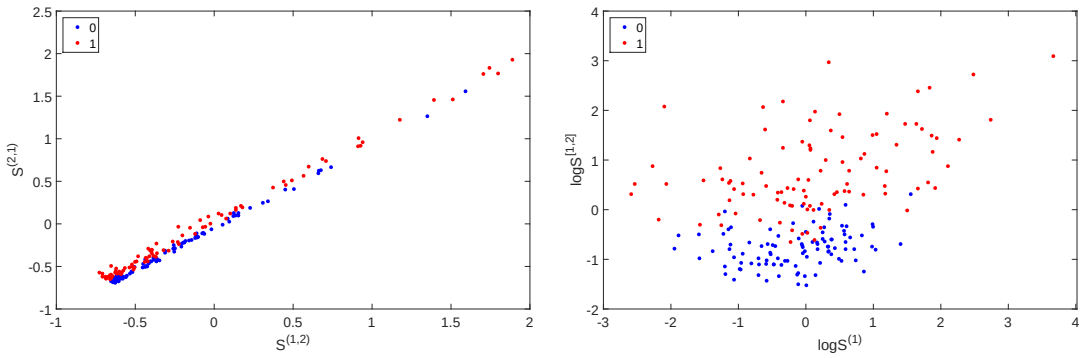
$$S(X_i) \;=\; \left(1, S_i^{(1)}, S_i^{(2)}, S_i^{(1,1)}, S_i^{(1,2)}, \ldots, S_i^{I}, \ldots\right). \tag{2.31}$$

The resulting feature matrix has the form:

| Feature Set | | | | | | Class |
|---|---|---|---|---|---|---|
| $\hat{S}_1^{(1)}$ | $\hat{S}_1^{(2)}$ | $\hat{S}_1^{(1,1)}$ | $\hat{S}_1^{(1,2)}$ | $\ldots$ | $\hat{S}_1^{I}$ | 0 |
| $\hat{S}_2^{(1)}$ | $\hat{S}_2^{(2)}$ | $\hat{S}_2^{(1,1)}$ | $\hat{S}_2^{(1,2)}$ | $\ldots$ | $\hat{S}_2^{I}$ | 0 |
| $\hat{S}_3^{(1)}$ | $\hat{S}_3^{(2)}$ | $\hat{S}_3^{(1,1)}$ | $\hat{S}_3^{(1,2)}$ | $\ldots$ | $\hat{S}_3^{I}$ | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\hat{S}_{500}^{(1)}$ | $\hat{S}_{500}^{(2)}$ | $\hat{S}_{500}^{(1,1)}$ | $\hat{S}_{500}^{(1,2)}$ | $\ldots$ | $\hat{S}_{500}^{I}$ | 0 |
| $\hat{S}_1^{(1)}$ | $\hat{S}_1^{(2)}$ | $\hat{S}_1^{(1,1)}$ | $\hat{S}_1^{(1,2)}$ | $\ldots$ | $\hat{S}_1^{I}$ | 1 |
| $\hat{S}_2^{(1)}$ | $\hat{S}_2^{(2)}$ | $\hat{S}_2^{(1,1)}$ | $\hat{S}_2^{(1,2)}$ | $\ldots$ | $\hat{S}_2^{I}$ | 1 |
| $\hat{S}_3^{(1)}$ | $\hat{S}_3^{(2)}$ | $\hat{S}_3^{(1,1)}$ | $\hat{S}_3^{(1,2)}$ | $\ldots$ | $\hat{S}_3^{I}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\hat{S}_{500}^{(1)}$ | $\hat{S}_{500}^{(2)}$ | $\hat{S}_{500}^{(1,1)}$ | $\hat{S}_{500}^{(1,2)}$ | $\ldots$ | $\hat{S}_{500}^{I}$ | 1 |

It is usual practice to standardise the signatures column-wise, where the *hat* corresponds to standardised values, computed by subtracting from each value $S_i^I$ the column mean and dividing by the column standard deviation.

For sake of simplicity, we compute the truncated signature up to level $L = 2$, remove the first constant term and thus produce a 6-dimensional feature space. The example of the signature and the log-signature terms projected on the two dimensional plane are presented in Fig. 16.




(a) (b)

**Figure 16**: Feature extraction from ARMA(1,1) time-series. The left panel (a) represents two signature terms $S^{(1,2)}$ and $S^{(2,1)}$, while the right panel (b) shows the first and the third terms of log-signature $\log S^1$,$\log S^{[1,2]}$.

Since we are not aiming to compare different classification algorithms, but rather to demonstrate the signature method at work, we arbitrarily choose a standard logistic regression algorithm with LASSO [25] penalization scheme. The relevant features which are selected by LASSO are:

$$\left\{S^{(1)}, S^{(2)}, S^{(1,2)}, S^{(2,1)}\right\}.$$

The classification results are summarised in Table 1.

| True \ Predicted | 0 | 1 |
|---|---|---|
| 0 | 326 | 20 |
| 1 | 35 | 319 |

(a) Training data set (70%) with accuracy 0.92

| True \ Predicted | 0 | 1 |
|---|---|---|
| 0 | 139 | 15 |
| 1 | 14 | 132 |

(b) Testing data set (30%) with accuracy 0.90

**Table 1**: Confusion matrix of the classification showing both the training and the testing data sets results.

Using the full signature for feature extraction is conceptually similar to working with basis functions. Having obtained the signature terms as representing features of data, one can continue further with standard methods in machine learning to solve problems.

### 2.2.2 Dealing with missing data in time-series

Sometimes sequential observations are not complete and some values might be missing due to various reasons. For example in medical follow up studies patients may not provide response data at required times due to personal reasons or clinical conditions. There are many ways to deal with missing data, for example imputing, various sophisticated interpolation methods, Gaussian smoothing kernels, and many more techniques are thoroughly studied and described in the literature. These methods might perform well and even result in robust analysis and predictions, but conceptually, imputing missing values may change the underlying information about initial data and thus create biased analysis which masks the original message.

One of the methods to treat missing values is to introduce a new binary variable: *indicator* matrix $R_{ij}$ along with observations $Y_{ij}$, where $i$ corresponds to individual time-series and $j$ denotes a time instance as explained in [6]. The elements of the indicator matrix are defined by:

$$R_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ observed,} \\ 1 & \text{if } Y_{ij} \text{ is missing.} \end{cases} \tag{2.32}$$

The data might be of three major types:

- MCAR - missing completely at random: missingness does not depend on data at all $P(M|Y,\theta) = P(M|\theta)$
- MAR - missing at random: missingness does not depend on the unobserved data, but does depend on the observed values $P(M|Y,\theta) = P(M|Y_{obs},\theta)$
- NMAR - not missing at random: missingness depends on the unobserved data $P(M|Y,\theta) = P(M|Y_{miss},\theta)$

A common approach to treat missing vales is to base inference on the likelihood function for the incomplete data by treating the indicator matrix $R_{ij}$ as a random variable. Specifying the joint distribution of $R_{ij}$ and $Y_{ij}$ together with the Expectation-Maximisation algorithm is usually the method of choice.

The signature framework naturally allows us to embed the indicator vector (a single row of the indicator matrix $R_{ij}$) into the path by lifting the path in to higher dimension. For example, consider a time-series with several missing values and its corresponding indicator vector (here the index $i$ is fixed):

$$\begin{aligned} Y_j &= \{1, 3, \star, 5, 3, \star, \star, 9, 3, 5\} \\ R_j &= \{0, 0, 1, 0, 0, 1, 1, 0, 0, 0\}, \end{aligned} \tag{2.33}$$

where the symbol "$\star$" denotes a missing value at the expected time point. The idea is to create a two dimensional path from this data. The evolution of the path is from the starting point towards the end and can be seen as propagation in three dimensional space; the first two dimensions are "observed data" and "missing data", while the third direction corresponds to time. At every time point where we have a missing point, we jump from the "observed" to the "missing" dimension and fill in the missing place with the same value as seen before (a.k.a *feed forward* method). The intuition is simple: unless we have any new information about the data, we continue walking along a "missing" axis direction, which is in another dimension than the "observed" data. The resulting path is given by (2.34)

$$\tilde{Y}_j = \{(0,1,0), (1,3,0), (2,\mathbf{3},1), (3,5,1), (4,3,1), (5,\mathbf{3},2), (6,\mathbf{3},3), (7,9,3), (8,3,3), (9,5,3)\}. \tag{2.34}$$

Here we introduce an auxiliary time parametrisation vector $t = \{0, 1, 2, \ldots, 9\}$ to account for the time propagation and create a path in three-dimensional space as depicted in Fig. 17. Red points represent the missing data. The path propagates from $t = 0$ to $t = 9$, and the observed points lie in the plane $\{t, Y\}$, while in the presence of a missing value, the path jumps along the increasing direction of $R$ axis.

This approach allows us to treat data streams with unobserved data at the same footing as complete data streams, which demonstrates another conceptual advantage of the signature framework.

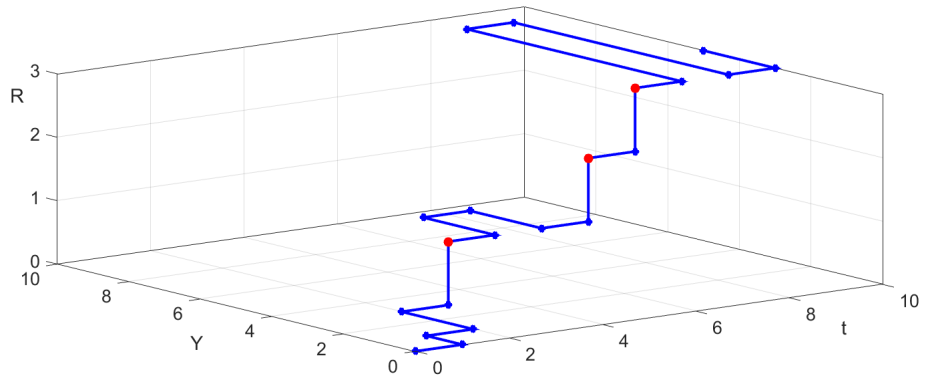



**Figure 17**: Example of embedding of data with missing values into a single path. The red dots represent unobserved data.

### 2.3 Computational considerations of the signature

So far, we have presented a theoretical background and practical applications of this method, but nothing has been mentioned about exact numerical computations of the signature and iterated integrals of paths. As it has been defined in (1.16), the terms of the signature are iterated integrals of a path, while the path is normally constructed by an interpolation of data points. One can compute such iterated integrals using several computational algorithms(cubature methods) which are generally straightforward to implement. The research group at the Oxford-Man Institute (University of Oxford, UK) works with a particular implementation of such algorithms written in C++ (*CoRoPa* project) [1] with Python wrapper package *sigtools*. Research groups from other universities have successfully implemented computations of iterated integrals with Matlab$^{TM}$.

### 2.4 Overview of recent progress of the signature method in machine learning

We would like to devote this final section to an overview of the applications of the signature method to various problems in machine learning and data analysis which have appeared in the literature. The application areas are quite wide, including financial data, sound compression, time-series analysis, medical data, and image recognition tasks.

#### 2.4.1 Extracting information from the signature of a financial data stream

Field *et al.* [7] have applied the signature transformation to financial data streams to find hidden patterns in trading strategies. The authors presented several examples of learning from data and further classification of unseen streams. The advantage of using the signature method is its ability to represent data using a small set of features which captures the most important properties of data in a non-parametric way without traditional statistical modelling. In their first experiment, the authors looked for atypical market behaviour

[1] http://coropa.sourceforge.net/

across standard 30-minute time buckets obtained from the WTI crude oil future market. In the second experiment, they aimed to distinguish between orders generated by two different trade execution algorithms. We will overview their methodology and results.

As we have already seen, the basics steps within the signature approach is to convert data into paths and then compute the iterated integrals of the resulting paths. In [7], the authors considered the following data streams:

- $P^a$ : best ask price
- $P^b$ : best bid price
- $V^a$ : number of orders at best ask price
- $V^b$ : number of orders at best bid price
- $C$ : cumulative traded volume

The final path was constructed from an embedding of normalisations of these streams. The normalisation was introduced in order to standardise the data and remove any spurious patterns. The ask and bid prices were transformed into the *mid price* $p_t$ and the *spread* $s_t$. Also, they included normalised time as $u_t$ and imbalance $d_t$. The details of the normalisation and transformation are described in [7]. The path is then:

$$X = \left\{(u_{t_i}, p_{t_i}, s_{t_i}, d_{t_i}, c_{t_i})_{i=0}^{N}\right\} \tag{2.35}$$

In order to capture the quadratic variation of the price, the path is extended by means of a lead-lag transform:

$$Z = \left\{\left(u_{t_i}^{lead}, p_{t_i}^{lead}, s_{t_i}^{lead}, d_{t_i}^{lead}, c_{t_i}^{lead}, p_{t_i}^{lag}\right)_{i=0}^{N}\right\} \tag{2.36}$$

Notice, if one is interested in the quadratic variation of any other variables, then the "lag" transformations of these variables should be included in the resulting path.

The aim is to learn a discriminant function which can classify new streams based on features extracted from signature of paths embedded from the data. The classification method was chosen as linear regression combined with LASSO shrinkage in order to select relevant terms in the signature. To measure the significance of separation, the authors used the non-parametric Kolmogorov-Smirnov test, and the receiver operating characteristic (ROC) curve and the area under the curve.

The objectives of the first experiment in [7] are WTI crude oil futures, sampled by minutes from standard 30-minutes interval between 09:30-10:00, 10:30-11.00, and 14:00-14:30. Considering the 6-dimensional input path (2.36) and its truncated signature at level $L = 4$, there are 1555 signature terms in total, but successful application of the LASSO algorithm allowed to select four terms which are the most informative. The results are illustrated in Fig 18. The upper panels show a visible separation between two groups plotted on the feature planes $S^{(1,5,1,5,)}, S^{(5,1,5,1)}$ (18a) and $S^{(5,1,5,1)}, S^{(1,5,5,1)}$ (18b) respectively. The

bottom panels demonstrate estimated densities of regressed values with KS distance 0.9 (training) and 0.91 (out-of-sample) and 95% accuracy 18c. The ROC curve of the classifier with AUC = 0.986 (out-of-sample AUC 0.984) is presented in 18d.

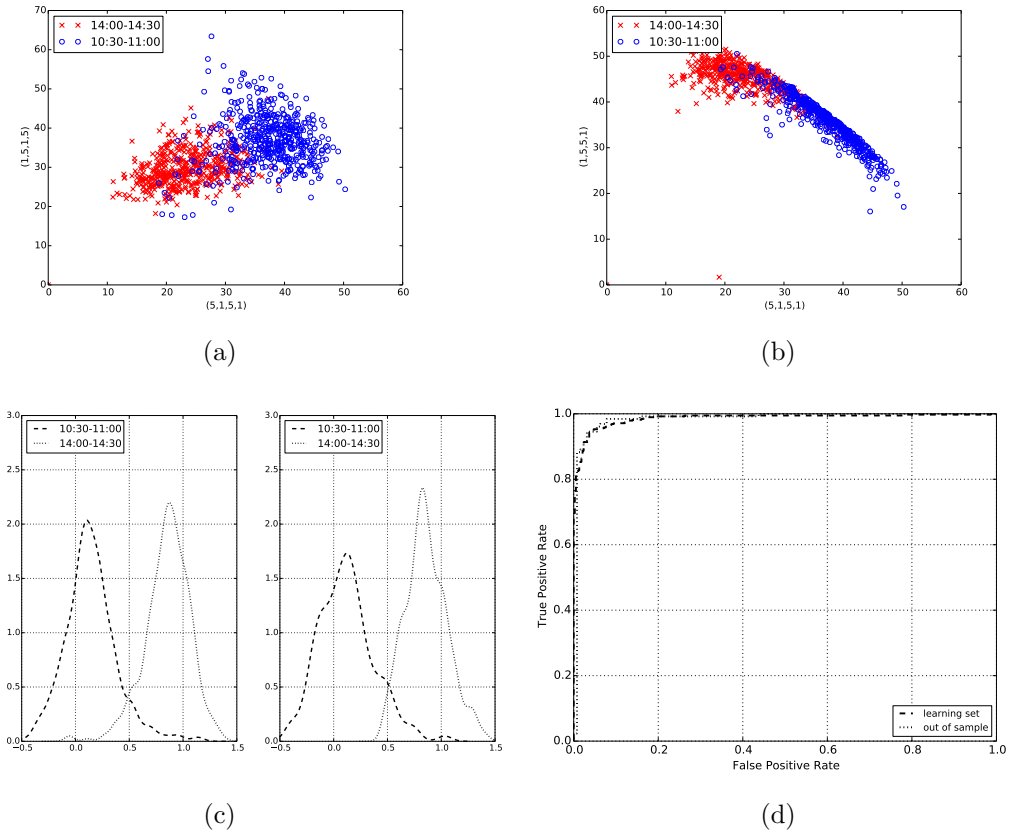
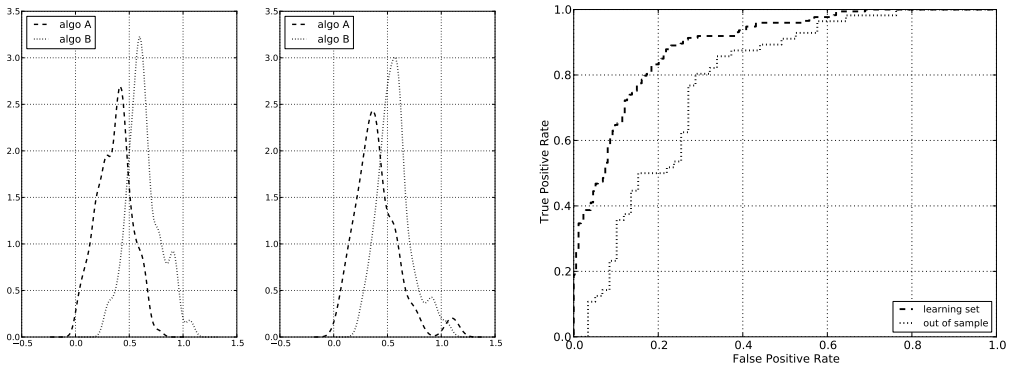



**Figure 18**: Visible separation between two groups projected onto $S^{(i,j,k,l)}$ feature plane (upper panels) and accuracy of separation (bottom panels). Courtesy of Field *et al.* [7].

The results from the second experiment of classification of trading algorithms are presented in Fig. 19. The aim is to classify the traces of trade execution algorithm. The data streams are sampled from the FTSE 100 index future market (NYSE Liffe) with the same transformation as described earlier. The beginning and the end of the streams are defined by the start and the end of parent orders generated by two different trade algorithms which are denoted by $A$ and $B$. Left panel 19a shows the estimated densities of regressed values with KS distance of 0.66 (0.518) with accuracy 82% (74.3%) for training (out-of-sample) data sets. The right panel shows the ROC curve of the classifier with the AUC 0.892 (0.777) for training (out-of-sample) sets respectively.




(a) (b)

**Figure 19**: Result of classification of two trading algorithms. Courtesy of Field *et al.* [7].

The conclusions and numerical results of this work demonstrate a great potential for applications of the signature method to various problems in financial data.

### 2.4.2 Sound compression - the rough paths approach

This research project dates back to 2005, when T. Lyons and N. Sidorova worked on the application of rough paths theory to the problem of sound compression [21]. They presented a new approach which turned out to be more effective than the traditional Fourier and wavelet transforms. The main difference between these methods is in the linearity of the Fourier approach, while the signature method accounts for non-linear dependencies. Consider a problem of digital compression of a continuous multi-dimensional signal $X$. Assuming that the signal is measured on a fine time scale, the signal can be seen as a continuous piece-wise linear path. Computing the signature of the resulting path, will produce signature terms $S^I(X)$ which represent the coefficients of compression of the original signal. The authors of [21] presented an efficient algorithm for reconstruction of the original signal $X$ from its signature terms.

### 2.4.3 Character recognition

Paths as geometric objects are legitimate candidates for applications to image recognition tasks. For example, hand written digits can be seen as continuous paths, and then the signature method becomes a natural approach to classification problems. The signature method has been successfully applied to the challenging problem of Chinese character recognition by B. Graham [12], and later by J. Lianwen [26] to a more general problem of hand written character recognition. Graham and Lianwen have shown (separately) that using signature terms as features of the image, for use in a convolutional neural network (CNN), significantly improved accuracy of online character recognition. This approach,

which combines the signature and CNN methods, has won the ICDAR2013 Online Isolated Chinese Character recognition competition receiving 0.95530 accuracy[1]. The group of J. Lianwen has developed the application[2] with graphical user interface for character recognition. This application is based on the signature method for feature extraction from elements of images.

These successful results indicate that applications of the signature method combined with deep neural networks represent a new and efficient way to tackle challenges in the field of image recognition.

### 2.4.4 Learning from the past, predicting the statistics for the future, learning an evolving system

In [16], Levin, Lyons, and Ni address an application of the signature method to time-series analysis and develop a new regression method based on the expected signature of paths. This method of linear regression on the signature of data streams is general and allows explicit computations and predictions. The main goal of this approach is to identify the specific feature set of the observed data and linearise the functional relationships between them (features). The authors introduced *the expected signature model* as a linear mapping: $L : X \mapsto Y$

$$Y = L(X) + \epsilon, \tag{2.37}$$

with $E[\epsilon] = 0$ and $X \in T((\mathbb{R}^d))$, $Y \in T((\mathbb{R}^e))$, where

$$T((\mathbb{R}^d)) = \bigoplus_{n=0}^{\infty} (\mathbb{R}^d)^{\otimes n} \tag{2.38}$$

is the algebra of tensor series over $\mathbb{R}^d$. We briefly outline the main idea of this model. Consider the univariate time-series (*returns*) denoted by $\{r_i\}_{i=1}^N$ and some fixed $k \in \mathbb{N}$, such that $1 < k < N$. Define the *p-past* returns up to time $t_k$ by $F_k = \{r_i\}_{i=k-p}^k$ and the *q-future* returns (from time $t_{k+1}$) by $G_{k+1} = \{r_i\}_{i=k+1}^{k+q}$. The signatures of $F_k$ and $G_{k+1}$ are respectively:

$$\begin{aligned} X_k &= S\left(\{t_i, r_i\}_{i=k-p}^{k}\right) \\ Y_k &= S\left(\{t_i, r_i\}_{i=k+1}^{k+q}\right). \end{aligned} \tag{2.39}$$

The definition of the Expected Signature model $ES(p, q, n, m)$ states that the stationary time-series $\{r_i\}_{i=1}^N$ satisfies the assumptions of the $ES$ model with parameters $p, q, n, m$ if there exists a linear functional $f : T^n\left(\mathbb{R}^2\right) \mapsto T^n\left(\mathbb{R}^2\right)$ such that:

$$\rho_m\left(S(\{r_{t+i}\}_{i=1}^q)\right) = f\left(\rho_m\left(S(\{r_{t-i}\}_{i=0}^p)\right)\right) + a_t, \tag{2.40}$$

[1] http://tinyurl.com/nfvbdlk
[2] http://www.deephcr.net/en.html

where $N \geq p+q$ is a positive integer, $m$ is the depth of truncation of signature, and the residual term $a_t$ satisfies $E[a_t|F_t] = 0$. Let $\mu_k$ be the expectation of $S(\{t_i, r_i\}_{i=k+1}^{k+q})$

$$\mu_k = E[S(\{t_i, r_i\}_{i=k+1}^{k+q})|F_k]. \tag{2.41}$$

It thus follows from the ES model that

$$\mu_k = f(X_k), \tag{2.42}$$

where $X_k$ is defined in (2.39) with both $\mu_k$ and $a_k$ in $T((\mathbb{R}^2))$. Next, the conditional covariance of the signature of the future return conditioned on $F_k$ is defined as

$$\Sigma_k^2(I,J) = Cov\left(S^I(Y_k), S^J(Y_k)|F_k\right), \tag{2.43}$$

where $I, J$ are two multi-indexes as defined previously in the text. In time-series analysis, (2.41) corresponds to the mean equation for $Y_k$, and (2.43) is the volatility equation for $Y_k$. The fundamental assumption of the ES model is the stationarity of the time-series $\{r_i\}$, which is common in time-series analysis.

It has been shown that many standard autoregressive models (AR, ARCH, GARCH, etc.) are special cases of the ES model. For example, given the time-series $\{t_i, r_i\}$ the mean equation and volatility are written as:

| | AR | ES |
|---|---|---|
| $m_k$ : | $E[r_{k+1}\|F_k]$ | $S^{(2)}(\mu_k)$ |
| $\sigma_k^2$ : | $Var[r_{k+1}\|F_k]$ | $2S^{(2,2)}(\mu_k) - \left(S^{(2)}(\mu_k)\right)^2$. |

Several important properties linking the autoregressive models and ES model are proven in [16]. One of them states that if a time-series $\{r_k\}$ satisfies the assumptions of ARCH(q) model and its mean equation is given by

$$\mu_k = \beta_0 + \sum_{i=1}^{Q} \beta_i r_{k-i}, \tag{2.44}$$

then there exists a sufficiently large integer $n$ such that the time-series $\{r_k\}$ satisfies the assumption of $ES(q+Q, 1, n, 2)$. Also, the model $ES(p, 1, n, 2)$ is considered as a classical time-series, where the signature of the past returns as an explanatory variable.

The work in [16] provides explicit numerical benchmarks between various autoregressive models and the expected signature model for various set-ups and configurations of the models.

### 2.4.5 Identifying patterns in MEG scans

The project of Gyurko *et al.* [13] is devoted to an application of the signature method to pattern recognition in medical studies and neuroimaging (**M**agneto**E**ncephalo**G**raphy). The patients are asked to press a button whenever they like, and their electromagnetic

brain activity is constantly monitored during this trial. The data streams of the electric signal are shown in Fig. 20 in the first two upper panels. For the sake of simplicity, Gyurko *et al.* considered only two channels from the MEG scanner as sources for the streaming data. The main goal was to identify a pattern in these two data streams which is able to predict the pressing of a button. The button-pressing instance was presented by a vector with binary outcome: "1" and "0" stand for the button being pressed and not being pressed respectively (Fig. 20, bottom panel). They used a rolling window analysis of time-series and regressed the button-pressing instances on signature of the data streams inside the rolling window.

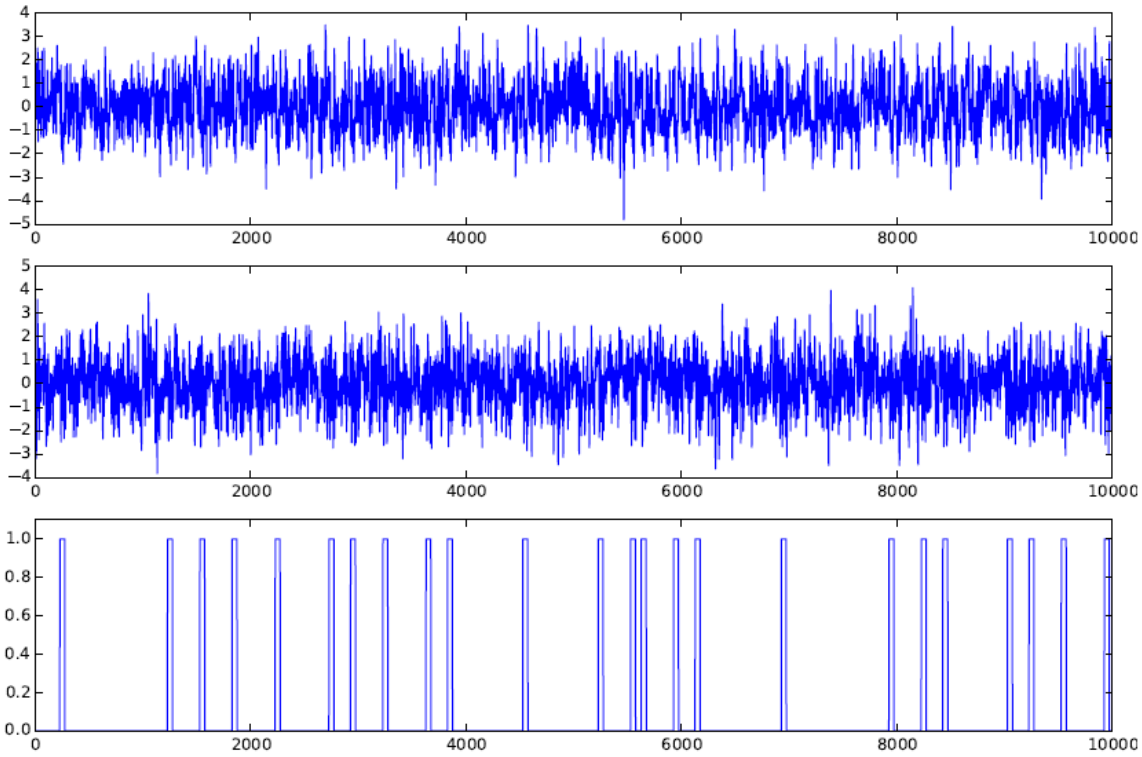



**Figure 20**: Finding patterns in two-dimensional data stream (two upper panels). The bottom panel shows the binary vector of the button-state (pressed or not pressed). The horizontal axis is in time units.

The detection and prediction results were good: AUC of ROC: 0.92 (learning set) and AUC of ROC (out-of-sample) 0.93, which indicates good separability between the two states (1 and 0). This example shows that applications of the signature approach are fairly general and are not constrained to particular cases.

### 2.4.6 Learning the effect of treatment on behavioural patterns of patients with bipolar disorder.

This is another example of an application of the signature method to analysis of medical data. Our main goal of this project is to learn the behavioural patterns of patients suffering from bipolar disorder. The data are follow-up self-reported assessment scale collected by the True Colours[1] web platform which is developed by the Department of Psychiatry at the University of Oxford and is used for the CONBRIO programme[2]. Patients are encouraged to submit via this platform their answers to the self-assessment questionnaires. These questionnaires are standard clinical tools for assessment of the severity of depression (QIDS-SR$_{16}$) [23] and mania (AMSR) [1] symptoms. These questionnaires contain 16 and 5 questions and scales are ranged [0 27] and [0 25] for QIDS-SR$_{16}$ and AMSR respectively. In Fig. 21, an example of the QIDS score for 50 weeks follow-up observations is displayed.

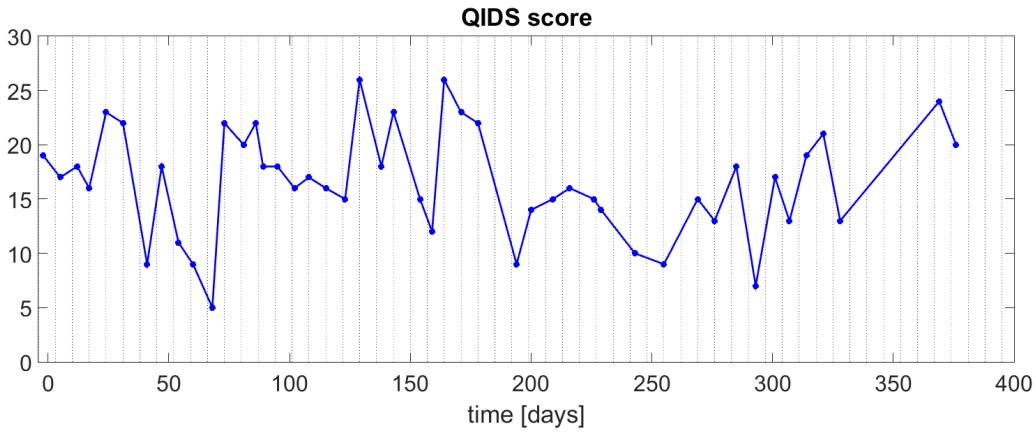



**Figure 21**: A typical example of the follow-up observations of the QIDS score.

As a part of our research programme, we analysed data from the CEQUEL clinical trial [10]. The primary objective is to compare the combination therapy of quetiapine plus lamotrigine with quetiapine monotherapy for treatment of bipolar depression. Among other interesting research questions, we focused on the behavioural pattern of patients in two different treatment groups: placebo and lamotrigine. Patients are asked to submit their self-assessment scores to the True Colours platform on a weekly basis. The system sends a scheduled prompt text every week and patients should reply within one week interval. If no response is received within a week, the system flags this week as a missing observation and sends a new prompt at the beginning of the next week. We analysed the delays - the time intervals between the prompt text and the patient's reply measured in integer units of days. The delays are presented schematically in the Fig. 22.

[1] https://oxfordhealth.truecolours.nhs.uk/www/en/
[2] http://conbrio.psych.ox.ac.uk/home

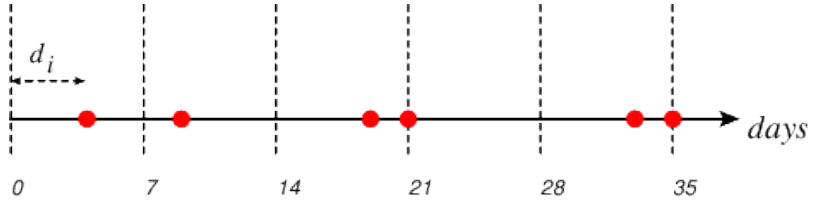



**Figure 22**: Definition of the delays $d_i$. The time line is divided by the dashed lines into weekly intervals. The delay is defined as the time difference between the beginning of the week (dashed line) and the time when the response is received (red dot). This particular configuration of delays corresponds to: $\{d_i\} = \{4, 2, 5, 0, 5, 0\}$.

The delays of the QIDS data from the Fig. 21 are shown in the Fig. 23.

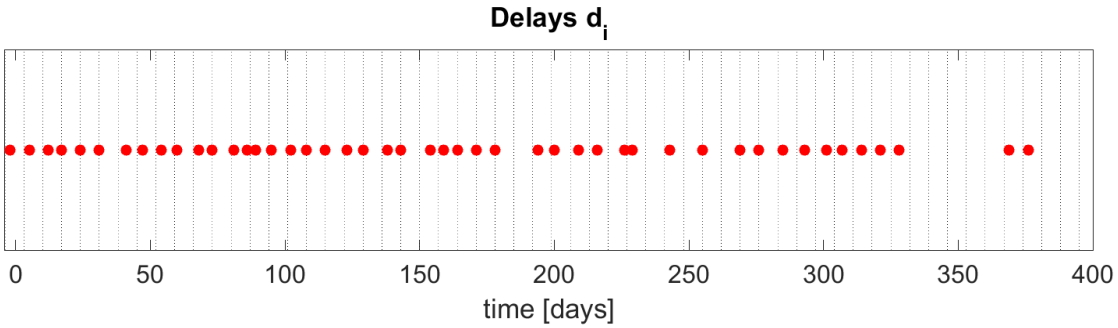



**Figure 23**: Example of the delays $d_i$ of the QIDS scores presented in the Fig. 21. Submissions are denoted by the red dots.

We were interested in the question, whether the distribution of delays $\{d_i\}$ is different between the two different treatment groups, namely placebo and lamotrigine. To answer this question, we applied the signature method to time-series constructed from the delay data $\{d_i\}$ and used regularised logistic regression to discriminate between the two groups. The sample sizes of the two groups are: 18 and 11 patients in the placebo and the lamotrigine group respectively. We constructed continuous paths from the sequence of integer delays $\{d_i\}$ by means of two types of embeddings: the first is the three dimensional simple lead-lag transformation of the raw data $\{d_i\}$ with time stamps:

$$X = \left\{ \left( t_i^{lead}, d_i^{lead}, d_i^{lag} \right)_i \right\}, \tag{2.45}$$

and the second embedding corresponds to the lead-lag transformation of the cumulative sum (partial sums) of the delays:

$$S = \left\{ \left( s_i^{lead}, s_i^{lag} \right)_i \right\}. \tag{2.46}$$

We computed signature terms from these two embeddings separately up to level $L$ and combined these terms in a single design matrix, which we used as the feature matrix for

the logistic regression. The design matrix exhibited some degree of multicollinearity, thus we applied the elastic net regularisation scheme [27] to the logistic regression. The elastic net allows a smooth regularisation, which benefits from both $L_1$ and $L_2$ norm shrinkage schemes, groups the correlated covariates, and removes the less relevant features. The choice of the $\alpha$ parameter of the elastic net is performed by the cross validated grid search, and estimated as $\alpha = 0.55$ giving the smallest deviance error as shown in Fig. 24.

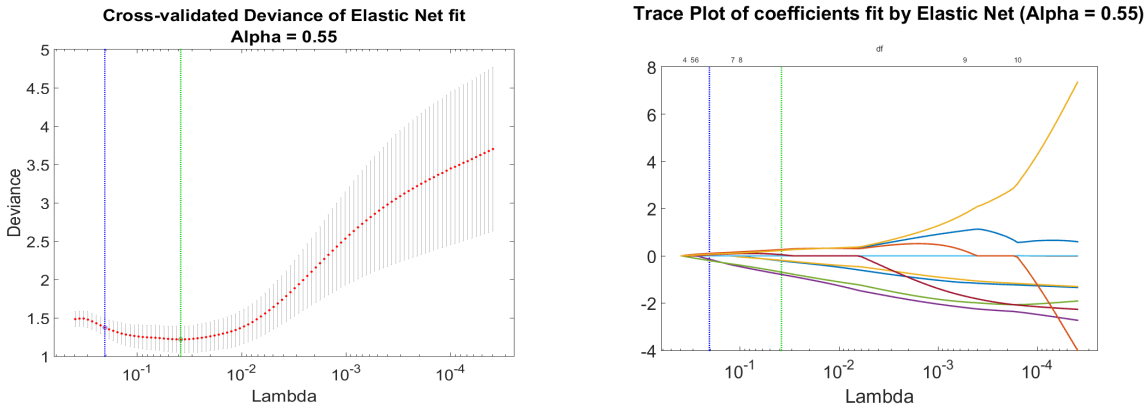



**Figure 24**: Statistical results of cross validated elastic net regularisation scheme.

The final classification model is chosen by the "one-standard-error" rule, denoted by the position of the blue dashed line, given the fact that the initial 10-dimensional feature set was shrunk to a 6-dimensional subset. Some of the features of this subset are displayed in the Fig. 25.

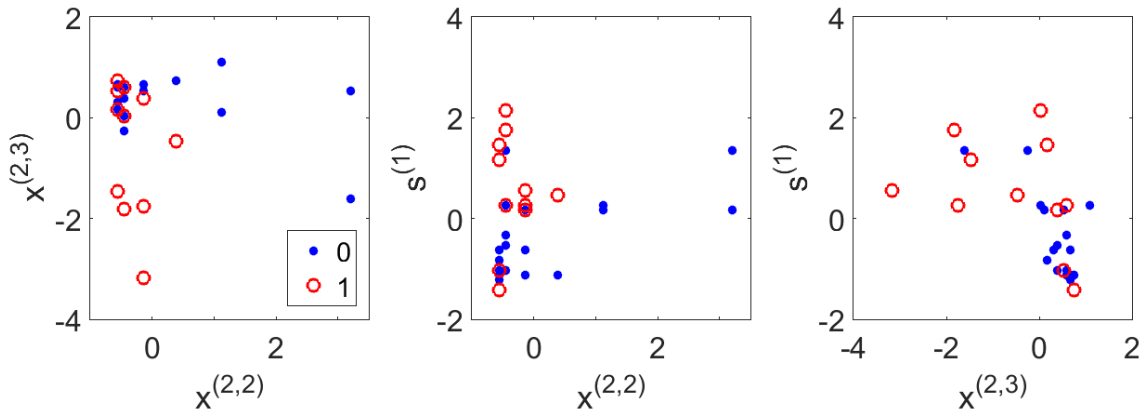



**Figure 25**: Examples of relevant features selected by the penalised logistic regression. These two groups denoted by "0" and "1" correspond to the placebo and the lamotrigine groups respectively.

The results of this project demonstrated the ability of the signature method to extract characteristic features from the data in a systematic way without introducing any *ad hoc* methods. The signature feature extraction method serves as an initial step in the machine

learning pipeline. Once we transformed the raw data into a set of signature terms, we naturally proceed further with standard machine learning techniques and methodologies using these features as inputs.

## Acknowledgement

A.K. wishes to thank the Oxford-Man Institute of Quantitative Finance for the hospitality during the course of this work and gratefully acknowledges the support of the Welcome Trust grant No: 102616/Z/13/Z, "CONBRIO".

## References

[1] Edward G Altman, Donald Hedeker, James L Peterson, and John M Davis. The altman self-rating mania scale. *Biological psychiatry*, 42(10):948–955, 1997.

[2] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: Uniqueness. arXiv:1406.7871, August 2014. Preprint.

[3] Horatio Boedihardjo, Hao Ni, and Zhongmin Qian. Uniqueness of signature for simple curves. *J. Funct. Anal.*, 267(6):1778–1806, 2014.

[4] Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Ann. of Math. (2)*, 65:163–178, 1957.

[5] Kuo-Tsai Chen. Integration of paths—a faithful representation of paths by non-commutative formal power series. *Trans. Amer. Math. Soc.*, 89:395–407, 1958.

[6] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

[7] Jonathann Field, Lajos Gergely Gyurkó, Mark Kontkowski, and Terry Lyons. Extracting information from the signature of a financial data stream. arXiv:1307.7244, July 2014. Preprint.

[8] Guy Flint, Ben Hambly, and Terry Lyons. Discretely sampled signals and the rough hoff process. *arXiv preprint arXiv:1310.4054*, 2015.

[9] Peter K. Friz and Martin Hairer. *A course on rough paths*. Universitext. Springer, Cham, 2014. With an introduction to regularity structures.

[10] John R Geddes, Alexandra Gardiner, Jennifer Rendell, Merryn Voysey, Elizabeth Tunbridge, Christopher Hinds, Ly-Mee Yu, Jane Hainsworth, Mary-Jane Attenburrow, Judit Simon, et al. Comparative evaluation of quetiapine plus lamotrigine combination versus quetiapine monotherapy (and folic acid versus placebo) in bipolar depression (cequel): a 2× 2 factorial randomised trial. *The Lancet Psychiatry*, 2015.

[11] Xi Geng. Reconstruction for the signature of a rough path. arXiv:1508.06890, August 2015. Preprint.

[12] Benjamin Graham. Sparse arrays of signatures for online character recognition. arXiv:1308.0371, December 2013. Preprint.

[13] Lajos Gergely Gyurko, Terry Lyons, and Harald Oberhauser. Identifying patterns via the signature for the comnbrio project. University of Oxford, 2014.

[14] 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(1):109–167, 2010.

[15] Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *arXiv preprint arXiv:1601.08169*, 2016.

[16] Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. arXiv:1309.0260, September 2015. Preprint.

[17] Terry Lyons and Weijun Xu. Hyperbolic development and inversion of signature. arXiv:1507.00286, July 2015. Preprint.

[18] Terry Lyons and Weijun Xu. Inverting the signature of a path. arXiv:1406.7833, July 2015. Preprint.

[19] Terry J. Lyons. Differential equations driven by rough signals. *Rev. Mat. Iberoamericana*, 14(2):215–310, 1998.

[20] Terry J. Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*, volume 1908 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[21] Terry J Lyons and Nadia Sidorova. Sound compression: a rough path approach. In *Proceedings of the 4th international symposium on Information and communication technologies*, pages 223–228. Trinity College Dublin, 2005.

[22] Rimhak Ree. Lie elements and an algebra associated with shuffles. *Ann. of Math. (2)*, 68:210–220, 1958.

[23] A John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, et al. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54(5):573–583, 2003.

[24] Christos Sapsanis, George Georgoulas, and Anthony Tzes. Emg based classification of basic hand movements based on time-frequency features. In *Control & Automation (MED), 2013 21st Mediterranean Conference on*, pages 716–722. IEEE, 2013.

[25] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[26] Weixin Yang, Lianwen Jin, and Manfei Liu. Deepwriterid: An end-to-end online text-independent writer identification system. *arXiv preprint arXiv:1508.04945*, 2015.

[27] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.