# A Primer on the Signature Method in Machine Learning

Lyu Chenxin

University of Hong Kong

December 28, 2022

# Overview

# About the article

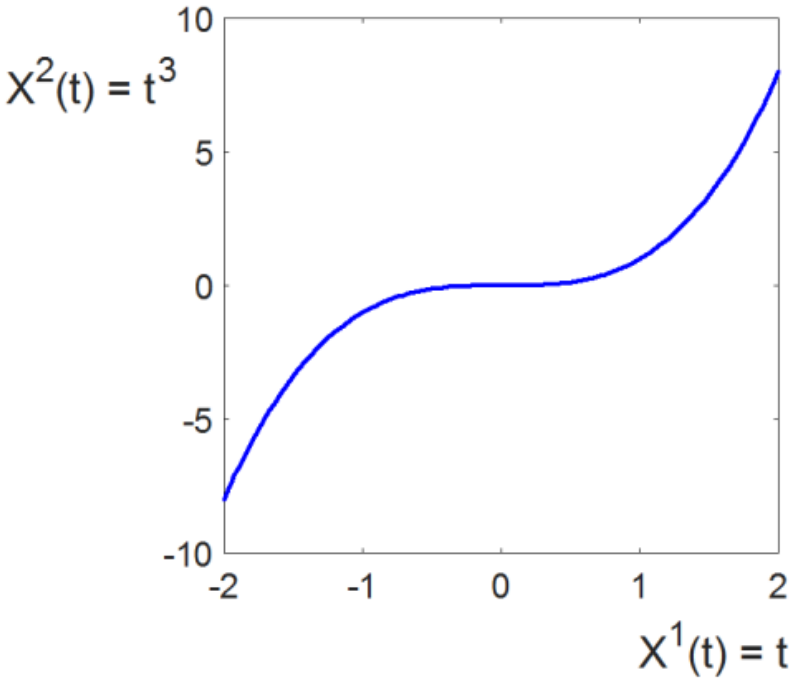
- Name of the article:
  A Primer on the Signature Method in Machine Learning
- Author(s): Ilya Chevyrev and Andrey Kormilitzin
- URL: https://arxiv.org/abs/1603.03788

# Paths in Euclidean space

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:

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

# Paths in Euclidean space



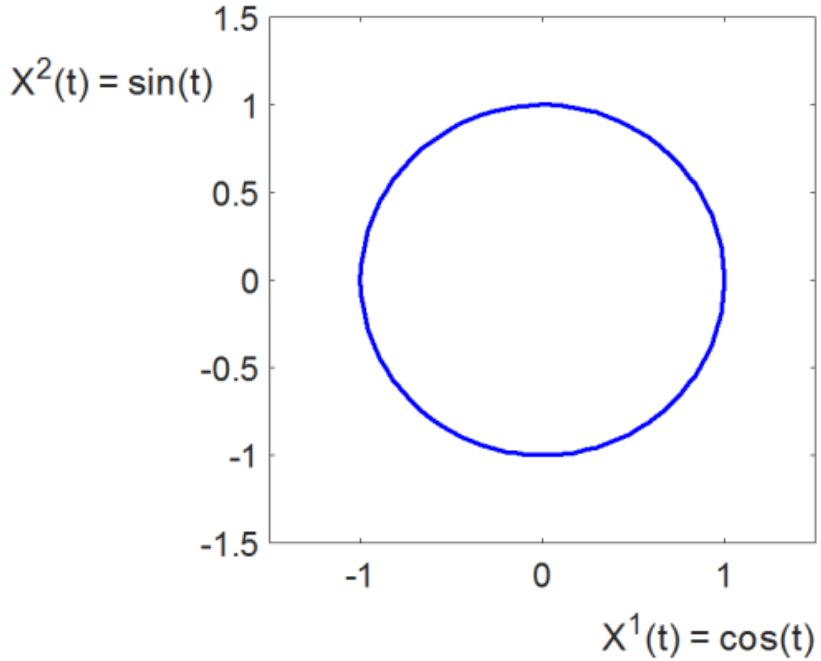




Figure 1

# Paths in Euclidean space

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

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

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

$$X_t = \left\{X_t^1, X_t^2\right\} = \{t, f(t)\}, t \in [0, 1], \tag{3}$$

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 $t$.

# Paths in Euclidean space

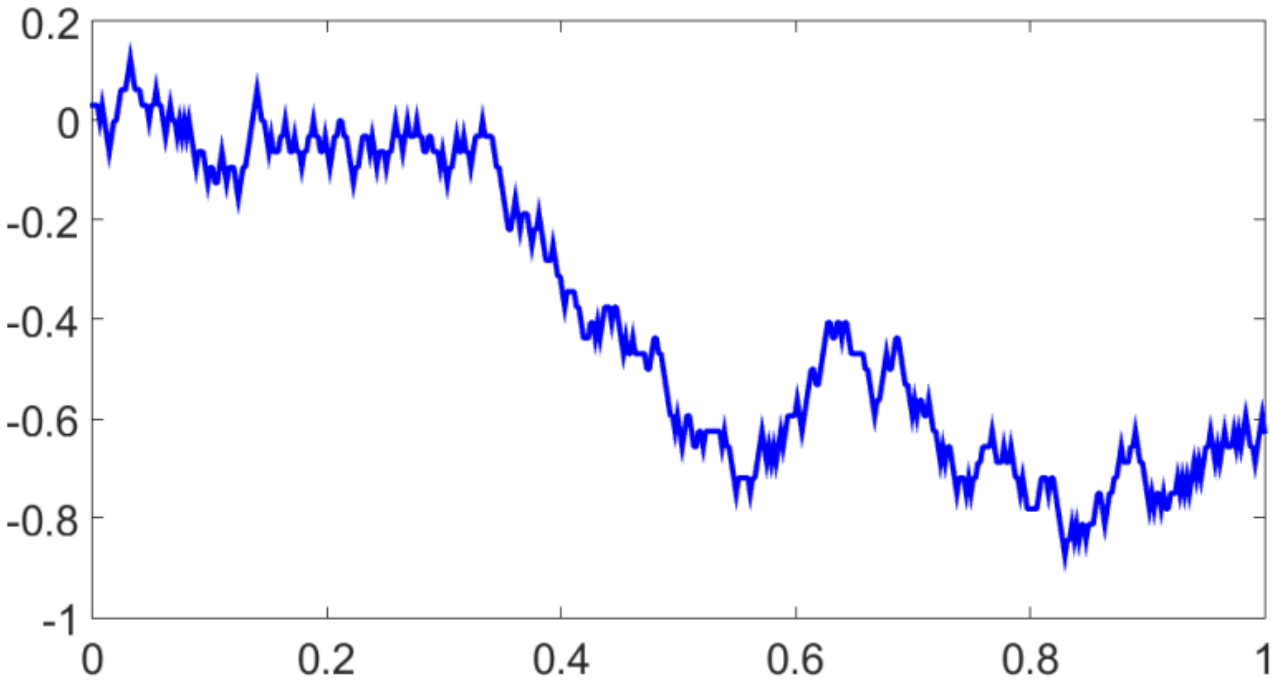



Figure 2

# 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

$$\int_a^b f(X_t)\, dX_t = \int_a^b f(X_t)\, \dot{X}_t dt \tag{4}$$

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: $\dot{X}_t = dX_t/dt$.

# Path integrals

In the expression (4), note that $f(X_t)$ is itself a real-valued path defined on $[a, b]$. 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

$$\int_a^b Y_t dX_t = \int_a^b Y_t \dot{X}_t dt. \tag{5}$$

We recover the usual path integral upon setting $Y_t = f(X_t)$.

# 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 $\left(X_t^1, \dots X_t^d\right)$, 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

$$S(X)_{a,t}^i = \int_{a<s<t} dX_s^i = X_t^i - X_0^i, \tag{6}$$

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 $[a, b]$.

# Definition

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

$$S(X)^{i,j}_{a,t} = \int_{a<s<t} S(X)^{i}_{a,s} dX^{j}_{s} = \int_{a<r<s<t} dX^{i}_{r} dX^{j}_{s}, \tag{7}$$

where $S(X)^{i}_{a,s}$ is given by (6) and the integration limits are simply:

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

We emphasise again that $S(X)^{i}_{a,s}$ and $X^{j}_{s}$ are simply real-valued paths, so the expression (7) is a special case of the path integral, and that $S(X)^{i,j}_{a,\cdot} : [a, b] \mapsto \mathbb{R}$ is itself a real-valued path.

# Definition

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

$$S(X)_{a,t}^{i,j,k} = \int_{a<s<t} S(X)_{a,s}^{i,j} dX_s^k = \int_{a<q<r<s<t} dX_q^i dX_r^j dX_s^k \tag{8}$$

Again, since $S(X)_{a,s}^{i,j}$ and $X_s^k$ are real-valued paths, the above is just a special case of the path integral, and $S(X)_{a,\cdot}^{i,j,k} : [a, b] \mapsto \mathbb{R}$ is itself a real-valued path.
We can continue recursively, and for any integer $k \geq 1$ and collection of indexes $i_1, \dots, i_k \in \{1, \dots, d\}$, we define

$$S(X)_{a,t}^{i_1,\dots,i_k} = \int_{a<s<t} S(X)_{a,s}^{i_1,\dots,i_{k-1}} dX_s^{i_k} \tag{9}$$

# Definition

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

$$S(X)_{a,t}^{i_1,\ldots,i_k} = \int_{a<t_k<t} \ldots \int_{a<t_1<t_2} dX_{t_1}^{i_1} \ldots dX_{t_k}^{i_k} \tag{10}$$

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

# Definition

## Definition: 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

$$S(X)_{a,b} = \left(1, S(X)^1_{a,b}, \dots, S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \dots\right) \tag{11}$$

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

$$W = \{(i_1, \dots, i_k) \mid k \geq 1, i_1, \dots, i_k \in \{1, \dots, d\}\}. \tag{12}$$

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

# Example

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

$$\begin{aligned} S(X)^1_{a,b} &= X_b - X_a \\ S(X)^{1,1}_{a,b} &= \frac{(X_b - X_a)^2}{2!} \\ S(X)^{1,1,1}_{a,b} &= \frac{(X_b - X_a)^3}{3!} \\ &\vdots \end{aligned} \tag{13}$$

Hence, for one-dimensional paths, the signature depends only on the increment $X_b - X_a$.

# Example

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

$$W = \{(i_1, \ldots, i_k) \mid k \geq 1, i_1, \ldots, i_k \in \{1, 2\}\}, \tag{14}$$

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

# Example

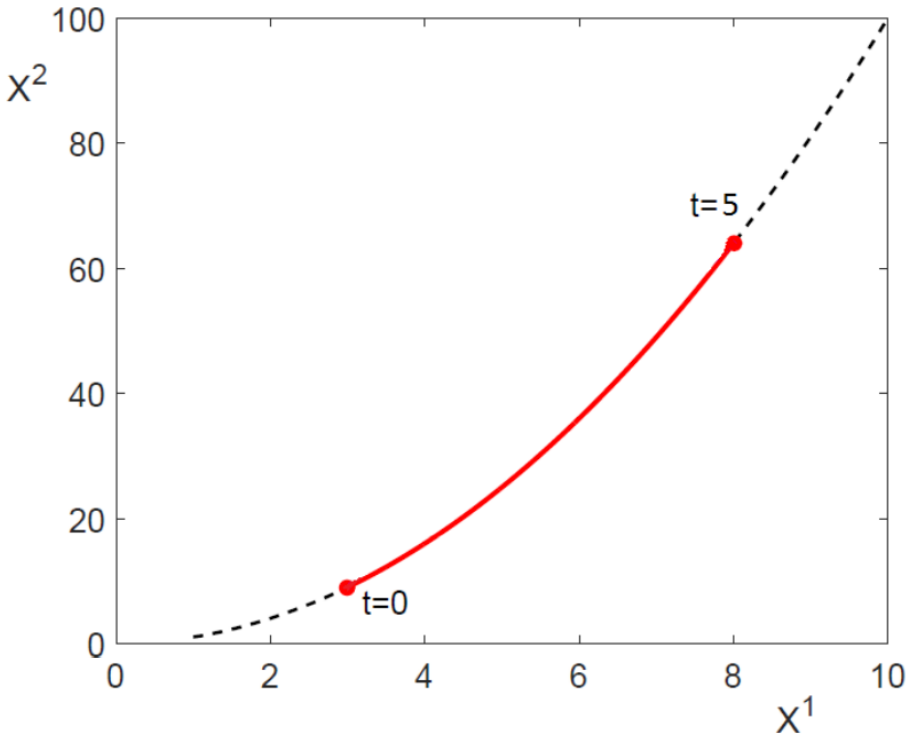



Figure 3

# Example

Explicitly:

$$\begin{aligned} X_t &= \left\{X_t^1, X_t^2\right\} = \left\{3+t, (3+t)^2\right\} \quad t \in [0,5], (a=0, b=5), \\ dX_t &= \left\{dX_t^1, dX_t^2\right\} = \{dt, 2(3+t)dt\} \end{aligned} \tag{15}$$

A straightforward computation gives:

$$\begin{aligned} S(X)_{0,5}^1 &= \int_{0<t<5} dX_t^1 = \int_0^5 dt = X_5^1 - X_0^1 = 5 \\ S(X)_{0,5}^2 &= \int_{0<t<5} dX_t^2 = \int_0^5 2(3+t)dt = X_5^2 - X_0^2 = 55 \\ S(X)_{0,5}^{1,1} &= \iint_{0<t_1<t_2<5} dX_{t_1}^1 dX_{t_2}^1 = \int_0^5 \left[\int_0^{t_2} dt_1\right] dt_2 = \frac{25}{2}, \\ S(X)_{0,5}^{1,2} &= \iint_{0<t_1<t_2<5} dX_{t_1}^1 dX_{t_2}^2 = \int_0^5 \left[\int_0^{t_2} dt_1\right] 2\left(3+t_2\right) dt_2 = \frac{475}{3}, \end{aligned} \tag{16}$$

# Example

$$
\begin{aligned}
S(X)^{2,1}_{0,5} &= \iint_{0<t_1<t_2<5} dX^2_{t_1} dX^1_{t_2} = \int_0^5 \left[ \int_0^{t_2} 2\,(3+t_1)\, dt_1 \right] dt_2 = \frac{350}{3}, \\
S(X)^{2,2}_{0,5} &= \iint_{0<t_1<t_2<5} dX^2_{t_1} dX^2_{t_2} = \int_0^5 \left[ \int_0^{t_2} 2\,(3+t_1)\, dt_1 \right] 2\,(3+t_2)\, dt_2 = \frac{3025}{2}, \\
S(X)^{1,1,1}_{0,5} &= \iiint_{0<t_1<t_2<t_3<5} dX^1_{t_1} dX^1_{t_2} dX^1_{t_3} = \int_0^1 \left[ \int_0^{t_3} \left[ \int_0^{t_2} dt_1 \right] dt_2 \right] dt_3 = \frac{125}{6}, \\
&\vdots
\end{aligned}
\tag{17}
$$

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

# Picard iteration

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

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

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

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

# Picard iteration

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

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

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

# Picard iteration

## Example

Consider the ODE:

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

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

$$y_0(x) = 1$$

$$y_1(x) = 1 + \int_0^x y_0(t)dt = 1 + x$$

$$y_2(x) = 1 + \int_0^x y_1(t)dt = 1 + x + \frac{1}{2}x^2$$

$$y_3(x) = 1 + \int_0^x y_2(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3$$

$$y_4(x) = 1 + \int_0^x y_3(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4$$

# Picard iteration

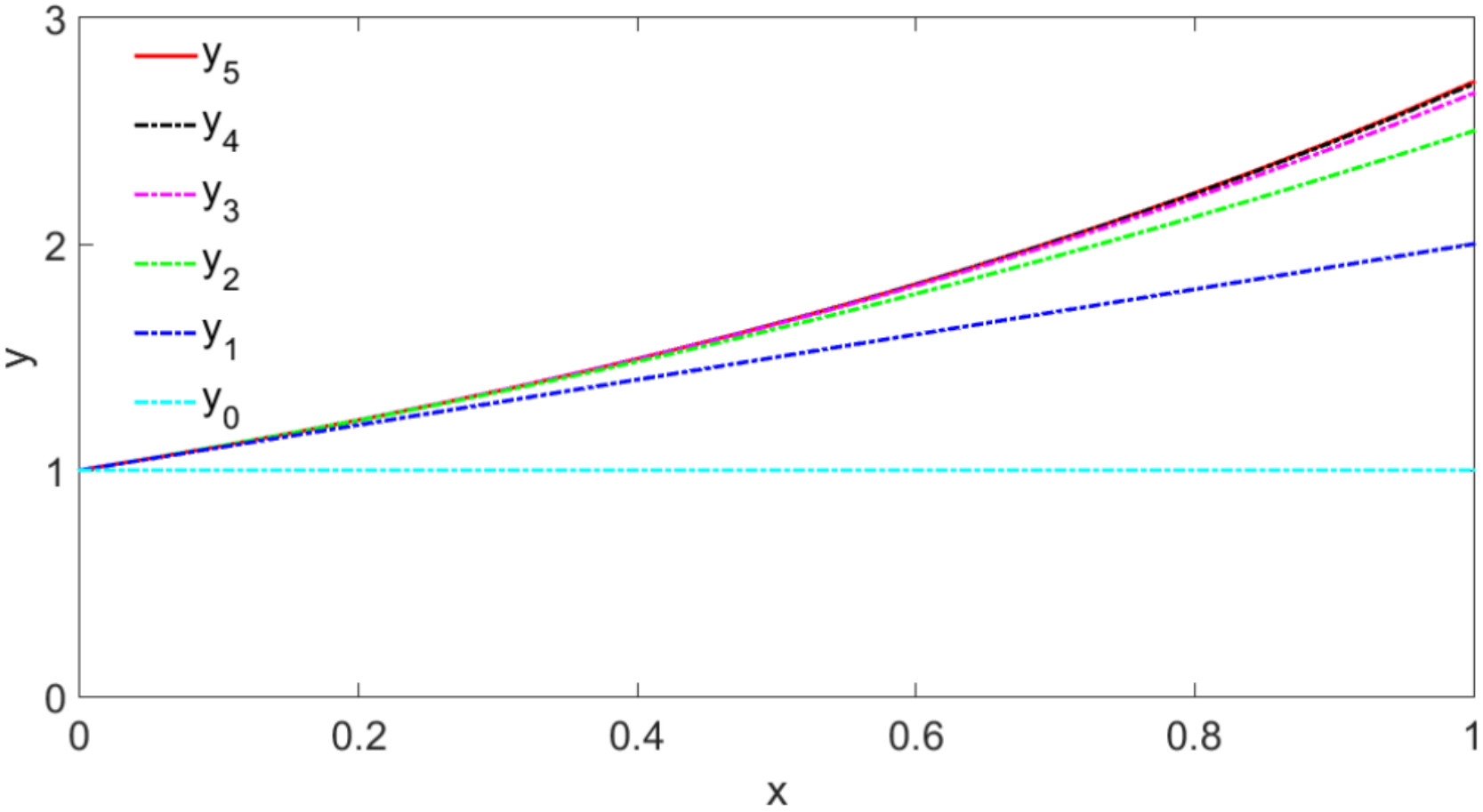
## Example

$$\vdots$$

$$y_k(x) = \sum_{n=0}^{k} \frac{1}{n!} x^n$$

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. 8.

# Picard iteration



# Picard iteration

Consider a path $X : [a, b] \mapsto \mathbb{R}^d$. Let $\mathbf{L}\left(\mathbb{R}^d, \mathbb{R}^e\right)$ denote the vector space of linear maps from $\mathbb{R}^d$ to $\mathbb{R}^e$. Equivalently, $\mathbf{L}\left(\mathbb{R}^d, \mathbb{R}^e\right)$ can be regarded as the vector space of $d \times e$ real matrices. For a path $Z : [a, b] \mapsto \mathbf{L}\left(\mathbb{R}^d, \mathbb{R}^e\right)$, note that we can define the integral

$$\int_a^b Z_t dX_t$$

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

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

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

$$Y_t = y + \int_a^t V\left(Y_s\right) dX_s. \tag{18}$$

# Picard iteration

A standard procedure to obtain a solution to (18) 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

$$F(Y)_t = y + \int_a^t V(Y_s)\, dX_s$$

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

# Picard iteration

Consider now the case when $V : \mathbb{R}^e \mapsto \mathbf{L}\left(\mathbb{R}^d, \mathbb{R}^e\right)$ is a linear map. Note that we may equivalently treat $V$ as a linear map $\mathbb{R}^d \mapsto \mathbf{L}\left(\mathbb{R}^e, \mathbb{R}^e\right)$, where $\mathbf{L}\left(\mathbb{R}^e, \mathbb{R}^e\right)$ 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 $t \in [a, b]$.
To better illustrate this idea, we can define $V$ as a function with two arguments $V(X, Y)$ where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^e$ and $V(X, Y) \in \mathbb{R}^e$. Then $V(X, \cdot) : \mathbb{R}^d \mapsto \mathbf{L}\left(\mathbb{R}^e, \mathbb{R}^e\right)$ , and $V(\cdot, Y) : \mathbb{R}^e \mapsto \mathbf{L}\left(\mathbb{R}^d, \mathbb{R}^e\right)$.

# Picard iteration

Denoting by $I_e$ the identity operator (or matrix) in $\mathbf{L}\left(\mathbb{R}^e, \mathbb{R}^e\right)$, it follows that the iterations of $F$ can be expressed as follows:

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

# Picard iteration

Each quantity

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

can naturally be defined as an element of $\mathbf{L}(\mathbb{R}^e, \mathbb{R}^e)$, which, one can check, is determined 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 (18) will also agree at time $t$ for any choice of the linear vector fields $V$.

# Geometric intuition

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 $\left(S(X)^1_{a,b}, \ldots, S(X)^d_{a,b}\right)$, 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 $\left(X^i_b - X^i_a\right)^2/2$.

To give meaning to the term $S(X)^{i,j}_{a,b}$ for $i \neq j$, consider the Lévy area, 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 $\left\{X^1_t, X^2_t\right\}$ is given by:

$$A = \frac{1}{2}\left(S(X)^{1,2}_{a,b} - S(X)^{2,1}_{a,b}\right). \tag{19}$$

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.

# Geometric intuition

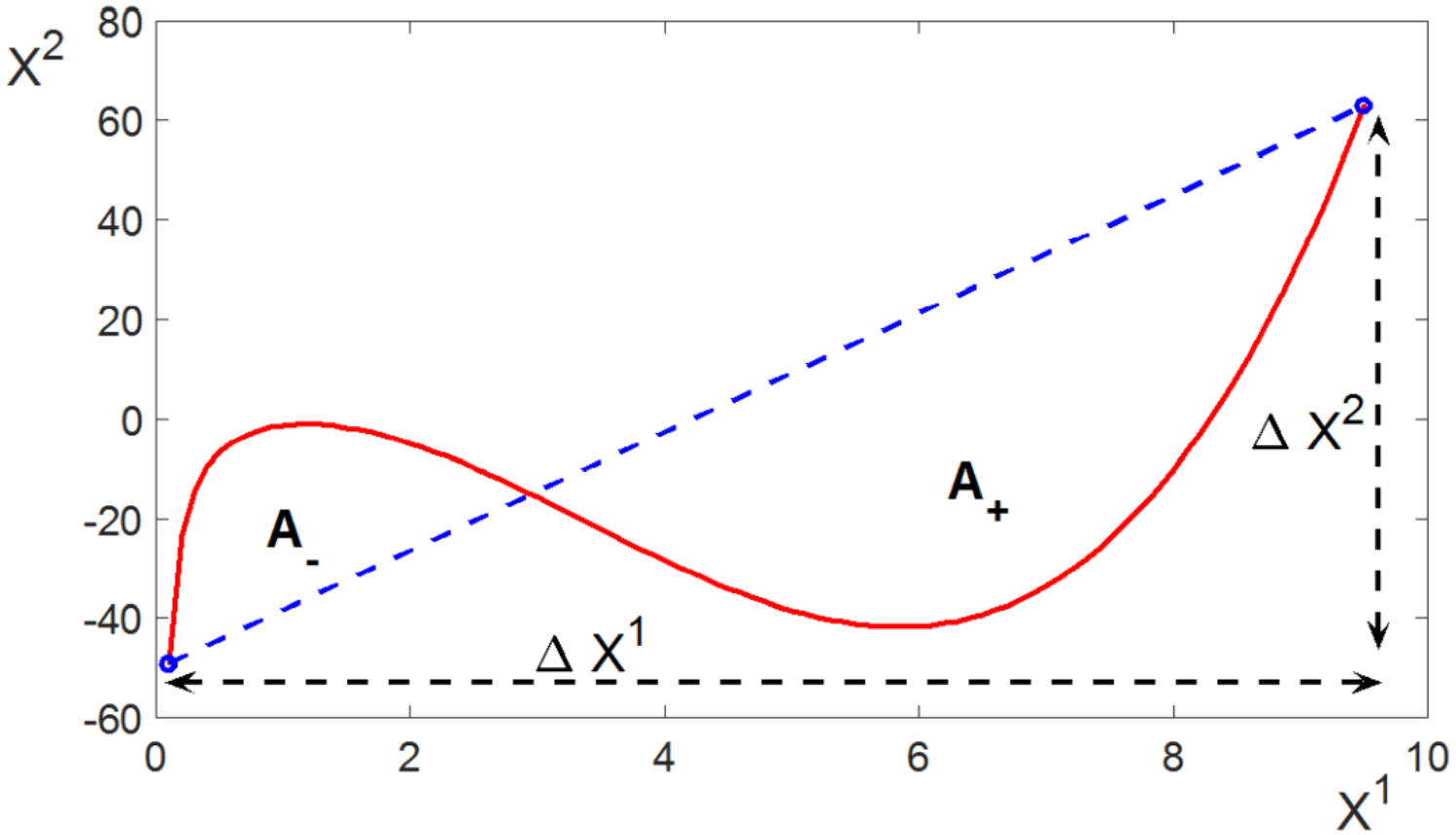



Figure 4 : Example of signed Levy area of a curve. Areas above and under the chord connecting two endpoints are negative and positive respectively.

# 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 $\tilde{X}, \tilde{Y} : [a, b] \mapsto \mathbb{R}$ by $\tilde{X}_t = X_{\psi(t)}$ and $\tilde{Y}_t = Y_{\psi(t)}$. Observe that

$$\dot{\tilde{X}}_t = \dot{X}_{\psi(t)}\dot{\psi}(t), \tag{20}$$

# Invariance under time reparametrisations

From which it follows that

$$\int_a^b \tilde{Y}_t d\widetilde{X}_t = \int_a^b Y_{\psi(t)} \dot{X}_{\psi(t)} \dot{\psi}(t) dt = \int_a^b Y_u dX_u \tag{21}$$

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.

# Invariance under time reparametrisations

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 $\tilde{X} : [a, b] \mapsto \mathbb{R}^d$ 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

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

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

# Shuffle product

One of the fundamental properties of the signature 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.

# Shuffle product

## Definition: Shuffle product

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

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

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

$$I \sqcup\!\sqcup J = \left\{ \left(r_{\sigma(1)}, \dots, r_{\sigma(k+m)}\right) \mid \sigma \in \text{Shuffles}(k, m) \right\}$$

# Shuffle product

## Theorem: 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

$$S(X)^I_{a,b} S(X)^J_{a,b} = \sum_{K \in I \sqcup\!\sqcup J} S(X)^K_{a,b}.$$

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

$$S(X)^1_{a,b} S(X)^2_{a,b} = S(X)^{1,2}_{a,b} + S(X)^{2,1}_{a,b}$$
$$S(X)^{1,2}_{a,b} S(X)^1_{a,b} = 2S(X)^{1,1,2}_{a,b} + S(X)^{1,2,1}_{a,b}.$$

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.

# Shuffle product

## Proof

$$S(X)^I_{a,b} S(X)^J_{a,b}$$

$$= \underbrace{\int_0^1 \cdots \int_0^1}_{I} 1_{\{0<t_{i_1}<\cdots t_{i_I}<1\}} dX^{i_1}_{t_{i_1}} \ldots dX^{i_I}_{t_{i_I}}$$

$$\underbrace{\int_0^1 \cdots \int_0^1}_{J} 1_{\{0<t_{j_1}<\cdots t_{j_J}<1\}} dX^{j_1}_{t_{j_1}} \ldots dX^{j_J}_{t_{j_J}}$$

$$= \underbrace{\int_0^1 \cdots \int_0^1}_{I+J} 1_{\{0<t_{i_1}<\cdots t_{i_I}<1, 0<t_{j_1}<\cdots t_{j_J}<1\}} dX^{i_1}_{t_{i_1}} \ldots dX^{i_I}_{t_{i_I}} dX^{j_1}_{t_{j_1}} \ldots dX^{j_J}_{t_{j_J}}$$

# 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.

## Definition: 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

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

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.

# Chen's identity

A (non-commuting) formal polynomial is a formal power series for which only a finite number of coefficients $\lambda_{i_1,\ldots,i_k}$ are non-zero. The terms $e_{i_1} \ldots e_{i_k}$ are called monomials. The term corresponding to $k = 0$ is simply just a real number $\lambda_0$. 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

$$\left( \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k \in \{1,\ldots d\}} \lambda_{i_1,\ldots,i_k} e_{i_1} \ldots e_{i_k} \right) + \left( \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k \in \{1,\ldots d\}} \mu_{i_1,\ldots,i_k} e_{i_1} \ldots e_{i_k} \right)$$
$$= \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k \in \{1,\ldots d\}} \left( \lambda_{i_1,\ldots,i_k} + \mu_{i_1,\ldots,i_k} \right) e_{i_1} \ldots e_{i_k} \tag{22}$$

# Chen's identity

and

$$c\left(\sum_{k=0}^{\infty}\sum_{i_1,\ldots,i_k\in\{1,\ldots d\}}\lambda_{i_1,\ldots,i_k}e_{i_1}\ldots e_{i_k}\right)=\sum_{k=0}^{\infty}\sum_{i_1,\ldots,i_k\in\{1,\ldots d\}}c\lambda_{i_1,\ldots,i_k}e_{i_1}\ldots e_{i_k}. \tag{23}$$

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

$$e_{i_1}\ldots e_{i_k}\otimes e_{j_1}\ldots e_{j_m}=e_{i_1}\ldots e_{i_k}e_{j_1}\ldots e_{j_m}. \tag{24}$$

# Chen's identity

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

$$\left(\sum_{k=0}^{\infty}\sum_{i_1,\ldots,i_k\in\{1,\ldots d\}}\lambda_{i_1,\ldots,i_k}e_{i_1}\ldots e_{i_k}\right)\otimes\left(\sum_{k=0}^{\infty}\sum_{i_1,\ldots,i_k\in\{1,\ldots d\}}\mu_{i_1,\ldots,i_k}e_{i_1}\ldots e_{i_k}\right)$$
$$=\lambda_0\mu_0+\sum_{i=1}^{d}\left(\lambda_0\mu_i+\lambda_i\mu_0\right)e_i+\sum_{i,j=1}^{d}\left(\lambda_0\mu_{i,j}+\lambda_i\mu_j+\lambda_{i,j}\mu_0\right)e_ie_j+\ldots \tag{25}$$

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

# Chen's identity

One 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

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

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.

# Chen's identity

## Definition: 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.

## Chen's identity

Let $X : [a, b] \mapsto \mathbb{R}^d$ and $Y : [b, c] \mapsto \mathbb{R}^d$ be two paths. Then

$$S(X * Y)_{a,c} = S(X)_{a,b} \otimes S(Y)_{b,c}. \tag{26}$$

# Chen's identity

## Proof

$$\sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k \in \{1,\cdots,d\}} \underbrace{\int_a^c \cdots \int_a^c}_{k} 1_{\{a<t_1<\cdots t_k<c\}} dZ_{t_{i_1}}^{i_1} \ldots dZ_{t_{i_l}}^{i_l} e^{i_1} \ldots e^{i_k}$$

$$= \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k \in \{1,\cdots,d\}} \Big( \sum_{1 \le l \le k} \underbrace{\int_a^b \cdots \int_a^b}_{l} 1_{\{a<t_1<\cdots t_l<b\}} dX_{t_{i_1}}^{i_1} \ldots dX_{t_{i_l}}^{i_l}$$

$$\underbrace{\int_b^c \cdots \int_b^c}_{k-l} 1_{\{b<t_{l+1}<\cdots t_k<c\}} dY_{t_{i_{l+1}}}^{i_{l+1}} \ldots dY_{t_{i_k}}^{i_k} \Big) e^{i_1} \ldots e^{i_k}$$

$$= \sum_{k=0}^{\infty} \sum_{1 \le l \le k} \Big( \sum_{i_1,\ldots,i_l \in \{1,\cdots,d\}} S(X)_{a,b}^{i_1,\ldots,i_l} e^{i_1} \ldots e^{i_l} \Big) *$$

$$\Big( \sum_{i_{l+1},\ldots,i_k \in \{1,\cdots,d\}} S(X)_{a,b}^{i_{l+1},\ldots,i_k} e^{i_{l+1}} \ldots e^{i_k} \Big)$$

# Time-reversal signature

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

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 $t \in [a, b]$

# Time-reversal signature

**Theorem**

For a path $X : [a, b] \mapsto \mathbb{R}^d$, it holds that

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

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 $\otimes$.

# Time-reversal signature

## Proof

Define a path $Y : [b, 2b-a] \mapsto \mathbb{R}^d$ such that $Y_t = \overleftarrow{X}_{t+b-a}$. Let's denote $X * Y$ as $Z$.

$$S(X)_{a,b} \otimes S(\overleftarrow{X})_{a,b} = S(X)_{a,b} \otimes S(Y)_{b,2b-a} = S(Z)_{b,2b-a}$$

$$S(Z)^{i_1,\cdot,i_n}_{a,2b-a} = \sum_{k=0}^{n} \int_{a<t_1<\cdots<t_k<b<t_{k+1}<\cdots<t_n<2b-a} dZ_{t_1} dZ_{t_2} \cdots dZ_{t_n}$$

$$= \sum_{k=0}^{n} \int_{a<t_1<\cdots<t_k<b} dZ_{t_1} \cdots dZ_{t_k} \otimes \int_{b<t_{k+1}<\cdots<t_n<2b-a} dZ_{t_k} \cdots dZ_{t_n}$$

$$= \sum_{k=0}^{n} \int_{a<t_1<\cdots<t_k<b} dX_{t_1} \cdots dX_{t_k} \otimes \int_{b<t_{k+1}<\cdots<t_n<2b-a} dY_{t_k} \cdots dY_{t_n}$$

$$= \sum_{k=0}^{n} \int_{a<t_1<\cdots<t_k<b} dX_{t_1} \cdots dX_{t_k} \otimes \int_{a<t_{k+1}<\cdots<t_n<b} d\overleftarrow{X}_{t_k} \cdots d\overleftarrow{X}_{t_n}$$

$$= \sum_{k=0}^{n} \int_{a<t_1<\cdots<t_k<b} dX_{t_1} \cdots dX_{t_k} \otimes (-1)^{n-k} \int_{a<t_{k+1}<\cdots<t_k<b} dX_{t_k} \cdots dX_{t_n}$$

# Time-reversal signature

## Proof

Let's define

$$J_k = \int_{a<t_1<\cdots<t_k<b} dX_{t_1}\cdots dX_{t_k} = \frac{1}{k!}\int_{[a,b]^k} (dX_t)^{\otimes k} = \frac{1}{k!}(X_b - X_a)^{\otimes k}$$

$$S(Z)^{i_1,\cdot,i_n}_{a,2b-a} = \sum_{k=0}^{n} J_k \otimes (-1)^{n-k} J_{n-k}$$

$$= \frac{(x_b - x_a)^{\otimes n}}{n!}\sum_{k=0}^{n}(-1)^{n-k}\frac{n!}{k!(n-k)!}$$

$$= \frac{(x_b - x_a)^{\otimes n}}{n!}\sum_{k=0}^{n}(-1)^{n-k}\cdot C_n^k$$

$$= \frac{(x_b - x_a)^{\otimes n}}{n!}(1-1)^n = 0$$

the last equation followed from $(x+y)^n = \sum_{k=0}^{n} C_n^k x^k y^{n-k} = 0$.

# 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

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

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

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

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

# Log signature

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

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

one can readily check that

$$\log x = \lambda e_1.$$

# Log signature

## Definition: 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

$$[x, y] = x \otimes y - y \otimes x.$$

# Log signature

## Theorem

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

$$\log S(X)_{a,b} = \sum_{k\geq 1} \sum_{i_1,\ldots,i_k \in \{1,\ldots,d\}} \lambda_{i_1,\ldots,i_k} \left[e_{i_1}, \left[e_{i_2}, \ldots, \left[e_{i_{k-1}}, e_{i_k}\right] \ldots\right]\right]$$

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

# Log signature

## Baker-Campbell-Hausdorff formula

Baker-Campbell-Hausdorff formula is the solution for $Z$ to the equation

$$e^X e^Y = e^Z$$

for possibly non-commutative $X$ and $Y$ in the Lie algebra of a Lie group.

$$\begin{aligned} Z(X,Y) &= \log(\exp X \exp Y) \\ &= X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}([X,[X,Y]] + [Y,[Y,X]]) \\ &\quad - \frac{1}{24}[Y,[X,[X,Y]]] \\ &\quad - \frac{1}{720}([Y,[Y,[Y,[Y,X]]]] + [X,[X,[X,[X,Y]]]]) \end{aligned}$$

# Log signature

## Baker-Campbell-Hausdorff formula: Continue

$$+\frac{1}{360}([X,[Y,[Y,[Y,X]]]]+[Y,[X,[X,[X,Y]]]])$$
$$+\frac{1}{120}([Y,[X,[Y,[X,Y]]]]+[X,[Y,[X,[Y,X]]]])$$
$$+\frac{1}{240}([X,[Y,[X,[Y,[X,Y]]]]])$$
$$+\frac{1}{720}([X,[Y,[X,[X,[X,Y]]]]]-[X,[X,[Y,[Y,[X,Y]]]]])$$
$$+\frac{1}{1440}([X,[Y,[Y,[Y,[X,Y]]]]]-[X,[X,[Y,[X,[X,Y]]]]])+\cdots$$

# Log signature

## Example

Consider the two-dimensional path

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

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

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

# Log signature

## Example

It follows by Chen's identity that

$$S(X)_{0,2} = S(Y)_{0,1} \otimes S(Z)_{1,2} = 1 + e_1 + e_2 + \frac{1}{2!}e_1 + \frac{1}{2!}e_2 + e_1e_2 + \dots$$

It can be easily seen that

$$e^{log(S(Y)_{0,1})} = S(Y)_{0,1}, e^{log(S(Z)_{1,2})} = S(Z)_{0,1}.$$

By Baker-Campbell-Hausdorff formula,

$$log(S(X)_{0,2}) = log(S(Y)_{0,1} \otimes S(Z)_{1,2})$$

$=\log(S(Y)_{0,1}) + log(S(Z)_{1,2}) + \frac{1}{2}[log(S(Y)_{0,1}), log(S(Z)_{1,2})] + \dots$
$= e_1 + e_2 + \frac{1}{2}[e_1, e_2] + \dots$

# Young's integral

We say that the Stieltjes integral

$$\int_a^b f(x)\,\mathrm{d}g(x)$$

exists in the Riemann sense with the value $I$, if the sum

$$\sum_{i=1}^{n} f\left(\xi_i\right)\left\{g\left(x_i\right)-g\left(x_{i-1}\right)\right\},$$

where $a = x_0 < x_1 < \ldots < x_n = b$ and $\xi_i \in [x_{i-1}, x_i]$ is as close to $I$ as we wish, provided the mesh of the partition $\pi = \{x_0, x_1, \ldots, x_n\}$,

$$\operatorname{mesh}(\pi) := \max_{1\le i\le n}\left(x_i - x_{i-1}\right),$$

is sufficiently small.

# Young's integral

There is no problem with the existence of the Riemann-Stieltjes integral if the total variation of the integrator

$$\mathrm{TV}(g,[a,b]) := \sup_n \sup_{a\le x_0<x_1<\ldots<x_n\le b} \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})|$$

is finite, the integrand $f$ is bounded and regulated (has left and right limits), and $f$ and $g$ have no common points of discontinuity. If $g$ is not of bounded variation, then there will be continuous functions which cannot be integrated with respect to $g$. By the integration by parts formula

$$\int_a^b f(x)\mathrm{d}g(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(x)\mathrm{d}f(x)$$

it is easy to see that the integral also exists whenever the total variation of $f$ is finite, $g$ is bounded and regulated, and $f$ and $g$ have no common points of discontinuity.

# Young's integral

Young was considering the functions of finite $p$ and $q$-variations. Let us recall the definition of $p$-variation $(p > 0)$ : for any $f : [a, b] \to \mathbb{R}$

$$\mathrm{V}^p(f, [a, b]) := \sup_n \sup_{a \le x_0 < x_1 < \ldots < x_n \le b} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|^p .$$

This may be generalized and defined for any $f : [a, b] \to E$ attaining its values in a metric space $E$ with the metric $d$ :

$$\mathrm{V}^p(f, [a, b]) := \sup_n \sup_{a \le x_0 < x_1 < \ldots < x_n \le b} \sum_{i=1}^{n} d(f(x_i), f(x_{i-1}))^p$$

# Young's integral

Strong $p$-variation may be viewed as a measure of path irregularity. If $\mathrm{V}^p(f,[a,b]) < +\infty$ for $p \geq 1$ then $\mathrm{V}^q(f,[a,b]) < +\infty$ for all $q > p$. This follows from inequality $(\sum_i |a_i|^q)^{1/q} \leq (\sum_i |a_i|^p)^{1/p}$. This is the reason to introduce the variation index:

$$\mathrm{Ind}_{\mathrm{var}}(f,[a,b]) := \inf\{p \geq 1 : \mathrm{V}^p(f,[a,b]) < +\infty\}$$

Another measure of path irregularity is Hölder exponent: we say that the function $f : [a,b] \to \mathbb{R}$ is Hölder continuous with the Hölder exponent $\alpha$ if

$$\sup_{a \leq s < t \leq b} \frac{|f(t) - f(s)|}{|t-s|^\alpha} < +\infty.$$

If $f$ is Hölder continuous with the Hölder exponent $0 < \alpha \leq 1$ then $\mathrm{V}^{1/\alpha}(f,[a,b]) < +\infty$

# Young's integral

## Theorem

The Stieltjes integral $\int_a^b f(x)\mathrm{d}g(x)$ exists in the Riemann sense whenever $f$ and $g$ have no common discontinuities, and $V^p(f,[a,b]) < +\infty$, $V^q(g,[a,b]) < +\infty$ for some $p > 0, q > 0$ such that $p^{-1} + q^{-1} > 1$. Moreover, for any $\xi \in [a,b]$ one has the following estimate

$$\left| \int_a^b f(x)\mathrm{d}g(x) - f(\xi)[g(b) - g(a)] \right|$$
$$\leq 2\left(1 + \zeta\left(\frac{1}{p} + \frac{1}{q}\right)\right)\left(V^p(f,[a,b])\right)^{1/p}\left(V^q(g,[a,b])\right)^{1/q}$$

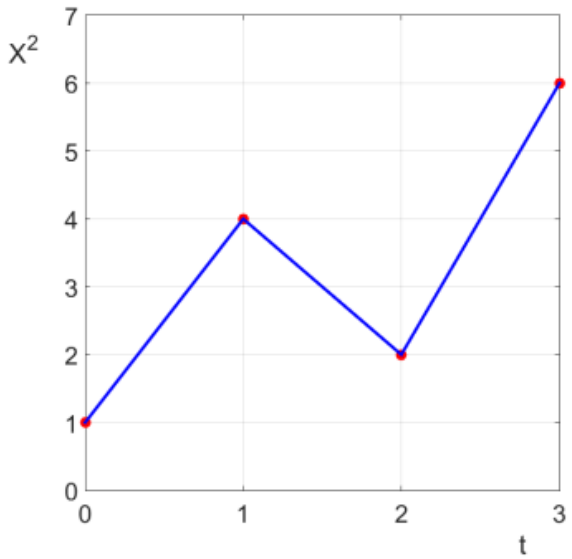
# Paths from discrete data

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

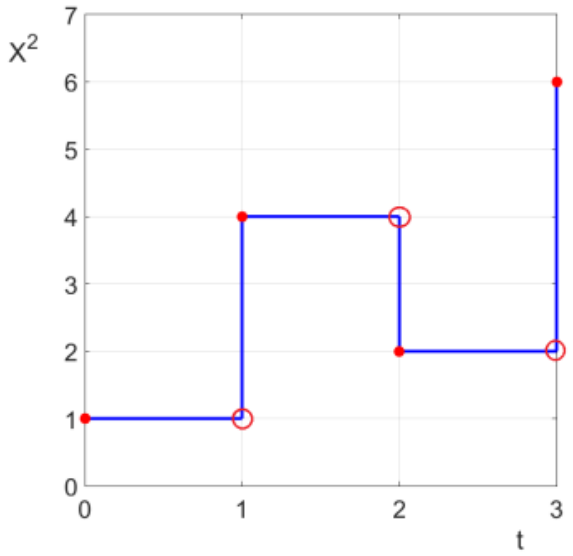
$$\{X_i^1\}_{i=1}^4 = \{1, 3, 5, 8\}$$
$$\{X_i^2\}_{i=1}^4 = \{1, 4, 2, 6\}$$
$$\{t_i\}_{i=1}^4 = \{0, 1, 2, 3\},$$

where the variable $\{t_i\}$ corresponds to 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).

# Paths from discrete data




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




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

Figure 5

# The signature of paths

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

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

and

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

where the last term in $\log S(X)$ is given by $\frac{1}{2}\left(S^{(1,2)} - S^{(2,1)}\right)$ 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. 7.

# The signature of paths

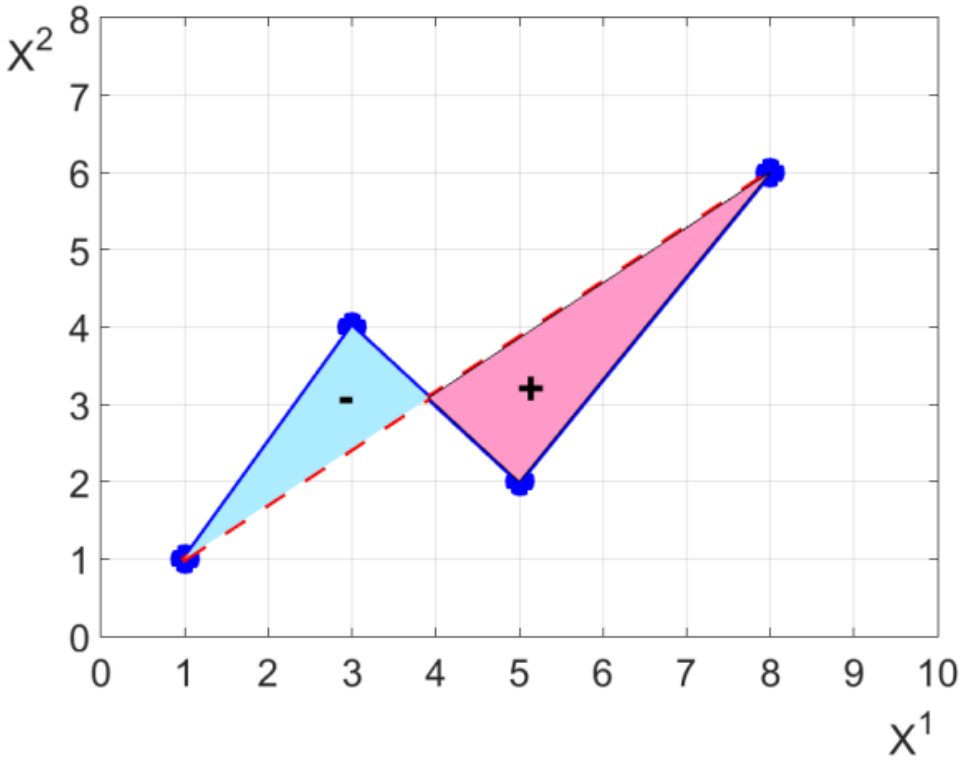
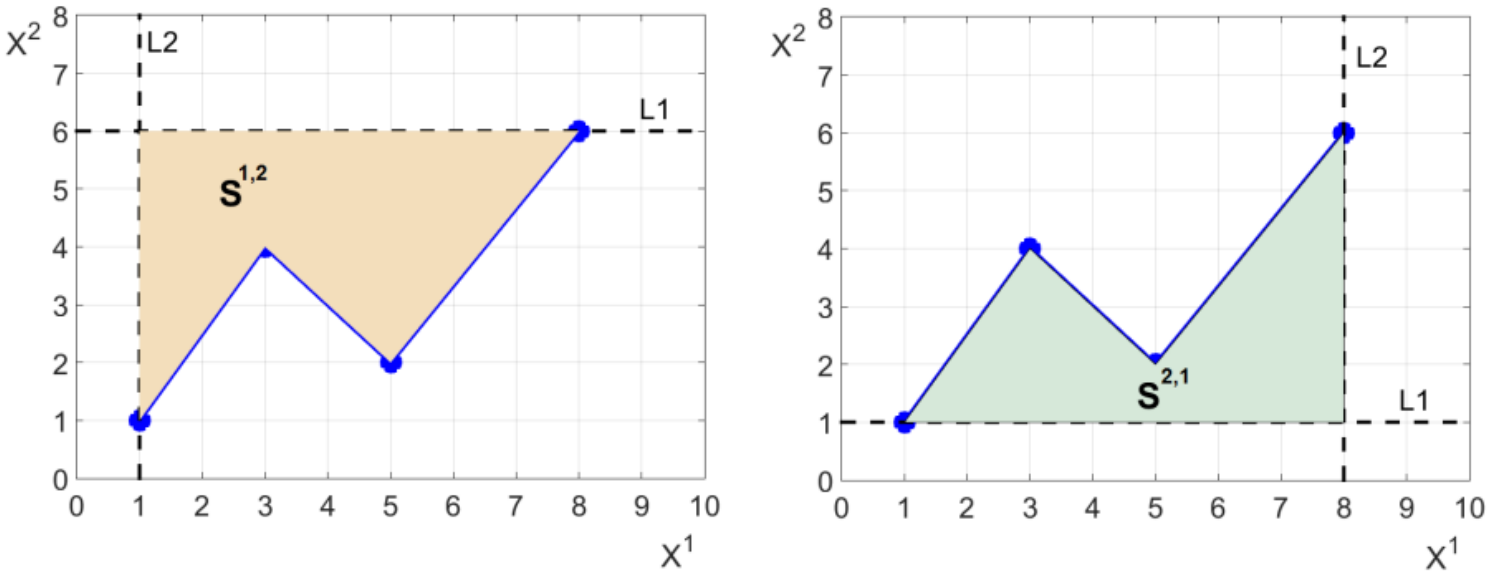



Figure 6: Example of signed area enclosed by the piece-wise linear path (blue) and the chord (red dashed line). The light blue area is negative and pink area is positive.

# The signature of paths




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

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

Figure 7: 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.

# The signature of paths

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:

$$S^{(1)} \cdot S^{(2)} = S^{(1,2)} + S^{(2,1)}$$

$$5 \cdot 7 = 19 + 16$$

# The lead-lag transformation

Now we are going to introduce the Lead-Lag transformation of data, which maps a one-dimensional path into a two-dimensional path. Considering the sequence $X_i$, the Lead-Lag mapping is given by:

$$X^2 = \{1, 4, 2, 6\} \mapsto \begin{cases} X^{2,\text{ Lead}} & = \{1, 4, 4, 2, 2, 6, 6\} \\ X^{2,\text{ Lag}} & = \{1, 1, 4, 4, 2, 2, 6\} \end{cases}$$

and the resulting embedded path is presented in Fig. 9 with three additional points $\{(1, 4), (4, 2), (2, 6)\}$

# The lead-lag transformation

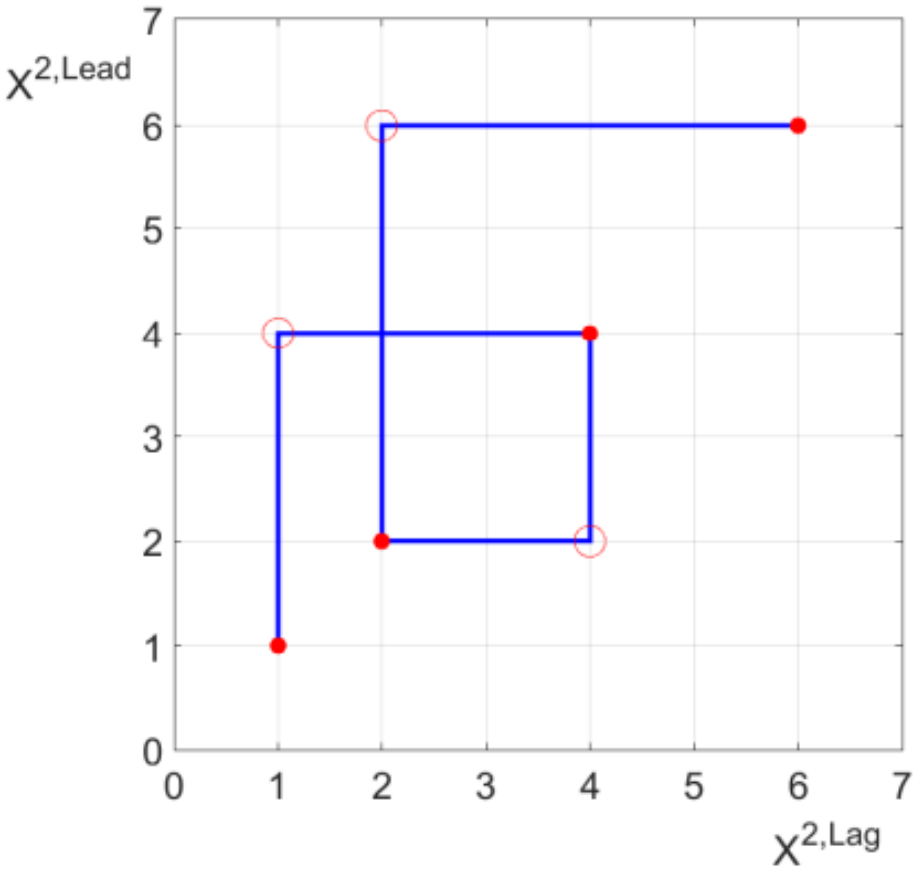



Figure 9: Lead-Lag transform of one-dimensional data $X_i^2$

# Sign of enclosed area

Consider the six possibilities in the following figure,

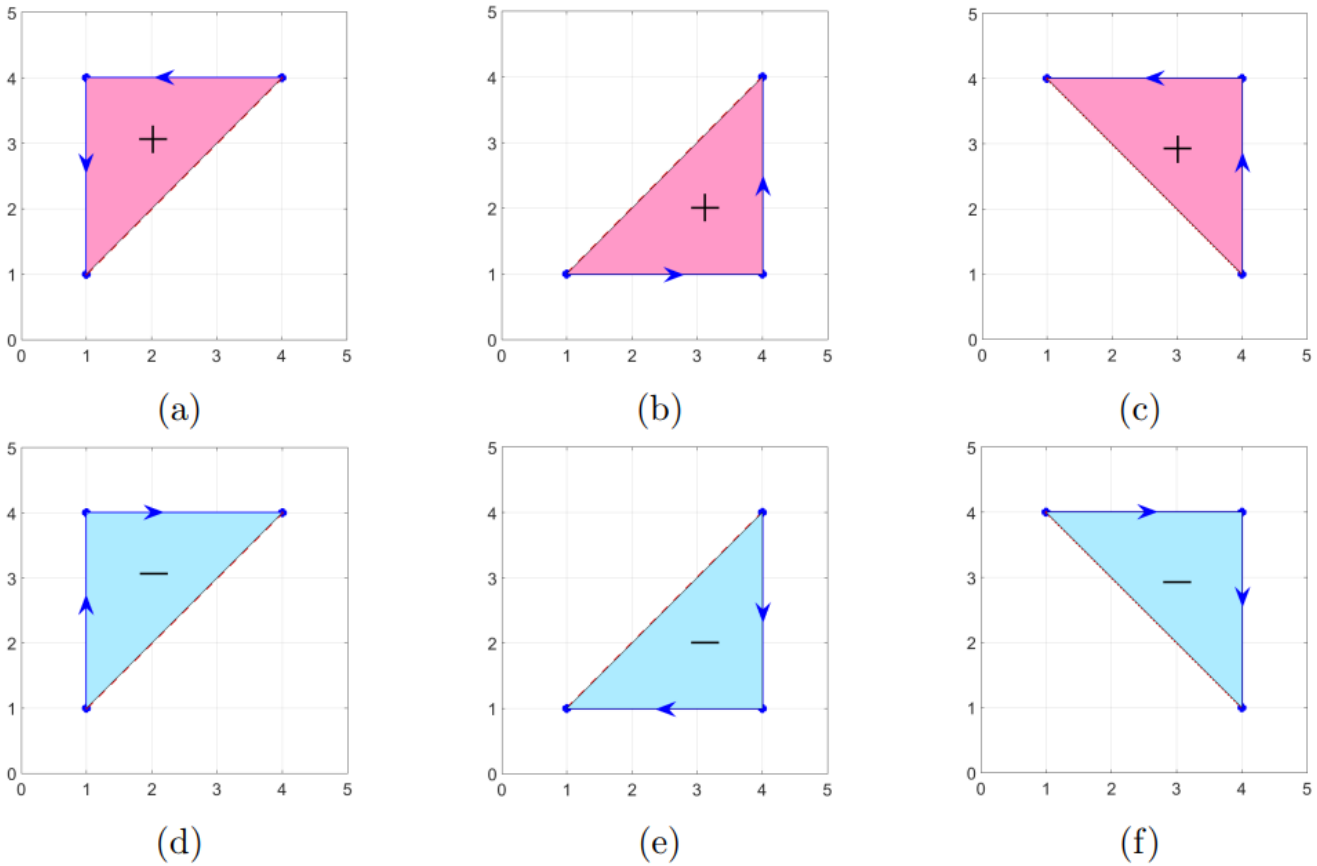



Figure 10: Various possibilities of signed area. The first row (a)-(c) corresponds to counter clock-wise movement along the path, and the bottom row (d)-(f) to clock-wise movement.

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

We can decompose figure 9 into three right-angled isosceles triangles Fig. 11. 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.

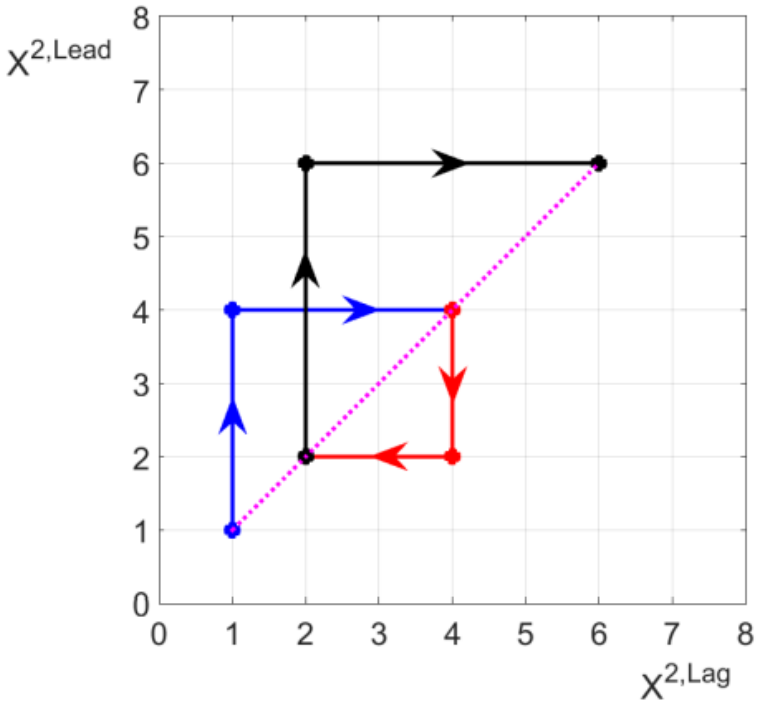



Figure 11

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

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

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

Let us write:

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

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.

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

Thus one can generally write for any sequence $\{X_i\}_{i=1}^{N}$

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

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

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

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

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

# The cumulative sum of a sequence

Next we explore certain properties of paths which originate from embedding points using cumulative sums. The cumulative sum of a sequence $X^2$ is:

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

Explicitly, for any general sequence $\{X_i\}$ :

$$\{X\}_i \to \left\{0, \{X\}_{i=1}^N\right\} \to CS\left(\{X\}_i\right) = \{\tilde{X}\}_{i=0}^N = \{0, X_1, X_1 + X_2, \ldots\}.$$

For an arbitrary series $\{X\}_{i=1}^N$, we define the following terms:

# The cumulative sum of a sequence

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

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

We can see that, for a collection of data points $\{X_i\}_{i=1}^{N}$ :

$$\text{Mean}(X) = E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{\Delta\tilde{X}}{N}$$

$$\begin{aligned} \text{Var}(X) &= E\left[(X - F[X])^2\right] && = E\left[X^2\right] - (E[X])^2 \\ &= \frac{1}{N}\left(QV(\tilde{X}) - \frac{1}{N}(\Delta\tilde{X})^2\right) \end{aligned}$$

# The cumulative sum of a sequence

Remember that the first two levels of the signature (log or full) correspond to the total increment and the Levy area, which shows how these signature terms determine the first two statistical moments.

# Signature terms as a function of data points

Consider the cumulative lead-lag embedding of $N$ data points $\{X_i\}_{i=1}^N$ into a continuous path, then the resulting truncated signature at level $L = 2$ is simply given by:

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

with

$$S^{(1)} = S^{(2)} = \sum_i X_i$$

$$S^{(1,1)} = S^{(2,2)} = \frac{1}{2}\left(\sum_i X_i\right)^2$$

$$S^{(1,2)} = \frac{1}{2}\left[\left(\sum_i X_i\right)^2 - \sum_i X_i^2\right]$$

# Signature terms as a function of data points

$$S^{(2,1)} = \frac{1}{2}\left[\left(\sum_i^N X_i\right)^2 + \sum_i^N X_i^2\right]$$

After some computations, we have the following result:

$$\text{Mean}(X) = \frac{1}{N} S^{(1)}$$

$$\text{Var}(X) = -\frac{N+1}{N^2} S^{(1,2)} + \frac{N-1}{N^2} S^{(2,1)},$$

where $N$ is the total number of data points in the sample.