*The Annals of Statistics*
2011, Vol. 39, No. 4, 2047–2073
DOI: 10.1214/11-AOS893


# PARAMETER ESTIMATION FOR ROUGH DIFFERENTIAL EQUATIONS[1]

BY ANASTASIA PAPAVASILIOU AND CHRISTOPHE LADROUE

*University of Warwick and University of Crete, and University of Bristol*

We construct the "expected signature matching" estimator for differential equations driven by rough paths and we prove its consistency and asymptotic normality. We use it to estimate parameters of a diffusion and a fractional diffusions, that is, a differential equation driven by fractional Brownian motion.

**1. Introduction.** Statistical inference for stochastic processes is a huge field, both in terms of research output and importance. In particular, a lot of work has been done in the context of diffusions (see [3, 15, 22] for a general overview and [2] for some recent developments). Nevertheless, the problem of statistical inference for diffusions still poses many challenges, as, for example, constructing the Maximum Likelihood Estimator (MLE) for the general multi-dimensional diffusion. An alternative method in this case is that of the Generalized Moment Matching Estimator (GMME). While, in general, less efficient compared to the MLE, the GMME is usually easier to use, more flexible and has been successfully applied to general Markov processes (see [1, 10]).

On the other hand, most methods of statistical inference in the context of non-Markovian continuous processes are restricted to specific classes of models. In the case of differential equations driven by fractional Brownian motion, some recent results can be found in [3, 11, 24]. In [12], the author discusses the problem of parameter estimation for differential equations driven by Volterra type processes—which include fractional Brownian motion. In all these papers, the analysis is restricted to models that depend linearly on the parameter and for parameters appearing in the drift. Finally, for non-Markovian processes coming from stochastic delay equations, see [14, 23].

The theory of rough paths provides a general framework for making sense of differential equations driven by any type of noise modelled as a rough path—this includes diffusions, differential equations driven by fractional Brownian motion, delay equations and even delay equation driven by fractional Brownian motion

Received June 2010; revised January 2011.

[1]Supported in part by EP/H019588/1 Grant "Parameter Estimation for Rough Differential Equations with Applications to Multiscale Modelling" and FP7-REGPOT-2009-1 project "Archimedes Center for Modeling, Analysis and Computation."

*MSC2010 subject classifications.* Primary 62F12, 62F12; secondary 62M99.

*Key words and phrases.* Rough paths, diffusions, fractional diffusions, generalized moment matching, parameter estimation.

(see [20]). The basic ideas have been developed in the 90s (see [18] and references within). However, the problem of statistical inference for differential equations driven by rough paths has not been addressed yet. This is exactly what we strive to do in this paper.

The exact setting of the statistical problem we consider is the following: *we observe many independent copies of specific iterated integrals of the response* $\{Y_t, 0 < t < T\}$ of a differential equation

$$dY_t = f(Y_t; \theta) \cdot dX_t, \qquad Y_0 = y_0,$$

driven by the *rough path* $X$. We will formally define what we mean by a rough path and a differential equation driven by it in Section 2.1. Two examples of interest are $X_t = (t, W_t)$ where $W_t$ is Brownian motion and the differential equation is a Stratonovich stochastic differential equation and $X_t = (t, B_t^H)$ where $B_t^H$ is fractional Brownian motion. *The iterated integrals are observed at a fixed time* $T$. In this sense, our setting is similar to [7]. However, if the response lives in more than one dimension, the iterated integrals could be functions of the whole path. For example, suppose that $Y_t = (Y_t^{(1)}, Y_t^{(2)})$ and we observe

$$\int\int_{0<u_1<u_2<T} dY_{u_1}^{(2)}\, dY_{u_2}^{(1)}$$

for fixed time $T$. We further assume that *the vector field* $f(y; \theta)$ *is polynomial in* $y$ *and depends on the unknown parameter* $\theta$. Finally, we assume that *we know the* expected signature *of the rough path* $X$ *on the interval* $[0, T]$, to be formally defined later. For now, let's just say that it is the set of all iterated integrals of $X$ and its expectation fully describes the distribution of the rough path $X$.

The first assumption is a bit unusual: it is much more common to assume that we observe one long path rather than many short ones. This setting is chosen for two reasons. The first is its simplicity: we develop here some basic tools for statistical inference of differential equation driven by rough paths.

The second reason was that such settings arise in the context of "equation-free" modelling of multiscale models (see [17]). Suppose that we have access to some code that simulates the dynamics of a complex system, such as molecular dynamics. We treat the code as a "black box." We are interested in the global behavior of a function of our system that "lives" in the slow scale, that is, in some limit its dynamics follow a diffusion, which is, however, unknown. The basic idea of "equation-free" modelling is to run the code for a *short time* and use the output to *locally estimate* the parameters of the differential equation. This process is repeated several times with carefully chosen initial conditions, so as to get an estimate of the global dynamics. To summarize, in this problem:

(a) we observe many independent paths;
(b) time is short;
(c) we locally approximate the vector field by a polynomial.

Currently, the estimation is done using the MLE approach, pretending that the data comes from the diffusion rather than the multiscale model (see [4]). However, for short time $T$ we cannot expect the diffusion approximation to be a good one. We believe that in the scale of $T$, we can always approximate the dynamics by a differential equation driven by a rough path (see [21]).

However, the method can be generalized to other settings, such as observing one continuous path, provided that some ergodicity conditions are fulfilled. We also describe the methodology for this setting and demonstrate it with an example. Note though that in the general setting of rough paths, ergodicity theory has not yet been developed and has to be checked for each case separately. For some recent results on the ergodicity of differential equations driven by fractional Brownian motion see [8].

The structure of the paper is the following: we start by reviewing some basic concepts and results from the theory of rough paths and we give a precise description of the problem we consider. In Section 3, we describe the methodology. The idea is simple: we want to match the theoretical and the expected signatures of the response. However, in general we cannot expect to get an explicit formula for the theoretical expected signature, so we construct an approximation of it. We go on to give a precise definition of the "expected signature matching estimator" using this approximation and prove its consistency and asymptotic normality.

In Section 4, we extend the methodology to the setting where we observe one path of a stationary ergodic process and we discuss optimality.

In Section 5, we apply the methods to three examples that represent the most common RDEs: diffusions and differential equations driven by fractional Brownian motion. We have written a package in *Mathematica* that is publicly available from http://chrisladroue.com/software/brownian-motion-and-iterated-integrals-on-mathematica/ and can be used to recreate the examples we include in the paper or try out new ones.

## 2. Setting.

2.1. *Some basic results from the theory of rough paths.* In this section, we review some of the basic results from the theory of rough paths. For more details, see [6, 19] and references within. The goal of this theory is to give meaning to the differential equation

$$dY_t = f(Y_t) \cdot dX_t, \qquad Y_0 = y_0, \tag{2.1}$$

for very general continuous paths $X$. More specifically, we think of $X$ and $Y$ as paths on a Euclidean space: $X : I \to \mathbb{R}^n$ and $Y : I \to \mathbb{R}^m$ for $I := [0, T]$, so $X_t \in \mathbb{R}^n$ and $Y_t \in \mathbb{R}^m$ for each $t \in I$. Also, $f : \mathbb{R}^m \to \mathrm{L}(\mathbb{R}^n, \mathbb{R}^m)$, where $\mathrm{L}(\mathbb{R}^n, \mathbb{R}^m)$ is the space of linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ which is isomorphic to the space of $m \times n$ matrices. For the sake of simplicity, we will assume that

$f(y)$ is a polynomial in $y$—however, the theory holds for more general $f$. The path $X$ is any path of finite $p$-variation, meaning that

$$\sup_{\mathcal{D}\subset[0,T]}\left(\sum_{\ell}\|X_{t_\ell}-X_{t_{\ell-1}}\|^p\right)^{1/p}<\infty,$$

where $\mathcal{D}=\{t_\ell\}_\ell$ goes through all possible partitions of $[0,T]$ and $\|\cdot\|$ is the Euclidean norm. Note that we will later define finite $p$-variation for multiplicative functionals, also to be defined later.

The fact the $X$ is allowed to have any finite $p$-variation is exactly what makes this theory so general: Brownian motion is an example of a path that has finite $p$-variation for any $p>2$ while fractional Brownian motion with Hurst index $h$ has finite $p$ variation for $p>\frac{1}{h}$. We will define fractional Brownian motion formally in the corresponding example—for now, let us just say that it is Gaussian, self-similar but not Markovian except for $h=1/2$ when it coincides with Brownian motion.

When $p\in[1,2)$, we say that $Y$ is a solution of (2.1) if

$$Y_t=Y_s+\int_s^t f(Y_u)\cdot dX_u \qquad \forall(s,t)\in\Delta_T,$$

where $\Delta_T:=\{(s,t);0\le s\le t\le T\}$. In this case, the integral is defined as the Young integral (see [18]). What does it mean for $Y$ to be a solution of (2.1) when $p\ge 2$? In order to answer this question, we first need to define the integral. To make this task possible, we rewrite the integral so that the integrand is a function of the integrator: set $f_{y_0}(\cdot):=f(\cdot+y_0)$. Define $h:\mathbb{R}^n\oplus\mathbb{R}^m\to\mathrm{End}(\mathbb{R}^n\oplus\mathbb{R}^m)$ by

$$h(x,y):=\begin{pmatrix} I_{n\times n} & \mathbf{0}_{n\times m}\\ f_{y_0}(y) & \mathbf{0}_{m\times m}\end{pmatrix}. \tag{2.2}$$

Instead of defining $\int_s^t f(Y_u)\cdot dX_u$, we will define the integral

$$\int_s^t h(Z_u)\cdot dZ_u \qquad \forall(s,t)\in\Delta_T, \tag{2.3}$$

where $Z=(X,Y)$. Note that if $f$ is a polynomial in $y$, then $h$ will also be a polynomial in $z$. More generally, we will define this integral for any path $Z$ in $\mathbb{R}^{\ell_1}$ of finite $p$-variation and any polynomial $h:\mathbb{R}^{\ell_1}\to\mathrm{L}(\mathbb{R}^{\ell_1},\mathbb{R}^{\ell_2})$ of degree $q$. Since $h$ is a polynomial, its Taylor expansion will be a finite sum:

$$h(z_2)=\sum_{k=0}^{q}h_k(z_1)\frac{(z_2-z_1)^{\otimes k}}{k!}\qquad \forall z_1,z_2\in\mathbb{R}^{\ell_1},$$

where $h_0=h$ and $h_k:\mathbb{R}^{\ell_1}\to\mathrm{L}(\mathbb{R}^{\ell_1 \otimes k},\mathrm{L}(\mathbb{R}^{\ell_1},\mathbb{R}^{\ell_2}))$ and for all $z\in\mathbb{R}^{\ell_1}$, $h_k(z)$ is a symmetric $k$-linear mapping from $\mathbb{R}^{\ell_1}$ to $\mathrm{L}(\mathbb{R}^{\ell_1},\mathbb{R}^{\ell_2})$, for $k\ge 1$.

Suppose that $Z$ is a path of bounded variation (i.e., $p = 1$). Then, using the symmetry of $h_k(z)$ and the "shuffle product property," we can write

$$h(Z_u) = \sum_{k=0}^{q} h_k(Z_s)\mathbf{Z}^k_{s,u} \qquad \forall (s,u) \in \Delta_T,$$

where for every $(s,t) \in \Delta_T$,

$$\mathbf{Z}^0 \equiv 1 \in \mathbb{R}$$

and

$$\mathbf{Z}^k_{s,t} = \left\{ \int \cdots \int_{s<u_1<\cdots<u_k<t} dZ^{(i_1)}_{u_1} \cdots dZ^{(i_k)}_{u_k} \right\}_{(i_1,\ldots,i_k)\in\{1,\ldots,n\}^k} \in \mathbb{R}^{\ell_1 \otimes k}.$$

More specifically, we use the notation

$$Z^{(i_1,\ldots,i_k)}_{s,t} := \int \cdots \int_{s<u_1<\cdots<u_k<t} dZ^{(i_1)}_{u_1} \cdots dZ^{(i_k)}_{u_k}.$$

The "shuffle product property" says that for any $(s,u) \in \Delta_T$ and any "words" $\sigma_1, \sigma_2 \in \bigcup_{k\geq 0}\{1,\ldots,\ell_1\}^k$, we can write

$$\mathbf{Z}^{\sigma_1}_{s,u}\mathbf{Z}^{\sigma_2}_{s,u} = \sum_{\sigma\in\sigma_1 \sqcup \sigma_2} \mathbf{Z}^{\sigma}_{s,u}, \tag{2.4}$$

where $\sigma_1 \sqcup \sigma_2$ is the *shuffle product* between the words $\sigma_1$ and $\sigma_2$, that is, it is the set of all words (with repetition) that we can create by mixing up the letters of $\sigma_1$ and $\sigma_2$ without changing the order of letters within each word. For example, $(1,2) \sqcup (2) = \{(1,2,2), (1,2,2), (2,1,2)\}$ (see [19]). This generalizes the "integration by parts" formula. Then, for all $(s,t) \in \Delta_T$,

$$\int_s^t h(Z_u)\,dZ_u = \sum_{k=0}^{q} h_k(Z_s)\mathbf{Z}^{k+1}_{s,t}.$$

EXAMPLE 1. Let us demonstrate what we have said so far with an example. Consider the ordinary differential equation

$$dY_t = Y_t\,dt + (Y_t^2 + 1)\,de^t, \qquad Y_0 = 0.$$

Then, $X_t = (t, e^t)$ is a path in $\mathbb{R}^2$, $Y_t \in \mathbb{R}$ and $f(y) = (y, y^2+1) \in \mathrm{L}(\mathbb{R}^2, \mathbb{R})$, which is polynomial of degree 2. In this case, $X$ is of bounded variation and $p = 1$. Following what we just mentioned, instead of defining the integral

$$\int_s^t f(Y_u)\,dX_u = \int_s^t \big(Y_u\,du + (Y_u^2+1)\,de^u\big)$$

directly, we set $Z_t = (X_t, Y_t)' = (t, e^t, Y_t)' \in \mathbb{R}^3$ and

$$h(Z_t) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ Z^{(3)}_t & (Z^{(3)}_t)^2 + 1 & 0 \end{pmatrix},$$

where $Z_t^{(3)} = Y_t$ is the projection of $Z_t$ to the third dimension. Then, the integral $\int_s^t h(Z_u)\,dZ_u$ becomes

$$\int_s^t h(Z_u)\,dZ_u = \left(0, 0, \int_s^t f(Y_u)\,dX_u\right),$$

so, defining $\int_s^t h(Z_u)\,dZ_u$ is equivalent to defining $\int_s^t f(Y_u)\,dX_u$. We now proceed to writing the integral as a linear combination of iterated integrals of $Z$, using the fact that $h$ is a quadratic polynomial. We define $h_k$ as

$$h_0(z) = h(z), \qquad h_1(z) = \{\partial_i h(z)\}_{i=1}^3, \qquad h_2(z) = \{\partial_{i_1,i_2} h(z)\}_{i_1,i_2=1}^3.$$

Also, we note that

$$\big((z_2 - z_1)^{\otimes 1}\big)_i = z_2^{(i)} - z_1^{(i)} \quad \text{and} \quad \big((z_2 - z_1)^{\otimes 2}\big)_{i_1,i_2} = \big(z_2^{(i_1)} - z_1^{(i_1)}\big)\big(z_2^{(i_2)} - z_1^{(i_2)}\big)$$

and thus the sum $\sum_{k=0}^2 h_k(z_1)\frac{(z_2-z_1)^{\otimes k}}{k!}$ becomes

$$\begin{pmatrix} 0 \\ 0 \\ z_1^{(3)} + \big(z_1^{(3)}\big)^2 + 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \big(1 + 2z_1^{(3)}\big)\big(z_2^{(3)} - z_1^{(3)}\big) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 2\dfrac{(z_2^{(3)} - z_1^{(3)})^2}{2} \end{pmatrix},$$

which is equal to $h(z_2)$. It is easy to see that for all $0 < s < t < T$,

$$\big(z_t^{(3)} - z_s^{(3)}\big) = \int_s^t dz_u^{(3)} \quad \text{and} \quad \frac{(z_t^{(3)} - z_s^{(3)})^2}{2} = \int_s^t \int_s^{u_1} dz_{u_1}^{(3)}\,dz_{u_2}^{(3)}.$$

Thus, using the notation of the iterated integral, we write

$$h(z_u) = h(z_s) + \partial_3 h(z_s) Z_{s,u}^{(3)} + \partial_{3,3}^2 h(z_s) Z_{s,u}^{(3,3)}$$

and if we integrate once more we get

$$\int_s^t h(z_u)\,du = h(z_s) Z_{s,t}^{(3)} + \partial_3 h(z_s) Z_{s,t}^{(3,3)} + \partial_{3,3}^2 h(z_s) Z_{s,t}^{(3,3,3)}.$$

Note that in the above example, we did not use the shuffle product formula because $m = 1$ ($Y_t \in \mathbb{R}$). If the response $Y$ lives in more that one dimensions, then the shuffle product formula is used, for example, to say that

$$\tfrac{1}{2}(z_t - z_s)^{(i_2)}(z_t - z_s)^{(i_1)} = \tfrac{1}{2} Z_{s,t}^{(i_2)} Z_{s,t}^{(i_1)} = Z_{s,t}^{(i_1,i_2)} + Z_{s,t}^{(i_2,i_1)}.$$

Below we give a concrete example to show how the shuffle product formula extends integration by parts.

EXAMPLE 2. Let us give here an example of the shuffle product. Let $z_t$ be a smooth path in $\mathbb{R}^m$ for some $m \geq 1$. Then, for any pair $i_1, i_2 \in \{1, \dots, m\}$ using the integration by parts formula, we get

$$\begin{aligned}
Z_{s,t}^{(i_1,i_2)} &= \int_s^t \int_s^u dz_{u_2}^{(i_1)}\, dz_u^{(i_2)} = \int_s^t \big(z_u^{(i_1)} - z_s^{(i_1)}\big)\, dz_u^{(i_2)} \\
&= \int_s^t z_u^{(i_1)}\, dz_u^{(i_2)} - z_s^{(i_1)}\big(z_t^{(i_2)} - z_s^{(i_2)}\big) \\
&= \big[z_u^{(i_1)} z_u^{(i_2)}\big]_s^t - \int_s^t z_u^{(i_2)}\, dz_u^{(i_1)} - z_s^{(i_1)}\big(z_t^{(i_2)} - z_s^{(i_2)}\big) \\
&= z_t^{(i_2)}\big(z_t^{(i_1)} - z_s^{(i_1)}\big) - \int_s^t z_u^{(i_2)}\, dz_u^{(i_1)} \\
&= \big(z_t^{(i_2)} - z_s^{(i_2)}\big)\big(z_t^{(i_1)} - z_s^{(i_1)}\big) - \int_s^t \big(z_u^{(i_2)} - z_s^{(i_2)}\big)\, dz_u^{(i_1)} \\
&= Z_{s,t}^{(i_1)} Z_{s,t}^{(i_2)} - Z_{s,t}^{(i_2,i_1)},
\end{aligned}$$

which is in agreement with the shuffle product formula, since the shuffle product of two letters is $(i_1) \sqcup (i_2) = \{(i_1, i_2), (i_2, i_1)\}$.

It is now clear that in order to extend this construction to any path $Z$ of finite $p$-variation, where $p \geq 2$, we will first need to define their iterated integrals $\mathbf{Z}_{s,t}^k$. These are not necessarily unique (e.g., if $Z$ is Brownian motion, then Itô and Stratonovich gave two different definitions for the integral). Then, we will need to find those integrals that respect the “shuffle product property.” Before going any further, we need to give some definitions.

DEFINITION 2.1. Let $\Delta_T := \{(s,t); 0 \leq s \leq t \leq T\}$. Let $p \geq 1$ be a real number. We denote by $T^{(k)}(\mathbb{R}^{\ell_1})$ the $k$th truncated tensor algebra

$$T^{(k)}(\mathbb{R}^{\ell_1}) := \mathbb{R} \oplus \mathbb{R}^{\ell_1} \oplus \mathbb{R}^{\ell_1 \otimes 2} \oplus \cdots \oplus \mathbb{R}^{\ell_1 \otimes k}.$$

(1) Let $\mathbf{Z}: \Delta_T \to T^{(k)}(\mathbb{R}^{\ell_1})$ be a continuous map. For each $(s,t) \in \Delta_T$, denote by $\mathbf{Z}_{s,t}$ the image of $(s,t)$ through $\mathbf{Z}$ and write

$$\mathbf{Z}_{s,t} = (\mathbf{Z}_{s,t}^0, \mathbf{Z}_{s,t}^1, \ldots, \mathbf{Z}_{s,t}^k) \in T^{(k)}(\mathbb{R}^{\ell_1}) \qquad \text{where } \mathbf{Z}_{s,t}^j = \{\mathbf{Z}_{s,t}^{(i_1,\ldots,i_j)}\}_{i_1,\ldots,i_j=1}^{\ell_1}.$$

The function $\mathbf{Z}$ is called a *multiplicative functional* of degree $k$ in $\mathbb{R}^{\ell_1}$ if $\mathbf{Z}_{s,t}^0 = 1$ for all $(s,t) \in \Delta_T$ and

$$\mathbf{Z}_{s,u} \otimes \mathbf{Z}_{u,t} = \mathbf{Z}_{s,t} \qquad \forall s,u,t \text{ satisfying } 0 \leq s \leq u \leq t \leq T,$$

that is, for every $(i_1, \ldots, i_l) \in \{1, \ldots, \ell_1\}^l$ and $l = 1, \ldots, k$:

$$(\mathbf{Z}_{s,u} \otimes \mathbf{Z}_{u,t})^{(i_1,\ldots,i_l)} = \sum_{j=0}^{l} \mathbf{Z}_{s,u}^{(i_1,\ldots,i_j)} \mathbf{Z}_{u,t}^{(i_{j+1},\ldots,i_l)}.$$

This is called *Chen’s identity*.

(2) A *p-rough path* $\mathbf{Z}$ in $\mathbb{R}^{\ell_1}$ is a multiplicative functional of degree $\lfloor p \rfloor$ in $\mathbb{R}^{\ell_1}$ that has finite $p$-variation, that is, $\forall i = 1, \dots, \lfloor p \rfloor$ and $(s, t) \in \Delta_T$, it satisfies

$$\|\mathbf{X}^i_{s,t}\| \leq \frac{(M(t-s))^{i/p}}{\beta(i/p)!},$$

where $\|\cdot\|$ is the Euclidean norm in the appropriate dimension and $\beta$ a real number depending only on $p$ and $M$ is a fixed constant. The space of $p$-rough paths in $\mathbb{R}^{\ell_1}$ is denoted by $\Omega_p(\mathbb{R}^{\ell_1})$.

(3) A *geometric p-rough path* is a $p$-rough path that can be expressed as a limit of 1-rough paths in the $p$-variation distance $d_p$, defined as follows: for any $\mathbf{X}, \mathbf{Y}$ continuous functions from $\Delta_T$ to $T^{(\lfloor p \rfloor)}(\mathbb{R}^{\ell_1})$,

$$d_p(\mathbf{X}, \mathbf{Y}) = \max_{1 \leq i \leq \lfloor p \rfloor} \sup_{\mathcal{D} \subset [0,T]} \left( \sum_{\ell} \|\mathbf{X}^i_{t_{\ell-1},t_\ell} - \mathbf{Y}^i_{t_{\ell-1},t_\ell}\|^{p/i} \right)^{i/p},$$

where $\mathcal{D} = \{t_\ell\}_\ell$ goes through all possible partitions of $[0, T]$. The space of geometric $p$-rough paths in $\mathbb{R}^n$ is denoted by $G\Omega_p(\mathbb{R}^{\ell_1})$.

One of the main results of the theory of rough paths is the following, called the "extension theorem."

THEOREM 2.2 (Theorem 3.7, [19]). *Let $p \geq 1$ be a real number and $k \geq 1$ be an integer. Let $\mathbf{X} : \Delta_T \to T^{(k)}(\mathbb{R}^n)$ be a multiplicative functional with finite p-variation. Assume that $k \geq \lfloor p \rfloor$. Then there exists a unique extension of $\mathbf{X}$ to a multiplicative functional $\hat{\mathbf{X}} : \Delta_T \to T^{(k+1)}(\mathbb{R}^n)$.*

Let $X : [0, T] \to \mathbb{R}^n$ be an $n$-dimensional path of finite $p$-variation for $n > 1$. One way of constructing a $p$-rough path is by considering the set of all iterated integrals of degree up to $\lfloor p \rfloor$. If $X_t = (X^{(1)}_t, \dots, X^{(n)}_t)$, we define $\mathbf{X} : \Delta_T \to T^{(\lfloor p \rfloor)}$ as follows:

$$\mathbf{X}^0 \equiv 1 \in \mathbb{R}$$

and

$$\mathbf{X}^k_{s,t} = \left\{ \int \cdots \int_{s < u_1 < \cdots < u_k < t} dX^{(i_1)}_{u_1} \cdots dX^{(i_k)}_{u_k} \right\}_{(i_1, \dots, i_k) \in \{1, \dots, n\}^k} \in \mathbb{R}^{n \otimes k}$$

for $k = 1, \dots, \lfloor p \rfloor$. Note that Chen's identity is an identity all iterated integrals satisfy. For example, for word $(i_1, i_2)$ Chen's identity says that

$$(\mathbf{Z}_{s,t})^{(i_1,i_2)} = (\mathbf{Z}_{s,u})^{(i_1,i_2)} + (\mathbf{Z}_{s,u})^{(i_1)} (\mathbf{Z}_{u,t})^{(i_2)} + (\mathbf{Z}_{u,t})^{(i_1,i_2)}.$$

This follows by breaking the domain of integration $\{u_1, u_2 : s < u_1 < u_2 < t\}$ into three domains $\{u_1, u_2 : s < u_1 < u_2 < u\}$, $\{u_1, u_2 : u < u_1 < u_2 < t\}$ and $\{u_1, u_2 : s < u_1 < u \text{ and } u < u_2 < t\}$.

When $p \in [1, 2)$, the iterated integrals are uniquely defined as Young integrals. However, as we already mentioned, when $p \geq 2$ there will be more than one way of defining them. What the extension theorem says is that if the path has finite $p$-variation and we define the first $\lfloor p \rfloor$ iterated integrals, the rest will be uniquely defined. So, if the path is of bounded variation ($p = 1$) we only need to know its increments, while for an $n$-dimensional Brownian path, we need to define the second iterated integrals by specifying the rules on how to construct them. In general, we can think of a $p$-rough path as a path $X : [0, T] \to \mathbb{R}^n$ of finite $p$-variation, together with a set of rules on how to define the first $\lfloor p \rfloor$ iterated integrals. Once we know how to construct the first $\lfloor p \rfloor$, we know how to construct all of them.

DEFINITION 2.3. Let $X : [0, T] \to \mathbb{R}^n$ be a path. The set of all iterated integrals is called the *signature of the path* and is denoted by $S(X)$.

We can now proceed to define the integral (2.3) when $Z$ is a path of finite $p$-variation with $p \geq 2$. First, it is clear that in order for the integral to be uniquely defined, we should define the first $\lfloor p \rfloor$ iterated integrals, so we define the integral not with respect to $Z$ but a corresponding $p$-rough path $\mathbf{Z}$. To extend the previous construction, we also need that $\mathbf{Z}$ satisfies the "shuffle product property." It is not hard to see that geometric $p$-rough paths do satisfy this property since they are limits of paths of bounded variation and for paths of bounded variation the property follows from the usual integration by parts formula (see also [18]). So, we will define $\int h(\mathbf{Z})\, d\mathbf{Z}$, where $\mathbf{Z}$ is a geometric $p$-rough path in $\mathbb{R}^{\ell_1}$, that is, $\mathbf{Z} \in G\Omega_p(\mathbb{R}^{\ell_1})$.

By definition, there exists a sequence $\mathbf{Z}(r) \in \Omega_1(\mathbb{R}^{\ell_1})$ such that $d_p(\mathbf{Z}(r), \mathbf{Z}) \to 0$ as $r \to \infty$. Then, for each $r > 0$, we define $\tilde{\mathbf{Z}}(r) := \int h(\mathbf{Z}(r))\, d\mathbf{Z}(r)$. These are also a 1-rough paths in $\mathbb{R}^{\ell_2}$ and thus, their higher iterated integrals are uniquely defined. In addition, it is possible to show that the map $\int h : \Omega_1(\mathbb{R}^{\ell_1}) \to \Omega_1(\mathbb{R}^{\ell_2})$ sending $\mathbf{Z}(r)$ to $\tilde{\mathbf{Z}}(r)$ is continuous in the $p$-variation topology.

We define $\tilde{\mathbf{Z}} := \int h(\mathbf{Z})\, d\mathbf{Z}$ as the limit of the $\tilde{\mathbf{Z}}(r)$ with respect to $d_p$—this is will also be a geometric $p$-rough path. In other words, the continuous map $\int h$ can be extended to a continuous map from $G\Omega_p(\mathbb{R}^{\ell_1})$ to $G\Omega_p(\mathbb{R}^{\ell_2})$, which are the closures of $\Omega_1(\mathbb{R}^{\ell_1})$ and $\Omega_1(\mathbb{R}^{\ell_2})$, respectively (see Theorem 4.12, [19]).

Note that this construction of the integral can be extended for any $h \in \mathrm{Lip}(\gamma - 1)$ for $\gamma > p$ (see [19]).

REMARK 2.4. We say that a sequence $\mathbf{Z}(r)$ of $p$-rough paths converges to a $p$-rough path $\mathbf{Z}$ in $p$-variation topology if there exists an $M \in \mathbb{R}$ and a sequence $a(r)$ converging to zero when $r \to \infty$, such that

$$\|\mathbf{Z}(r)^i_{s,t}\|, \|\mathbf{Z}^i_{s,t}\| \leq \big(M(t-s)\big)^{i/p}$$

and

$$\|\mathbf{Z}(r)^i_{s,t} - \mathbf{Z}^i_{s,t}\| \leq a(r)\big(M(t-s)\big)^{i/p}$$

for $i = 1, \dots, \lfloor p \rfloor$ and $(s,t) \in \Delta_T$. Note that this is not exactly equivalent to convergence in $d_p$: while convergence in $d_p$ implies convergence in the $p$-variation topology, the opposite is not true. Convergence in the $p$-variation topology implies that there is a *subsequence* that converges in $d_p$.

We can now give the precise meaning of the solution of (2.1), when driven not by a path $X$ but a geometric $p$-rough path $\mathbf{X}$:

DEFINITION 2.5. Consider $\mathbf{X} \in G\Omega_p(\mathbb{R}^n)$ and $y_0 \in \mathbb{R}^m$. Set $f_{y_0}(\cdot) := f(\cdot + y_0)$ and define $h : \mathbb{R}^n \oplus \mathbb{R}^m \to \mathrm{End}(\mathbb{R}^n \oplus \mathbb{R}^m)$ as in (2.2). We call $\mathbf{Z} \in G\Omega_p(\mathbb{R}^n \oplus \mathbb{R}^m)$ a *solution* of (2.1) if the following two conditions hold:

(i) $\mathbf{Z} = \int h(\mathbf{Z})\, d\mathbf{Z}$.
(ii) $\pi_{\mathbb{R}^n}(\mathbf{Z}) = \mathbf{X}$, where by $\pi_{\mathbb{R}^n}$ we denote the projection of $\mathbf{Z}$ to $\mathbb{R}^n$.

As in the case of ordinary differential equations ($p = 1$), it is possible to construct the solution using Picard iterations: we define $\mathbf{Z}(\mathbf{0}) := (\mathbf{X}, \mathbf{e})$, where by $\mathbf{e}$ we denote the trivial rough path $\mathbf{e} = (1, \mathbf{0}_{\mathbb{R}^n}, \mathbf{0}_{\mathbb{R}^{n \otimes 2}}, \dots)$. Then, for every $r \geq 1$, we define $\mathbf{Z}(r) = \int h(\mathbf{Z}(r-1))\, d\mathbf{Z}(r-1)$. The following theorem, called the "Universal Limit theorem," gives the conditions for the existence and uniqueness of the solution to (2.1). The theorem holds for any $f \in \mathrm{Lip}(\gamma)$ for $\gamma > p$ but we will assume that $f$ is a polynomial. The proof is based on the convergence of the Picard iterations.

THEOREM 2.6 (Theorem 5.3, [19]). *Let $p \geq 1$. For all $\mathbf{X} \in G\Omega_p(\mathbb{R}^n)$ and all $y_0 \in \mathbb{R}^m$, equation (2.1) admits a unique solution $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \in G\Omega_p(\mathbb{R}^n \oplus \mathbb{R}^m)$, in the sense of Definition 2.5. This solution depends continuously on $\mathbf{X}$ and $y_0$ and the mapping $I_f : G\Omega_p(\mathbb{R}^n) \to G\Omega_p(\mathbb{R}^m)$ which sends $(\mathbf{X}, y_0)$ to $\mathbf{Y}$ is continuous in the $p$-variation topology.*

*The rough path $\mathbf{Y}$ is the limit of the sequence $\mathbf{Y}(r)$, where $\mathbf{Y}(r)$ is the projection of the $r$th Picard iteration $\mathbf{Z}(r)$ to $\mathbb{R}^m$. For all $\rho > 1$, there exists $T_\rho \in (0, T]$ such that*

$$\|\mathbf{Y}(r)^i_{s,t} - \mathbf{Y}(r+1)^i_{s,t}\| \leq 2^i \rho^{-r} \frac{(M(t-s))^{i/p}}{\beta(i/p)!} \qquad \forall (s,t) \in \Delta_{T_\rho}, \forall i = 0, \dots, \lfloor p \rfloor.$$

*The constant $T_\rho$ depends only on $f$ and $p$.*

2.2. *The problem.* We now describe the problem that we are going to study in the rest of the paper. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbf{X} : \Omega \to G\Omega_p(\mathbb{R}^n)$ a random variable, taking values in the space of geometric $p$-rough paths endowed with the $p$-variation topology. For each $\omega \in \Omega$, the rough path $\mathbf{X}(\omega)$ drives the following differential equation

$$dY_t(\omega) = f(Y_t(\omega);\theta) \cdot dX_t(\omega), \qquad Y_0 = y_0, \tag{2.5}$$

where $\theta \in \Theta \subseteq \mathbb{R}^d$, $\Theta$ being the parameter space and for each $\theta \in \Theta$. As before, $f : \mathbb{R}^m \times \Theta \to \mathrm{L}(\mathbb{R}^n, \mathbb{R}^m)$ and $f_\theta(y) := f(y;\theta)$ is a polynomial in $y$ for each $\theta \in \Theta$. According to Theorem 2.6, we can think of equation (2.5) as a map

$$I_{f_\theta, y_0} : G\Omega_p(\mathbb{R}^n) \to G\Omega_p(\mathbb{R}^m), \tag{2.6}$$

sending a geometric $p$-rough path $\mathbf{X}$ to a geometric $p$-rough path $\mathbf{Y}$ and is continuous with respect to the $p$-variation topology. Consequently,

$$\mathbf{Y} := I_{f_\theta, y_0} \circ \mathbf{X} : \Omega \to G\Omega_p(\mathbb{R}^m)$$

is also a random variable, taking values in $G\Omega_p(\mathbb{R}^m)$ and if $\mathbb{P}^T$ is the distribution of $\mathbf{X}_{0,T}$, the distribution of $\mathbf{Y}_{0,T}$ will be

$$\mathbb{Q}_\theta^T = \mathbb{P}^T \circ I_{f_\theta, y_0}^{-1}. \tag{2.7}$$

Suppose that we know the *expected signature* of $\mathbf{X}$ at $[0, T]$, that is, we know

$$\mathbb{E}(\mathbf{X}_{0,T}^{(i_1,\dots,i_k)}) := \mathbb{E}\Big(\int \cdots \int_{0<u_1<\cdots<u_k<T} dX_{u_1}^{(i_1)} \cdots dX_{u_k}^{(i_k)}\Big)$$

for all $i_j \in \{1, \dots, n\}$ where $j = 1, \dots, k$ and $k \geq 1$. Our goal will be to estimate $\theta$, given several realizations of $\mathbf{Y}_{0,T}$, that is, $\{\mathbf{Y}_{0,T}(\omega_i)\}_{i=1}^N$.

REMARK 2.7. We are assuming that we are observing many independent copies of the signature at just one point $T$. In the case of scalar response, this is equivalent to observing many independent realizations of the response at just one point in time. In the case the response lives in more than one dimensions, the elements of the signature might depend on either the whole path up to time $T$ or just on $T$, making the method appropriate for both cases of discrete or continuous observations.

**3. Method.** In order to estimate $\theta$, we are going to use a method that is similar to the "Method of Moments." The idea is simple: we will try to (partially) match the empirical expected signature of the observed $p$-rough path with the theoretical one, which is a function of the unknown parameters. Remember that the data we have available is several realizations of the $p$-rough path $\mathbf{Y}_{0,T}$ described in Section 2.2. To make this more precise, let us introduce some notation: let

$$E^\tau(\theta) := \mathbb{E}_\theta(\mathbf{Y}_{0,T}^\tau) \tag{3.1}$$

be the *theoretical expected signature* corresponding to parameter value $\theta$ and word $\tau$ and

$$M_N^\tau := \frac{1}{N} \sum_{i=1}^N \mathbf{Y}_{0,T}^\tau(\omega_i) \tag{3.2}$$

be the *empirical expected signature*, which is a Monte Carlo approximation of the actual one. The word $\tau$ is constructed from the alphabet $\{1, \dots, m\}$, that is, $\tau \in W_m$ where $W_m := \bigcup_{k \geq 0} \{1, \dots, m\}^k$. The idea is to find $\hat{\theta}$ such that

$$E^{\tau}(\hat{\theta}) = M_N^{\tau} \qquad \forall \tau \in V \subset W_m$$

for some choice of a set of words $V$. Then $\hat{\theta}$ will be our estimate.

REMARK 3.1. When $m = 1$, the expected signature of $\mathbf{Y}$ is equivalent to its moments, since

$$\mathbf{Y}_{0,T}^{\overbrace{(1, \dots, 1)}^{m}} = \frac{1}{m!}(Y_T - Y_0)^m.$$

When $m = 2$, one example is to consider the word $\tau = (1, 2)$. Then, one needs to compute the iterated integral (or an approximation of, if the path is discretely observed)

$$\mathbf{Y}_{0,T}^{(1,2)}(\omega_i) = \int_0^T \int_0^s dY_u^{(1)}(\omega_i)\, dY_s^{(2)}(\omega_i)$$

for each path $Y_t(\omega_i) = (Y_t^{(1)}(\omega_i), Y_t^{(1)}(\omega_i))$, for $i = 1, \dots, N$. Then

$$M_N^{(1,2)} := \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_{0,T}^{(1,2)}(\omega_i).$$

Note that this is closely related to the correlation of the two one-dimensional paths $\{Y_t^{(1)}\}_{t \in [0,T]}$ and $\{Y_t^{(2)}\}_{t \in [0,T]}$ since, by the shuffle product,

$$\mathbf{Y}_{0,T}^{(1,2)}(\omega_i) + \mathbf{Y}_{0,T}^{(2,1)}(\omega_i) = \mathbf{Y}_{0,T}^{(1)}(\omega_i)\mathbf{Y}_{0,T}^{(2)}(\omega_i)$$

and by the law of large numbers,

$$\lim_{N \to \infty} \bigl(M_N^{(1,2)} + M_N^{(2,1)}\bigr) = \mathbb{E}\bigl(\bigl(Y_T^{(1)} - Y_0^{(1)}\bigr)\bigl(Y_T^{(2)} - Y_0^{(2)}\bigr)\bigr).$$

Several questions arise:

(i) How can we get an analytic expression for $E^{\tau}(\theta)$ as a function of $\theta$?
(ii) What is a good choice for $V$ or, for $m = 1$, how do we choose which moments to match?
(iii) How good is $\hat{\theta}$ as an estimate?

We will try to answer these questions below.

3.1. *Computing the theoretical expected signature.* We want to get an analytic expression for the expected signature of the $p$-rough path $\mathbf{Y}$ at $(0,T)$, where $\mathbf{Y}$ is the solution of (2.5) in the sense described above. In other words, we want to compute (3.1). We are given the expected signature of the $p$-rough path $\mathbf{X}$ which is driving the equation, again at $(0,T)$, that is, we are given

$$\mathbb{E}(\mathbf{X}_{0,T}^{\sigma}) \qquad \forall\sigma \in \{1,\dots,n\}^k, \qquad k \in \mathbb{N}.$$

In addition, we know the vector field $f_\theta(y) = f(y;\theta)$ in (2.5), up to parameter $\theta$ and we know that it is polynomial.

It turns out that we cannot compute (3.1), in general. We need to make one more approximation since the solution $\mathbf{Y}$ will not usually be available: we will approximate the solution by the $r$th Picard iteration $\mathbf{Y}(r)$, described in the Universal Limit theorem (Theorem 2.6). Finally, we will approximate the expected signature of the solution corresponding to a word $\tau$, $E^\tau(\theta)$, by the expected signature of the $r$th Picard iteration at $\tau$, which we will denote by $E_r^\tau(\theta)$:

$$E_r^\tau(\theta) := \mathbb{E}_\theta(\mathbf{Y}(r)_{0,T}^{\tau}). \tag{3.3}$$

The good news is that when $f_\theta$ is a polynomial of degree $q$ on $y$, for any $q \in \mathbb{N}$, the $r$th Picard iteration of the solution is a linear combination of iterated integrals of the driving force $\mathbf{X}$. More specifically, for any realization $\omega$ and any time interval $(s,t) \in \Delta_T$, we can write

$$\mathbf{Y}(r)_{s,t}^{\tau} = \sum_{|\sigma|\le|\tau|(q^r-1)/(q-1)} \alpha_{r,\sigma}^{\tau}(y_0,s;\theta)\mathbf{X}_{s,t}^{\sigma}, \tag{3.4}$$

where $\alpha_{r,\sigma}^{\tau}(y;\theta)$ is a polynomial in $y$ of degree $q^r$ and $|\cdot|$ gives the length of a word. Thus,

$$E_r^\tau(\theta) = \sum_{|\sigma|\le|\tau|(q^r-1)/(q-1)} \alpha_{r,\sigma}^{\tau}(y_0,s;\theta)\mathbb{E}(\mathbf{X}_{s,t}^{\sigma}). \tag{3.5}$$

We will prove (3.4), first for $p = 1$ and then for any $p \ge 1$ by taking limits with respect to $d_p$. We will need the following lemma.

LEMMA 3.2. *Suppose that* $\mathbf{X} \in G\Omega_1(\mathbb{R}^n)$, $\mathbf{Y} \in G\Omega_1(\mathbb{R}^m)$ *and it is possible to write*

$$\mathbf{Y}_{s,t}^{(j)} = \sum_{\sigma\in W_n, q_1\le|\sigma|\le q_2} \alpha_\sigma^{(j)}(y_s)\mathbf{X}_{s,t}^{\sigma} \qquad \forall(s,t) \in \Delta_T \text{ and } \forall j = 1,\dots,m, \tag{3.6}$$

*where* $\alpha_\sigma^{(j)} : \mathbb{R}^m \to \mathrm{L}(\mathbb{R},\mathbb{R})$ *is a polynomial of degree* $q$ *with* $q, q_1, q_2 \in \mathbb{N}$ *and* $q_1 \ge 1$. *Then*

$$\mathbf{Y}_{s,t}^{\tau} = \sum_{\sigma\in W_n, |\tau|q_1\le|\sigma|\le|\tau|q_2} \alpha_\sigma^{\tau}(y_s)\mathbf{X}_{s,t}^{\sigma} \tag{3.7}$$

*for all* $(s,t) \in \Delta_T$ *and* $\tau \in W_m$. $\alpha_\sigma^\tau : \mathbb{R}^m \to \mathrm{L}(\mathbb{R},\mathbb{R})$ *are polynomials of degree* $\le q|\tau|$.

PROOF. We will prove (3.7) by induction on $|\tau|$, that is, the length of the word. By hypothesis, it is true when $|\tau| = 1$. Suppose that it is true for any $\tau \in W_m$ such that $|\tau| = k \geq 1$. First, note that from (3.6), we get that

$$dY_u^{(j)} = \sum_{\sigma \in W_n, q_1 \leq |\sigma| \leq q_2} \alpha_\sigma^{(j)}(y_s) \mathbf{X}_{s,u}^{\sigma-} \, dX_u^{\sigma_\ell} \qquad \forall u \in [s, t],$$

where $\sigma-$ is the word $\sigma$ without the last letter and $\sigma_\ell$ is the last letter. For example, if $\sigma = (i_1, \ldots, i_{b-1}, i_b)$, then $\sigma- = (i_1, \ldots, i_{b-1})$ and $\sigma_\ell = i_b$. Note that this cannot be defined when $\sigma$ is the empty word $\varnothing$ ($b = 0$). Now suppose that $|\tau| = k+1$, so $\tau = (j_1, \ldots, j_k, j_{k+1})$ for some $j_1, \ldots, j_{k+1} \in \{1, \ldots, m\}$. Then

$$\begin{aligned}
\mathbf{Y}_{s,t}^\tau &= \int_s^t \mathbf{Y}_{s,u}^{\tau-} \, dY_u^{(j_{k+1})} \\
&= \int_s^t \biggl( \sum_{kq_1 \leq |\sigma_1| \leq kq_2} \alpha_{\sigma_1}^{\tau-}(y_s) \mathbf{X}_{s,u}^{\sigma_1} \biggr) \sum_{q_1 \leq |\sigma_2| \leq q_2} \alpha_{\sigma_2}^{(j_{k+1})}(y_s) \mathbf{X}_{s,u}^{\sigma_2-} \, dX_u^{\sigma_{2\ell}} \\
&= \sum_{kq_1 \leq |\sigma_1| \leq kq_2, q_1 \leq |\sigma_2| \leq q_2} \bigl( \alpha_{\sigma_1}^{\tau-}(y_s) \alpha_{\sigma_2}^{(j_{k+1})}(y_s) \bigr) \int_s^t \mathbf{X}_{s,u}^{\sigma_1} \mathbf{X}_{s,u}^{\sigma_2-} \, dX_u^{\sigma_{2\ell}}.
\end{aligned}$$

Now we use the fact that for any geometric rough path $\mathbf{X}$ and any $(s, u) \in \Delta_T$, we can write

$$\mathbf{X}_{s,u}^{\sigma_1} \mathbf{X}_{s,u}^{\sigma_2-} = \sum_{\sigma \in \sigma_1 \sqcup (\sigma_2-)} \mathbf{X}_{s,u}^{\sigma}, \tag{3.8}$$

where $\sigma_1 \sqcup (\sigma_2-)$ is the shuffle product between the words $\sigma_1$ and $\sigma_2-$. Applying (3.8) above, we get

$$\mathbf{Y}_{s,t}^\tau = \sum_{\sigma \in W_n, (k+1)q_1 \leq |\sigma| \leq (k+1)q_2} \alpha_\sigma^\tau(y_s) \mathbf{X}_{s,t}^\sigma,$$

where

$$\alpha_\sigma^\tau(y_s) = \sum_{(\sigma_1 \sqcup \sigma_2-) \ni \sigma-, \sigma_\ell = \sigma_{2\ell}} \alpha_{\sigma_1}^{\tau-}(y_s) \alpha_{\sigma_2}^{\tau_\ell}(y_s)$$

is a polynomial of degree $\leq kq + q = (k+1)q$. Note that the above sum is over all $\sigma_1, \sigma_2 \in W_n$ such that $kq_1 \leq |\sigma_1| \leq kq_2$ and $q_1 \leq |\sigma_1| \leq q_2$. □

We now prove (3.4) for $p = 1$.

LEMMA 3.3. *Suppose that* $\mathbf{X} \in G\Omega_1(\mathbb{R}^n)$ *is driving system* (2.1), *where* $f : \mathbb{R}^m \to \mathrm{L}(\mathbb{R}^n, \mathbb{R}^m)$ *is a polynomial of degree* $q$. *Let* $\mathbf{Y}(r)$ *be the projection of the* $r$*th Picard iteration* $\mathbf{Z}(r)$ *to* $\mathbb{R}^m$, *as described above. Then,* $\mathbf{Y}(r) \in G\Omega_1(\mathbb{R}^m)$ *and it satisfies*

$$\mathbf{Y}(r)_{s,t}^\tau = \sum_{|\sigma| \leq |\tau|(q^r - 1)/(q-1)} \alpha_{r,\sigma}^\tau(y_0, s) \mathbf{X}_{s,t}^\sigma \tag{3.9}$$

*for all* $(s, t) \in \Delta_T$ *and* $\tau \in W_m$. $\alpha_{r,\sigma}^\tau(y, s)$ *is a polynomial of degree* $\leq |\tau| q^r$ *in* $y$.

PROOF. For every $r \geq 0$, $\mathbf{Z}(r) \in G\Omega_1(\mathbb{R}^{n+m})$ since $\mathbf{Z}(\mathbf{0}) := (\mathbf{X}, \mathbf{e})$, $\mathbf{X} \in G\Omega_1(\mathbb{R}^n)$, and integrals preserve the roughness of the integrator. So, $\mathbf{Y}(r) \in G\Omega_1(\mathbb{R}^m)$. We will prove the claim by induction on $r$.

For $r = 0$, $\mathbf{Y}(\mathbf{0}) = \mathbf{e}$ and thus (3.9) becomes

$$\mathbf{Y}(\mathbf{0})^{\tau}_{s,t} = \alpha^{\tau}_{0,\varnothing}(y_0, s)$$

and it is true for $\alpha^{\varnothing}_{0,\varnothing} \equiv 1$ and $\alpha^{\tau}_{0,\varnothing} \equiv 0$ for every $\tau \in W_m$ such that $|\tau| > 0$.

Now suppose it is true for some $r \geq 0$. Remember that $\mathbf{Z}(r) = (\mathbf{X}, \mathbf{Y}(r))$ and that $\mathbf{Z}(r+1)$ is defined by

$$\mathbf{Z}(r+1) = \int h(\mathbf{Z}(r))\, d\mathbf{Z}(r),$$

where $h$ is defined in (2.2) and $f_{y_0}(y) = f(y_0 + y)$. Since $f$ is a polynomial of degree $q$, $h$ is also a polynomial of degree $q$ and, thus, it is possible to write

$$h(z_2) = \sum_{k=0}^{q} h_k(z_1) \frac{(z_2 - z_1)^{\otimes k}}{k!} \qquad \forall z_1, z_2 \in \mathbb{R}^{\ell}, \tag{3.10}$$

where $\ell = n + m$. Then, the integral is defined to be

$$\mathbf{Z}(r+1)_{s,t} := \int_s^t h(\mathbf{Z}(r))\, d\mathbf{Z}(r) = \sum_{k=0}^{q} h_k(Z(r)_s) \mathbf{Z}(r)^{k+1}_{s,t} \qquad \forall (s,t) \in \Delta_T.$$

Let's take a closer look at functions $h_k : \mathbb{R}^{\ell} \to \mathrm{L}(\mathbb{R}^{\ell \otimes k}, \mathrm{L}(\mathbb{R}^{\ell}, \mathbb{R}^{\ell}))$. Since (3.10) is the Taylor expansion for polynomial $h$, $h_k$ is the $k$th derivative of $h$. So, for every word $\beta \in W_{\ell}$ such that $|\beta| = k$ and every $z = (x, y) \in \mathbb{R}^{\ell}$, $(h_k(z))^{\beta} = \partial_{\beta} h(z) \in \mathrm{L}(\mathbb{R}^{\ell}, \mathbb{R}^{\ell})$. By definition, $h$ is independent of $x$ and thus the derivative will always be zero if $\beta$ contains any letters in $\{1, \dots, n\}$.

Remember that $\mathbf{Y}(r+1)$ is the projection of $\mathbf{Z}(r+1)$ onto $\mathbb{R}^m$. So, for each $j \in \{1, \dots, m\}$,

$$\begin{aligned}
\mathbf{Y}(r+1)^{(j)}_{s,t} = \mathbf{Z}(r+1)^{(n+j)}_{s,t} &= \sum_{k=0}^{q} (h_k(Z(r)_s) \mathbf{Z}(r)^{k+1}_{s,t})^{(n+j)} \\
&= \sum_{i=1}^{\ell} \sum_{\tau \in W_m(0,q)} \partial_{\tau+n} h_{n+j,i}(Z(r)_s) \mathbf{Z}(r)^{(\tau+n,i)}_{s,t} \qquad (3.11) \\
&= \sum_{i=1}^{n} \sum_{\tau \in W_m(0,q)} \partial_{\tau} f_{j,i}\big(y_0 + Y(r)_s\big) \mathbf{Y}(r)^{(\tau,i)}_{s,t},
\end{aligned}$$

where $W_m(k_1, k_2) = \{\tau \in W_m; k_1 \leq |\tau| \leq k_2\}$ for any $k_1, k_2 \in \mathbb{N}$, that is, it is the set of all words of length between $k_1$ and $k_2$. By the induction hypothesis, we know that for every $\tau \in W_m$,

$$\mathbf{Z}(r)^{\tau+n}_{s,t} = \mathbf{Y}(r)^{\tau}_{s,t} = \sum_{|\sigma| \leq |\tau|(q^r - 1)/(q-1)} \alpha^{\tau}_{r,\sigma}(y_0, s) \mathbf{X}^{\sigma}_{s,t}$$

and thus, for every $i = 1, \dots, n$,

$$\mathbf{Z}(r)_{s,t}^{(\tau+n,i)} = \sum_{|\sigma| \le |\tau|(q^r-1)/(q-1)} \alpha_{r,\sigma}^{\tau}(y_0, s)\mathbf{X}_{s,t}^{(\sigma,i)}. \tag{3.12}$$

Putting this back to the equation above, we get

$$\mathbf{Y}(r+1)_{s,t}^{(j)} = \sum_{i=1}^{n} \sum_{|\tau| \le q} \partial_\tau f_{j,i}\big(y_0 + Y(r)_s\big) \sum_{|\sigma| \le |\tau|(q^r-1)/(q-1)} \alpha_{r,\sigma}^{\tau}(y_0, s)\mathbf{X}_{s,t}^{(\sigma,i)}$$

and by reorganizing the sums, we get

$$\mathbf{Y}(r+1)_{s,t}^{(j)} = \sum_{|\sigma| \le q(q^r-1)/(q-1)+1 = (q^{r+1}-1)/(q-1)} \alpha_{r+1,\sigma}^{(j)}(y_0, s)\mathbf{X}_{s,t}^{\sigma}, \tag{3.13}$$

where $\alpha_{r+1,\varnothing}^{(j)} \equiv 0$ and for every $\sigma \in W_n - \varnothing$,

$$\alpha_{r+1,\sigma}^{(j)}(y_0, s) = \sum_{|\sigma-|(q-1)/(q^r-1) \le |\tau| \le q} \partial_\tau f_{j,\sigma_\ell}\big(y_0 + Y(r)_s\big)\alpha_{r,\sigma-}^{\tau}(y_0, s).$$

If $\alpha_{r,\sigma}^{\tau}$ are polynomials of degree $\le |\tau| q^r$, then $\alpha_{r,\sigma}^{(j)}$ are polynomials of degree $\le q^r$. The result follow by applying Lemma 3.2. Notice that (in the notation of Lemma 3.2) $q_1 \ge 1$ since $\alpha_{r+1,\varnothing}^{(j)} \equiv 0$. □

We will now prove (3.4) for any $p \ge 1$.

THEOREM 3.4. *The result of Lemma 3.3 still holds when $\mathbf{X} \in G\Omega_p(\mathbb{R}^n)$, for any $p \ge 1$.*

PROOF. Since $\mathbf{X} \in G\Omega_p(\mathbb{R}^n)$, there exists a sequence $\{\mathbf{X}(k)\}_{k \ge 0}$ in $G\Omega_1(\mathbb{R}^n)$, such that $\mathbf{X}(k) \overset{k\to\infty}{\to} \mathbf{X}$ in the $p$-variation topology. We denote by $\mathbf{Z}(k, r)$ and $\mathbf{Z}(r)$ the $r$th Picard iteration corresponding to equation (2.1) driven by $\mathbf{X}(k)$ and $\mathbf{X}$, respectively.

First, we show that $\mathbf{Z}(k, r) \overset{k\to\infty}{\to} \mathbf{Z}(r)$ and consequently $\mathbf{Y}(k, r) \overset{k\to\infty}{\to} \mathbf{Y}(r)$ in the $p$-variation topology, for every $r \ge 0$. It is clearly true for $r = 0$. Now suppose that it is true for some $r \ge 0$. By definition, $\mathbf{Z}(r+1) = \int h(\mathbf{Z}(r))\, d\mathbf{Z}(r)$. Remember that the integral is defined as the limit in the $p$-variation topology of the integrals corresponding to a sequence of 1-rough paths that converge to $\mathbf{Z}(r)$ in the $p$-variation topology. By the induction hypothesis, this sequence can be $\mathbf{Z}(k, r)$. It follows that $\mathbf{Z}(k, r+1) = \int h(\mathbf{Z}(k, r))\, d\mathbf{Z}(k, r)$ converges to $\mathbf{Z}(r+1)$, which proves the claim. Convergence of the rough paths in $p$-variation topology implies convergence of each of the iterated integrals, that is,

$$\mathbf{Y}(k, r)_{s,t}^{\tau} \overset{k\to\infty}{\to} \mathbf{Y}(r)_{s,t}^{\tau}$$

for all $r \geq 0$, $(s, t) \in \Delta_T$ and $\tau \in W_m$.

By Lemma 3.3, since $\mathbf{X}(k) \in G\Omega_1(\mathbb{R}^n)$ for every $k \geq 1$, we can write

$$\mathbf{Y}(k, r)^{\tau}_{s,t} = \sum_{|\sigma| \leq |\tau|(q^r-1)/(q-1)} \alpha^{\tau}_{r,\sigma}(y_0, s)\mathbf{X}(k)^{\sigma}_{s,t}$$

for every $\tau \in W_m$, $(s, t) \in \Delta_T$ and $k \geq 1$. Since $\mathbf{X}(k) \stackrel{k\to\infty}{\to} \mathbf{X}$ in the $p$-variation topology and the sum is finite, it follows that

$$\mathbf{Y}(k, r)^{\tau}_{s,t} \stackrel{k\to\infty}{\to} \sum_{|\sigma| \leq |\tau|(q^r-1)/(q-1)} \alpha^{\tau}_{r,\sigma}(y_0, s)\mathbf{X}^{\sigma}_{s,t}.$$

The statement of the theorem follows. □

3.2. *The expected signature matching estimator.* We can now give a precise definition of the estimator, which we will formally call the *expected signature matching estimator* (ESME): suppose that we are in the setting of the problem described in Section 2.2 and $M^{\tau}_N$ and $E^{\tau}_r(\theta)$ are defined as in (3.2) and (3.3), respectively, for every $\tau \in W_m$. Let $V \subset W_m$ be a set of $d$ words constructed from the alphabet $\{1, \dots, m\}$. For each such $V$, we define the ESME $\hat{\theta}^V_{r,N}$ as the solution to

$$E^{\tau}_r(\theta) = M^{\tau}_N \qquad \forall \tau \in V. \tag{3.14}$$

This definition requires that (3.14) has a *unique* solution. This will not be true in general. Let $\mathcal{V}_r$ be the set of all $V$ containing $d$ words, such that $E^{\tau}_r(\theta) = M$, $\forall \tau \in V$, has a unique solution for all $M \in S_\tau \subseteq \mathbb{R}$ where $S_\tau$ is the set of all possible values of $M^{\tau}_N$, for any $N \geq 1$. We will assume the following.

ASSUMPTION 1 (Observability). The set $\mathcal{V}_r$ is nonempty and known (at least up to a nonempty subset).

Then $\hat{\theta}^V_{r,N}$ can be defined for every $V \in \mathcal{V}_r$.

REMARK 3.5. In order to achieve uniqueness of the estimator, we might need some extra information that we could get by looking at time correlations. We can fit this into our framework by considering scaled versions of (2.5) together with the original one: for example, consider the equation

$$dY_t(\omega) = f(Y_t(\omega); \theta) \cdot dX_t(\omega), \qquad Y_0 = y_0,$$
$$dY(c)_t(\omega) = f(Y(c)_t(\omega); \theta) \cdot dX_{ct}(\omega), \qquad Y(c)_0 = y_0,$$

for some appropriate constant $c$. Then $Y(c)_t = Y_{ct}$ and the expected signature at $[0, T]$ will also contain information about $\mathbb{E}(Y^{(j_1)}_T Y^{(j_2)}_{cT})$ for any $j_1, j_2 = 1, \dots, m$.

It is very difficult to say anything about the solutions of system (3.14), as it is very general. However, under the assumption that $f$ is also a polynomial in $\theta$, (3.14) becomes a system of polynomial equations and the problem of identifiability becomes equivalent to the problem of existence and uniqueness of solutions for that system.

3.3. *Properties of the ESME.* It is possible to show that the ESME defined as the solution of (3.14) will converge to the true value of the parameter and will be asymptotically normal. More precisely, the following holds.

THEOREM 3.6. *Let $\hat{\theta}^V_{r,N}$ be the expected signature matching estimator for the system described in Section 2.2 and $V \in \mathcal{V}_r$. Assume that the expected signature of $\mathbf{Y}_{0,T}$ is finite and that $f(y;\theta)$ is a polynomial of degree $q$ with respect to $y$ and twice differentiable with respect to $\theta$. Let $\theta_0$ be the "true" parameter value, meaning that the distribution of the observed signature $\mathbf{Y}_{0,T}$ is $\mathbb{Q}^T_{\theta_0}$, defined in (2.7). Set*

$$D^V_r(\theta)_{i,\tau} = \frac{\partial}{\partial\theta_i} E^\tau_r(\theta) \quad \text{and} \quad \Sigma_V(\theta_0)_{\tau,\tau'} = \operatorname{cov}(\mathbf{Y}^\tau_{0,T}, \mathbf{Y}^{\tau'}_{0,T}) \tag{3.15}$$

*and assume that* $\inf_{r>0,\theta\in\Theta} \|D^V_r(\theta)\| > 0$, *that is, $D^V_r(\theta)$ is uniformly nondegenerate with respect to $r$ and $\theta$. Then, for $r \propto \log N$ and $T$ are sufficiently small,*

$$\hat{\theta}^V_{r,N} \to \theta_0 \qquad \textit{with probability } 1 \tag{3.16}$$

*and*

$$\sqrt{N}\Phi_V(\theta_0)^{-1}(\hat{\theta}^V_{r,N} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I) \tag{3.17}$$

*as $N \to \infty$, where*

$$\Phi_V(\theta_0) = D^V(\theta_0)^{-1}\Sigma_V(\theta_0)^{1/2} \tag{3.18}$$

*with $D^V(\theta)_{i,\tau} = \frac{\partial}{\partial\theta_i} E^\tau(\theta)$.*

PROOF. By Theorem 3.4 and the definition of $E^\tau_r(\theta)$,

$$E^\tau_r(\theta) = \sum_{|\sigma|\le|\tau|(q^r-1)/(q-1)} \alpha^\tau_{r,\sigma}(y_0;\theta)\mathbb{E}(\mathbf{X}^\sigma_{0,T}),$$

where functions $\alpha^\tau_{r,\sigma}(y_0;\theta)$ are constructed recursively, as in Lemmas 3.2 and 3.3. Since $f$ is twice differentiable with respect to $\theta$, functions $\alpha$ and consequently $E^\tau_r$ will also be twice differentiable with respect to $\theta$. Thus, we can write

$$E^\tau_r(\theta) - E^\tau_r(\theta_0) = D^V_r(\tilde{\theta})_{\cdot,\tau}(\theta - \theta_0) \qquad \forall\theta \in \Theta \subseteq \mathbb{R}^d$$

for some $\tilde{\theta}$ within a ball of center $\theta_0$ and radius $\|\theta - \theta_0\|$ and the function $D^V_r(\theta)$ is continuous. By inverting $D^V_r$ and for $\theta = \hat{\theta}^V_{r,N}$, we get

$$(\hat{\theta}^V_{r,N} - \theta_0) = D^V_r(\tilde{\theta}^V_{r,N})^{-1}\big(E^V_r(\hat{\theta}^V_{r,N}) - E^V_r(\theta_0)\big), \tag{3.19}$$

where $E_r^V(\theta) = \{E_r^\tau(\theta)\}_{\tau \in V}$. By definition

$$E_r^V(\hat{\theta}_{r,N}^V) = \{M_N^\tau\}_{\tau \in V} = \left\{\frac{1}{N}\sum_{i=1}^N \mathbf{Y}_{0,T}^\tau(\omega_i)\right\}_{\tau \in V}, \tag{3.20}$$

where $\mathbf{Y}_{0,T}(\omega_i)$ are independent realizations of the random variable $\mathbf{Y}_{0,T}$. Suppose that $T$ is small enough, so that the above Monte Carlo approximation satisfies both the Law of Large Numbers and the Central Limit theorem, that is, the covariance matrix satisfies $0 < \|\Sigma_V(\theta_0)\| < \infty$. Then, for $N \to \infty$

$$|E_r^\tau(\hat{\theta}_{r,N}^V) - E^\tau(\theta_0)| = |E_r^\tau(\hat{\theta}_{r,N}^V) - \mathbb{E}(\mathbf{Y}_{0,T}^\tau)| \to 0 \qquad \forall \tau \in V$$

with probability 1. Note that the convergence does not depend on $r$. Also, for $r \to \infty$

$$E_r^\tau(\theta_0) \to E^\tau(\theta_0)$$

as a result of Theorem 2.6. Thus, for $r \propto \log N$

$$|E_r^\tau(\hat{\theta}_{r,N}^V) - E_r^\tau(\theta_0)| \to 0 \qquad \text{with probability } 1, \forall \tau \in V.$$

Combining this with (3.19) and the uniform nondegeneracy of $D_r^V$, we get (3.16). From (3.16) and the continuity and uniform nondegeneracy of $D_r^V$, we conclude that

$$D^V(\theta_0) D_r^V(\tilde{\theta}_{r,N}^V)^{-1} \to I \qquad \text{with probability 1}$$

provided that $T$ is small enough, so that $E^V(\theta_0) < \infty$. Now, since

$$\Phi_V(\theta_0)^{-1}(\hat{\theta}_{r,N}^V - \theta_0) = \Sigma_V(\theta_0)^{-1/2}(D^V(\theta_0)D_r^V(\tilde{\theta}_{r,N}^V)^{-1})\big(E_r^V(\hat{\theta}_{r,N}^V) - E_r^V(\theta_0)\big)$$

to prove (3.17) it is sufficient to prove that

$$\sqrt{N}\Sigma_V(\theta_0)^{-1/2}\big(E_r^V(\hat{\theta}_{r,N}^V) - E_r^V(\theta_0)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I).$$

It follows directly from (3.20) that

$$\sqrt{N}\Sigma_V(\theta_0)^{-1/2}\big(E_r^V(\hat{\theta}_{r,N}^V) - E^V(\theta_0)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I).$$

It remains to show that

$$\sqrt{N}\Sigma_V(\theta_0)^{-1/2}\big(E_r^V(\theta_0) - E^V(\theta_0)\big) \to 0.$$

It follows from Theorem 2.6 that

$$\|E_r^V(\theta_0) - E^V(\theta_0)\| \le C\rho^{-r}$$

for any $\rho > 1$ and sufficiently small $T$. The constant $C$ depends on $V$, $p$ and $T$. Suppose that $r = a\log N$ for some $a > 0$ and choose $\rho > \exp(\frac{1}{2c})$. Then

$$\sqrt{N}\big\|\big(E_r^V(\theta_0) - E^V(\theta_0)\big)\big\| \le CN^{(1/2 - c\log\rho)},$$

which proves the claim. $\square$

**4. Extensions.** In this section, we discuss how to extend the method described in Section 3 in two different directions. First, we generalize the ESME by matching linear combinations of the elements of the signature and considering issues of optimality. Then, we extend the method to a different setting where we observe one path of the signature of the response at many points in time rather than many independent signatures of the response at one fixed time $T$.

4.1. *Generalized ESME and discussion of optimality.* In some sense, our method is standard: we assumed that we observe $N$ realizations of the random variable $\mathbf{Y}_{0,T}$ with distribution $\mathbb{Q}_T^{\theta}$ defined in (2.7). Then, we estimate $\theta$ by matching the empirical and theoretical expectation of that random variable, with the challenging part being computing the theoretical expectation.

Similarly, we can generalize the method and study its optimality in the standard way (see [9]): we consider functions $g(\mathbf{Y}_{0,T}, \cdot) : \Theta \to \mathbb{R}^d$ such that

$$\mathbb{E}_{\theta}(g(\mathbf{Y}_{0,T}, \theta)) \equiv 0.$$

An obvious choice is

$$g(\mathbf{Y}_{0,T}, \theta) = \{\mathbf{Y}_{0,T}^{\tau} - \mathbb{E}_{\theta}(\mathbf{Y}_{0,T}^{\tau})\}_{\tau \in V} \tag{4.1}$$

for $V$ as before. More generally, we can consider linear combinations of iterated integrals $\mathbf{Y}_{0,T}^{\tau}$. This is sufficient since products of iterated integrals can be written as linear combinations of iterated integrals. Then, the generalized moment matching estimator is defined as the solution to

$$\frac{1}{N} \sum_{i=1}^{N} g(\mathbf{Y}_{0,T}(\omega_i), \theta) = 0.$$

We define the generalized ESME to be the solution to the system above with expectations being approximated by the expectations of Picard iterations. For $g$ defined in (4.1), we get back the ESME.

The asymptotically optimal choice of function $g$ among all linear combinations of iterated integrals $\mathbf{Y}_{0,T}^{\tau}$ with $\tau \in V$ is the one minimizing asymptotic variance. The optimization can be done iteratively: suppose that we want to choose parameters $\{\alpha_{\sigma}\}_{\sigma \in V}$ such that the estimator constructed by solving

$$\sum_{\sigma \in V} \alpha_{\sigma} \left( \mathbb{E}_{\theta} \mathbf{Y}_{0,T}^{\sigma} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_{0,T}^{\sigma}(\omega_i) \right)$$

in term of $\theta$, has minimal variance among all linear combinations of $\mathbf{Y}_{0,T}^{\sigma}$. We go through the following steps:

(0) Choose an initial value $\alpha_{\sigma}(0)$ for the $\alpha_{\sigma}$'s.
(1) For the current value of the $\alpha_{\sigma}$'s, solve for $\theta$.
(2) Compute the asymptotic variance as a function of the $\alpha$'s and $\theta$.

(3) Set the $\alpha$'s equal to the argument minimizing the asymptotic variance in terms of the $\alpha$'s for $\theta$ the solution of step (1).

(4) Go to step (1).

This is also discussed in [9].

4.2. *Observing one path.* Suppose that we observe one realization of the solution of (2.5), namely $\{\mathbf{Y}_{0,t}(\omega)\}_{0\leq t\leq T}$. We are going to say that the rough path $\mathbf{Y}$ is ergodic if, for $T\to\infty$,

$$\frac{1}{T}\int_0^T \delta_{\mathbf{Y}_{0,t}(\omega)}\,dt \to \mu_{\theta_0} \text{ weakly}, \qquad \mathbb{Q}_{\theta_0}\text{-a.s.} \tag{4.2}$$

for $\theta_0$ in the parameter space $\Theta$. The limit $\mu_{\theta_0}$ is a distribution on the space of geometric rough paths $G\Omega_p(\mathbb{R}^m)$ and we call it *the invariant distribution*. Then, if $\mathbf{Y}_0\sim\mu_{\theta_0}$, the process will be stationary. In particular, for all $t\geq 0$ and words $\tau\in W_m$,

$$\mathbb{E}_{\theta_0}(\mathbf{Y}^{\tau}_{0,t}) = \mathbb{E}_{\theta_0}(\mathbf{Y}^{\tau}_{0}), \tag{4.3}$$

where the expectation $\mathbb{E}_{\theta_0}$ is with respect to $\mu_{\theta_0}$.Thus, for $T$ large and any $S\geq 0$

$$\frac{1}{T}\int_0^T \mathbf{Y}^{\tau}_{0,t}(\omega)\,dt \approx \mathbb{E}_{\theta_0}(\mathbf{Y}^{\tau}_{0,S}), \qquad \mathbb{Q}_{\theta_0}\text{-a.s.}, \tag{4.4}$$

where the left-hand side can be computed from the observations and the right-hand side is a function of $\theta_0$. However, as before, the expectation $\mathbb{E}_{\theta_0}(\mathbf{Y}^{\tau}_{0,S})$ will not be know in general. We approximate it using (3.4). We get

$$\mathbb{E}_{\theta_0}(\mathbf{Y}(r)^{\tau}_{0,S}) \approx \sum_{|\sigma|\leq|\tau|(q^r-1)/(q-1)} \mathbb{E}_{\theta_0}(\alpha^{\tau}_{r,\sigma}(\mathbf{Y}_0,0;\theta))\mathbb{E}(\mathbf{X}^{\sigma}_{0,S}).$$

According to Theorem 3.4, functions $\alpha^{\tau}_{r,\sigma}(y,0;\theta)$ are polynomials in $y$ of degree $|\tau|q^r$ where $q$ is the degree of polynomial $f$ in (2.5) with respect to $y$. Thus, we write

$$\mathbb{E}_{\theta_0}(\alpha^{\tau}_{r,\sigma}(\mathbf{Y}_0,0;\theta)) = \mathbb{E}_{\theta_0}\left(\sum_{k=1}^{|\tau|q^r} \mathbf{Y}^k_0\cdot c^{r,\sigma,\tau}_k(\theta)\right) = \sum_{k=1}^{|\tau|q^r}\mathbb{E}_{\theta_0}(\mathbf{Y}^k_0)\cdot c^{r,\sigma,\tau}_k(\theta).$$

The expectation $\mathbb{E}_{\theta_0}(\mathbf{Y}^k_0)$ is still unknown but can be approximated using (4.2). We end up with the equation

$$\begin{aligned}\frac{1}{T}\int_0^T &\mathbf{Y}^{\tau}_{0,t}(\omega)\,dt \\ &\approx \sum_{|\sigma|\leq|\tau|(q^r-1)/(q-1)}\left(\sum_{k=1}^{|\tau|q^r}\left(\frac{1}{T}\int_0^T \mathbf{Y}^k_{0,t}(\omega)\,dt\right)\cdot c^{r,\sigma,\tau}_k(\theta)\right)\mathbb{E}(\mathbf{X}^{\sigma}_{0,S}).\end{aligned}$$

The coefficients $c^{r,\sigma,\tau}_k(\theta)$ are polynomials with respect to $\theta$. By considering several different words $\tau\in W_m$, we construct a polynomial system of equations of $\theta$. As before, if $\hat{\theta}$ is a solution of the system, we call it the *expected signature matching estimator*.

Note that $S$ and $T$ do not need to be the same. In fact, $T$ should be large so that (4.4) holds while $S$ needs to be small in order for the local approximation of the expectation by Picard iterations to be valid.

REMARK 4.1. At the moment, there is no unified theory of ergodicity for rough paths. Some interesting results in this direction can be found in [8]. Note that we have assumed that the system is initialized by a rough path $\mathbf{Y}_0$ rather than a point $Y_0 \in \mathbb{R}^m$. This is consistent with the results in [8], where the authors point our the need to consider all the past $\{Y_t; -\infty < t \leq 0\}$ as an initializer of the system in order to make sense of ergodicity.

**5. Examples.** In this section, we use the ESME in specific examples of diffusions and fractional diffusions. The code that was used in these examples is written in *Mathematica* and can be found in http://chrisladroue.com/software/brownian-motion-and-iterated-integrals-on-mathematica/. It can be used to generate more examples corresponding to different choices of drift and diffusion coefficient.

5.1. *Diffusions.* First, we apply the ESME to estimate the parameters of the following Stratonovich SDE:

$$dY_t = a(1 - Y_t)\,dX_t^{(1)} + bY_t^2\,dX_t^{(2)}, \qquad Y_0^{(1)} = 0, \tag{5.1}$$

where $X_t^{(1)} = t$ and $X_t^{(2)} = W_t$. We chose an SDE because the expected signature of $(t, W_t)$ can easily be computed explicitly.

After three Picard iterations and replacing the expected signature of $(t, W_t)$ by its value (see [13, 16]), we get

$$\begin{aligned}
\mathbb{E}(\mathbf{Y}(3)_{0,t}^{(1)}) &= at - \frac{a^2t^2}{2} + \frac{a^3t^3}{6} + \frac{1}{4}a^3b^2t^4 - \frac{1}{10}a^4b^2t^5, \\
\mathbb{E}(2\mathbf{Y}(3)_{0,t}^{(1,1)}) &= a^2t^2 - a^3t^3 + \frac{7a^4t^4}{12} - \frac{a^5t^5}{6} + \frac{7}{10}a^4b^2t^5 \\
&\quad + \frac{a^6t^6}{36} - \frac{17}{20}a^5b^2t^6 + \frac{191}{420}a^6b^2t^7 \\
&\quad - \frac{11}{105}a^7b^2t^8 + \frac{21}{80}a^6b^4t^8 + \frac{1}{144}a^8b^2t^9 \\
&\quad - \frac{43}{180}a^7b^4t^9 + \frac{33}{700}a^8b^4t^{10} + \frac{1}{50}a^8b^6t^{11}.
\end{aligned}$$

This gives us an approximation of the moments of the solution as polynomials of the parameters.

The empirical moments are computed from the data. We generate 2,000 approximate realizations of paths of the solution using Milstein's method with discretization step 0.001. We use these paths to approximate the iterated integrals over the

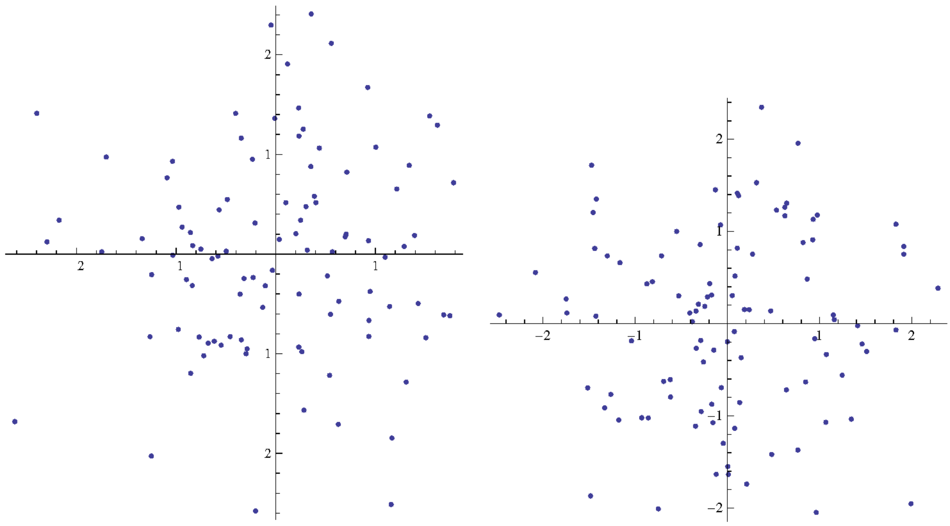


FIG. 1. 100 *realizations of the expected signature matching estimator, after centering and normalizing by the asymptotic variance*. Left: *from fractional Brownian motion paths* (*Hurst index* $= 11/24$), right: *from Brownian motion paths*.

interval $[0, \frac{1}{4}]$. We use the values $a = 1$ and $b = 2$. Then we get an approximation to the empirical moments at $T = \frac{1}{4}$ by averaging the different realizations of the iterated integrals of $\mathbf{Y}_{[0,1/4]}$.

Finally, by equating the empirical and theoretical approximations to the moments for $t = \frac{1}{4}$, we get a system of polynomials of $(a, b)$ of degree 14. We get two exact real solutions to this system: $(0.996353, -2.12892)$ and $(0.996353, 2.12892)$. As expected, the sign of $b$ cannot be identified. The estimates are very close to the true values.

We repeat this process 100 times and get 100 different estimates of $(a, b)$. In figure, we normalize the 100 positive solutions by the asymptotic variance (3.18), where $D_r^V(\theta)_{i,\tau}$ and $\Sigma_V(\theta_0)_{\tau,\tau'}$ in (3.15) are computed, the first using approximation of the theoretical moments from Picard iterations and the second is computed from the data by Monte Carlo. The normalized estimates are shown in Figure 1(right). Their covariance matrix is

$$\begin{pmatrix} 0.97172 & 0.0243445 \\ 0.0243445 & 0.954654 \end{pmatrix},$$

which is very close to the identity.

5.2. *Fractional diffusions*. We now apply the ESME to estimate the parameters of the differential equation driven by fractional diffusion with Hurst parameter $h > 1/4$. We choose the same vector field as before. Let

$$dY_t = a(1 - Y_t)\, dX_t^{(1)} + bY_t^2\, dX_t^{(2)}, \qquad Y_0^{(1)} = 0, \tag{5.2}$$

where $X_t^{(1)} = t$ and $X_t^{(2)} = B_t^h$, where $B_t^h$ is fractional Brownian motion with Hurst parameter $h$. Fractional Brownian motion generalizes Brownian motion, in the sense that it is a self-similar Gaussian process. It is defined as the Gaussian process with correlation given by

$$\mathbb{E}(B_s^h B_t^h) = \frac{1}{2}(|s|^{2h} + |t|^{2h} - |t-s|^{2h}).$$

Clearly, for $h = \frac{1}{2}$ we get independent intervals and Brownian motion. For $h > \frac{1}{2}$ the intervals are positively correlated and "smoother" than Brownian motion while for $h < \frac{1}{2}$ they are negatively correlated and they get more and more "rough" as $h$ gets smaller. In particular, the paths of fractional Brownian motion possess finite $p$-variation for every $p > \frac{1}{h}$.

Defining integration with respect to fractional Brownian motion is necessary in order for (5.2) to make sense. This is nontrivial and it is a very active area of research. One of the most successful approach is given by rough paths—but it is limited to $h > \frac{1}{4}$ (see [18] or [25] for a more recent approach), that is, to paths of finite $p$-variation for $p < 4$.

Having defined (5.2) as a differential equation driven by the rough path $(t, B_t^h)$, we can proceed to estimate the parameters $a$ and $b$. As in the diffusion case, we first construct an approximation to the theoretical moments, using Picard iterations. One difference is that up to this moment, an analytic expression for the expected signature is not known. Instead, we get a numerical approximation by simulating many paths of fractional Brownian motion, computing their iterated integral and then averaging.

We need to set some parameters: we choose $T = \frac{1}{4}$ as before and $h = \frac{11}{24}$. We use 1,000 paths of fractional Brownian motion with Hurst parameter $h = \frac{11}{24}$—these are exact simulations with discretization step $10^{-3}$—to compute the iterated integrals appearing in the Picard iteration and then average to approximate their expectations. We get the following formulas for the theoretical approximation of the first two moments of the response $Y$:

$$\begin{aligned}
\mathbb{E}\big(\mathbf{Y}(3)_{0,1/4}^{(1)}\big) &= 0.25a - 0.03125a^2 + 0.00260417a^3 \\
&\quad + 0.00044726a^2 b - 0.000111815a^3 b \\
&\quad + 4.97138 \times 10^{-6} a^4 b + 0.00116494a^3 b^2 \\
&\quad - 0.000115953a^4 b^2 + 2.53676 \times 10^{-6} a^4 b^3, \\
\mathbb{E}\big(2\mathbf{Y}(3)_{0,1/4}^{(1,1)}\big) &= 0.0625a^2 - 0.015625a^3 + 0.00227865a^4 \\
&\quad - 0.00016276a^5 + 6.78168 \times 10^{-6} a^6 \\
&\quad + 0.00022363a^3 b - 0.0000838612a^4 b \\
&\quad + 0.0000118036a^5 b - 8.93081 \times 10^{-7} a^6 b
\end{aligned}$$

$$
\begin{aligned}
&+ 2.58926 \times 10^{-8} a^7 b + 0.000814373 a^4 b^2 \\
&- 0.000246738 a^5 b^2 + 0.000033084 a^6 b^2 \\
&- 1.92279 \times 10^{-6} a^7 b^2 + 3.27969 \times 10^{-8} a^8 b^2 \\
&+ 4.39419 \times 10^{-6} a^5 b^3 - 1.24474 \times 10^{-6} a^6 b^3 \\
&+ 1.21202 \times 10^{-7} a^7 b^3 - 3.26456 \times 10^{-9} a^8 b^3 \\
&+ 5.74363 \times 10^{-6} a^6 b^4 - 1.31226 \times 10^{-6} a^7 b^4 \\
&+ 6.56898 \times 10^{-8} a^8 b^4 + 1.3868 \times 10^{-8} a^7 b^5 \\
&- 1.39803 \times 10^{-9} a^8 b^5 + 8.47574 \times 10^{-9} a^8 b^6.
\end{aligned}
$$

We create the data by numerically simulating 2,000 paths of the solution of (5.2) for $h = \frac{11}{24}$, $a = 1$ and $b = 2$ and discretization step $\delta = 10^{-3}$. We use a method proposed by Davie that is the equivalent of Milstein's method for differential equations driven by fractional Brownian motion (see [5] and references within). The error is of order $\delta^{3h-1}$, which for our choices of discretization step $\delta$ and Hurst parameter $h$ is 0.075.

Finally, we match the theoretical moments that are polynomials of $(a, b)$ with the empirical moments and solve the system. As in the diffusion case, we get two solutions corresponding to $b$ positive or negative. Since fractional Brownian motion is mean zero Gaussian process, we cannot expect to identify the sign of $b$.

We repeat the process 100 times to get 100 realizations of the estimates. These are shown in Figure 1(left), after normalization.

5.3. *Parameter estimation from one path.* As described in Section 4, we can apply this method on a single stationary path. We consider the fractional Orstein–Uhlenbeck process:

$$
dY_t = a(Y_t - b)\,dt + c\,dB_t^h, \qquad Y_0^{(1)} = Y_0. \tag{5.3}
$$

Through Picard iteration we compute the expansion of the two first moments. If $B_t^h$ is a Brownian motion, we obtain the polynomials

$$
\mathbb{E}(Y_t) = \mathrm{Y0} - abt - \tfrac{1}{2}a^2bt^2 - \tfrac{1}{6}a^3bt^3 - \tfrac{1}{24}a^4bt^4 - \tfrac{1}{120}a^5bt^5 + \cdots,
$$

$$
\mathbb{E}(Y_t^2) = \mathrm{Y0}^2 + c^2t + a^2b^2t^2 + ac^2t^2 + a^3b^2t^3 + \tfrac{2}{3}a^2c^2t^3 + \tfrac{7}{12}a^4b^2t^4 + \cdots.
$$

If $B_t^h$ is a fractional Brownian motion, the two moments also have an analytic expression.

We use Davie's method (see [5]) to numerically simulating one paths of the solution of (5.3) for $h = \frac{11}{24}$ and $h = \frac{1}{2}$, $a = -5, b = 2, c = 1$ and discretization step $\delta = 10^{-2}$. The error is of order $\delta^{3h-1}$, which for our choices of discretization step $\delta$ and Hurst parameter $h$ is 0.17.

Using the simulated data, we apply the method described in Section 4 for $T = 7$ and $S = 0.01$ in order to estimate $b$ and $c$. Figure 2 shows 500 realizations of the estimations of the mean $b$ and volatility $c$.

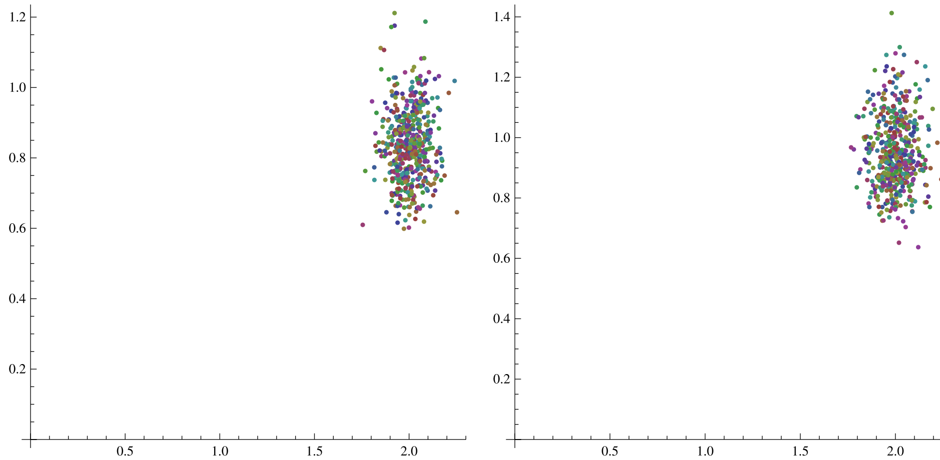


FIG. 2. 500 *realizations of the ESME for single paths*. Left: *from fractional O.U. paths* (*Hurst index* $= 11/24$), right: *from O.U. paths*.

## REFERENCES


[1] AÏT-SAHALIA, Y. and MYKLAND, P. A. (2008). An analysis of Hansen–Scheinkman moment estimators for discretely and randomly sampled diffusions. *J. Econometrics* **144** 1–26. MR2439920

[2] BESKOS, A., PAPASPILIOPOULOS, O. and ROBERTS, G. (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *Ann. Statist.* **37** 223–245. MR2488350

[3] BISHWAL, J. P. N. (2008). *Parameter Estimation in Stochastic Differential Equations*. *Lecture Notes in Math.* **1923**. Springer, Berlin. MR2360279

[4] CALDERON, C. P. (2007). Fitting effective diffusion models to data associated with a "glassy" potential: Estimation, classical inference procedures, and some heuristics. *Multiscale Model. Simul.* **6** 656–687 (electronic). MR2338498

[5] DEYA, A., NEUENKIRCH, A. and TINDEL, S. (2010). A Milstein-type scheme without Lévy area terms for SDEs driven by fractional Brownian motion. Available at arXiv:1001.3344.

[6] FRIZ, P. K. and VICTOIR, N. B. (2010). *Multidimensional Stochastic Processes as Rough Paths*: *Theory and Applications*. *Cambridge Studies in Advanced Mathematics* **120**. Cambridge Univ. Press, Cambridge. MR2604669

[7] GOBET, E., HOFFMANN, M. and REISS, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32** 2223–2253. MR2102509

[8] HAIRER, M. and PILLAI, N. S. (2009). Ergodicity of hypoelliptic SDEs driven by fractional Brownian motion. Available at arXiv:0909.4505.

[9] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

[10] HANSEN, L. P. and SCHEINKMAN, J. A. (1995). Back to the future: Generating moment implications for continuous-time Markov processes. *Econometrica* **63** 767–804. MR1343081

[11] HU, Y. and NUALART, D. (2010). Parameter estimation for fractional Ornstein–Uhlenbeck processes. *Statist. Probab. Lett.* **80** 1030–1038. MR2638974

[12] HULT, H. (2003). Approximating some Volterra type stochastic integrals with applications to parameter estimation. *Stochastic Process. Appl.* **105** 1–32. MR1972287

[13] KLOEDEN, P. E. and PLATEN, E. (1992). *Numerical Solution of Stochastic Differential Equations*. *Applications of Mathematics* (*New York*) **23**. Springer, Berlin. MR1214374

[14] KÜCHLER, U. and SØRENSEN, M. (1997). *Exponential Families of Stochastic Processes*. Springer, New York. MR1458891

[15] KUTOYANTS, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer, London. MR2144185

[16] LADROUE, C. (2010). Expectation of Stratonovich iterated integrals of Wiener processes. Available at arXiv:1008.4033.

[17] LI, J., KEVREKIDIS, P. G., GEAR, C. W. and KEVREKIDIS, I. G. (2007). Deciding the nature of the coarse equation through microscopic simulations: The baby-bathwater scheme. *SIAM Rev.* **49** 469–487 (electronic). MR2353808

[18] LYONS, T. and QIAN, Z. (2002). *System Control and Rough Paths*. Oxford Univ. Press, Oxford. MR2036784

[19] LYONS, T. J., CARUANA, M. and LÉVY, T. (2007). *Differential Equations Driven by Rough Paths*. *Lecture Notes in Math*. **1908**. Springer, Berlin. MR2314753

[20] NEUENKIRCH, A., NOURDIN, I. and TINDEL, S. (2008). Delay equations driven by rough paths. *Electron*. *J*. *Probab.* **13** 2031–2068. MR2453555

[21] PAPAVASILIOU, A. (2010). Coarse-grained modelling of multiscale diffusions: The $p$-variation estimates. In *Stochastic Analysis* 2010 (D. Crisan, ed.). Springer, Berlin.

[22] PRAKASA RAO, B. L. S. (1999). Statistical inference for diffusion type processes. In *Kendall's Library of Statistics* **8**. Oxford Univ. Press, New York. MR1717690

[23] REISS, M. (2005). Adaptive estimation for affine stochastic delay differential equations. *Bernoulli* **11** 67–102. MR2121456

[24] TUDOR, C. A. and VIENS, F. G. (2007). Statistical aspects of the fractional stochastic calculus. *Ann*. *Statist*. **35** 1183–1212. MR2341703

[25] UNTERBERGER, J. (2009). Stochastic calculus for fractional Brownian motion with Hurst exponent $H > \frac{1}{4}$: A rough path method by analytic extension. *Ann*. *Probab.* **37** 565–614. MR2510017

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
CV4 7AL, COVENTRY
UNITED KINGDOM
AND
DEPARTMENT OF APPLIED MATHEMATICS
UNIVERSITY OF CRETE
P.O. BOX 2208, 71409 HERAKLION
CRETE, GREECE
E-MAIL: A.Papavasiliou@warwick.ac.uk

BRISTOL GENETIC EPIDEMIOLOGY LABORATORIES
MRC CENTRE FOR CAUSAL ANALYSES
IN TRANSLATIONAL EPIDEMIOLOGY
SCHOOL OF SOCIAL AND COMMUNITY MEDICINE
UNIVERSITY OF BRISTOL
BRISTOL BS8 2BN
UNITED KINGDOM