Eur. J. Stat. 2 (2022) 13
doi: 10.28924/ada/stat.2.13

MLE Evolution Equation for Fractional Diffusions and Berry-Esseen Inequality of Stochastic Gradient Descent Algorithm for American Option

Jaya P. N. Bishwal

Department of Mathematics and Statistics, University of North Carolina at Charlotte,
376 Fretwell Bldg, 9201 University City Blvd. Charlotte, NC 28223-0001, USA
Correspondence: J.Bishwal@uncc.edu

Abstract. We study recursive parameter estimation in fractional diffusion processes. First, stability and asymptotic properties of the global maximum likelihood estimator (MLE) of the drift parameter are obtained under some regularity conditions. Then we obtain an evolution equation for the MLE of the drift parameter in nonhomogeneous fractional stochastic differential equation (fSDE) driven by fractional Brownian motion. This equation is then modified to yield an algorithm which is consistent, asymptotically efficient and converges to the MLE. The gradient and Newton type algorithm are first-order approximations. Finally we study the Berry-Esseen inequality for stochastic gradient descent in continuous time (SGDCT) algorithm for American option. We compare it with Longstaff-Schwartz regression based method.

1. Introduction and Preliminaries

Online parameter estimation is a challenging problem that appear frequently in fields such as robotics, neuroscience and finance in order to design adaptive filters and optimal controllers for unknown or changing systems. The approach here is based on modification of the offline maximum likelihood estimation.

First we introduce some basic tools from fractional stochastic calculus.

1.1 Fractional Brownian Motion

The fractional Brownian motion (fBm, in short), which provides a suitable generalization of the Brownian motion, is one of the simplest stochastic processes exhibiting long range-dependence. It was introduced by Kolmogorov (1940) in a Hilbert space framework and later on studied by Levy (1948) and in detail by Mandelbrot and Van Ness (1968).

Received: 20 Feb 2022.

Key words and phrases. fractional Brownian motion; maximum likelihood estimator process; evolution equation; recursive estimation; anticipative stochastic integral, Duhamel's principle; stochastic gradient descent algorithm; Berry-Esseen inequality; statistical learning, machine learning in finance; deep learning; American option.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables and processes below are defined.

A fractional Brownian motion $\{W_t^H, t \geq 0\}$ with Hurst parameter $H \in (0,1)$ is a centered Gaussian process with continuous sample paths whose covariance kernel is given by

$$E(W_t^H W_s^H) = \frac{V_H}{2}(s^{2H} + t^{2H} - |t-s|^{2H}), \quad s, t \geq 0$$

where

$$V_H := var(W_1^H) = \frac{1}{[\Gamma(H+\frac{1}{2})]^2}\left\{\frac{1}{2H} + \int_1^\infty \left[u^{H-\frac{1}{2}} - (u-1)^{H-\frac{1}{2}}\right]^2 du\right\}.$$

With $V_H = 1$, fBm is called a *normalized fBm*.

Properties

(P1) It has stationary increments: $E(W_t^H - W_s^H)^2 = |t-s|^{2H}, t, s \geq 0$.

(P2) $W_0^H = 0, E(W_t^H) = 0, E(W_t)^2 = |t|^{2H}, t \geq 0$.

(P3) When $H = \frac{1}{2}, W_t^{\frac{1}{2}}$ is the standard Brownian motion. The increments are independent.

(P4) The process is self similar or scale invariant, i.e., $(W_{\alpha t}^H, t \geq 0) =^d (\alpha^H W_t^H, t \geq 0), \alpha > 0$. H is also called the self similarity parameter.

(P5) The increments of the fBm are negatively correlated for $H < \frac{1}{2}$ and positively correlated for $H > \frac{1}{2}$.

(P6) For $H > \frac{1}{2}$, fBm is a long memory process since the covariance between far apart increments decrease to zero as a power law: $r(n) := E[W_1^H(W_{1+n}^H - W_n^H)] \sim C_H n^{2H-2}$ and $\sum_{n=1}^\infty r(n) = \infty$.

This property is also called long range dependence or long memory. The parameter H, measures the intensity of the long range dependence. Note that the estimation of the parameter H based on observation of fractional Brownian motion has already been paid some attention, see, e.g., Peltier and Levy Vehel (1994) and the references there in. However we assume H to be known.

(P7) The sample paths of W^H are almost surely Hölder continuous of any order less than H, but not Hölder continuous of any order greater than H, hence continuous but nowhere differentiable.

(P8) For any H, it has a finite $\frac{1}{H}$ variation, i.e.,

$$0 < \sup_\Pi E \sum_{t_i \in \Pi} \left[\left|W_{t_{i+1}}^H - W_{t_i}^H\right|^{\frac{1}{H}}\right] < \infty.$$

(P9) Law of the Iterated Logarithm (Arcones (1995)):

$$P\left(\overline{\lim}_{t\to 0+} \frac{W_t^H}{t^H(\log\log t^{-1})^{\frac{1}{2}}} = \sqrt{V_H}\right) = 1.$$

Self similarity of fBm leads to

$$P\left(\overline{\lim}_{t\to 0+} \frac{W_{\frac{1}{t}}^H}{(\log\log t^{-1})^{\frac{1}{2}}} = \sqrt{V_H}\right) = 1.$$

Setting $u = \frac{1}{t}$,

$$P\left(\overline{\lim}_{u\to\infty}\frac{W_u^H}{u^H(\log\log u^{-1})^{\frac{1}{2}}} = \sqrt{V_H}\right) = 1.$$

Strong Law of Large Numbers:

$$\lim_{u\to\infty}\frac{W_u^H}{u} = 0 \quad a.s.$$

(P10) fBm can be represented as a stochastic integral with respect to standard Brownian motion B (Mandelbrot and van Ness (1968)). For $H > \frac{1}{2}$,

$$W_t^H = \frac{1}{\Gamma(H+\frac{1}{2})}\left\{\int_{-\infty}^0 [(t-s)^{H-\frac{1}{2}} - (-s)^{H-\frac{1}{2}}]dB_s + \int_0^t (t-s)^{H-\frac{1}{2}}dB_s\right\}.$$

Standard Brownian motion can be written as a stochastic integral w.r.t W_t^H (see, Igloi and Terdik (1999)):

$$B_t = \frac{1}{\Gamma(\frac{3}{2}-H)}\left\{\int_{-\infty}^0 [(t-s)^{-H+\frac{1}{2}} - (-s)^{-H+\frac{1}{2}}]dW_s^H + \int_0^t (t-s)^{-H+\frac{1}{2}}dW_s^H\right\}.$$

(P11) With topological dimension n, the fractal dimension of fBm is $n+1-H$. Hausdorff dimension of one dimensional fBm is $2-H$.

(P12) *Existence of fBm*:

(i) It can be defined by a stochastic integral w.r.t. Brownian motion.

(ii) It can be constructed by Kolmogorov extension theorem (see, Samorodnitsky and Taqqu (1994)).

(iii) It can be defined as the weak limit of some random walks with strong correlations (see, Taqqu (1975)).

(P13) For $H \neq \frac{1}{2}$, the fBm is not a semimartingale and not a Markov process, but a Dirichlet process.

(P14) *Dirichlet Process*: A process is called a Dirichlet process if it can be decomposed as the sum of a local martingale and an adapted process of zero quadratic variation (zero energy). Obviously is a larger class of processes than semimartingales.

(P15) For $H < \frac{1}{2}$, the quadratic variation of W^H is infinite. For $H > \frac{1}{2}$, the quadratic variation of W^H is zero. Hence for $H > \frac{1}{2}, W^H$ is a Dirichlet process.

(P16) Fractional Brownian motion can be simulated using Cholesky decomposition method of the covariance matrix.

1.2 Stochastic Integral w.r.t. fBm

For $H \neq \frac{1}{2}$, the classical theory of stochastic integration with respect to semimartingales is not applicable to stochastic integration with respect to fBm. However, since fBm is a Gaussian process, stochastic integration with respect to Gaussian process is applicable.

For integration questions related to fractional Brownian motion, see Pipiras and Taqqu (2000). Now there exists several approaches to stochastic integration with respect to fBm:
(i) Classical Riemann sum approach : Lin (1995), Dai and Heyde (1996), Kleptsyna, Kloeden and Anh (1998c);
(ii) Malliavin calculus approach : Decreusefond and Ustunel (1998, 1999), Coutin and Decreusefond (1999a), Alos, Mazet and Nualart (2000, 2001);
(iii) Wick calculus approach : Duncan, Hu and Pasik-Duncan (1999);
(iv) Pathwise calculus : Young (1936), Zahle (1998, 1999), Ruzmaikina (2000);
(v) Dirichlet calculus : Lyons and Zhang (1994);
(vi) Rough path analysis : Lyons (1998), Lyons and Victoir (2007).

Lin (1995) introduced the stochastic integral as follows: Let $\pi : 0 < t_1 < t_2 < \cdots < t_n = 1$ be a partition of $[0,1]$. Let ϕ be a left continuous bounded Lebesgue measurable function with right limits, called *sure processes*. Then

$$\int_0^1 \psi(t)dW_t^H = \text{l.i.m.}_{|\pi|\to\infty} \sum_{t_i\in\pi} \psi(t_{i-1})(W_{t_i}^H - W_{t_{i-1}}^H).$$

The indefinite integral is defined as

$$\int_0^t \psi(s)dW_s^H = \int_0^1 \psi(t)I_{[0,t]}dW_t^H.$$

This integral has a continuous version and a Gaussian process. However,

$$E\left(\int_0^t \psi(s)dW_s^H\right) \neq 0.$$

To overcome this situation, Duncan, Hu and Pasik-Duncan (2000) introduced an integral using *Wick calculus* for which

$$E\left(\int_0^t f(s)dW_s^H\right) = 0.$$

They defined integrals of both Itô and Stratonovich type.

We shall discuss the Wick calculus approach here. Wiener integral for deterministic kernel was defined by Gripenberg and Norros (1996).

Let $\phi : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ be a Borel measurable deterministic function. Let

$$L_\phi^2(\mathbb{R}_+) := \left\{ f : |f|_\phi^2 = \int_0^\infty \int_0^\infty f(s)f(t)\phi(s,t)dsdt < \infty \right\}.$$

The inner product in the Hilbert space L_ϕ^2 is denoted by $\langle\cdot,\cdot\rangle_\phi$.

If $f,g \in L_\phi^2$, then $\int_0^\infty f_s dW_s^H$ and $\int_0^\infty g_s dW_s^H$ are well defined zero mean, Gaussian random variables with variances $|f|_\phi^2$ and $|g|_\phi^2$ respectively and covariance

$$E\left(\int_0^\infty f_s dW_s^H \int_0^\infty g_s dW_s^H\right) = \int_0^\infty \int_0^\infty f_s g_s \phi(s,t)dsdt =: \langle f,g\rangle_\phi.$$

Let $(\Omega, \mathcal{F}, P)$ be the probability space on which W^H is defined. For $f \in L^2_\phi$, define $\epsilon : L^2_\phi \to L^1(\Omega, \mathcal{F}, P)$ as

$$\begin{aligned} \epsilon(f) &:= \exp\left\{\int_0^\infty f_t dW_t^H - \frac{1}{2}\int_0^\infty \int_0^\infty f_s f_t \phi(s,t) ds dt\right\} \\ &= \exp\left\{\int_0^\infty f_t dW_t^H - \frac{1}{2}\int_0^\infty |f|^2_\phi\right\} \end{aligned}$$

which is called an exponential function.

Let $\mathcal{E}$ be the linear span of exponentials, i.e.,

$$\mathcal{E} = \left\{\sum_{k=1}^n a_k \epsilon(f_k) : n \in \mathbb{N}, a_k \in \mathbb{R}, f_k \in L^2_\phi(\mathbb{R}_+), k = 1, 2, \dots, n.\right\}$$

The Wick product of two exponentials is defined as

$$\epsilon(f) \diamond \epsilon(g) = \epsilon(f + g).$$

For distinct $f_1, f_2, \cdots, f_n \in L^2_\phi$, the exponentials $\epsilon(f_1), \epsilon(f_2), \cdots, \epsilon(f_n)$ are independent. It can be extended to define the Wick product of two functionals F and G in $\mathcal{E}$.

An analogue of Malliavin Derivative: Wick Derivative

The ϕ-derivative of a random variable $F \in L^p$ in the direction of Φg where $g \in L^2_\phi$ is defined as

$$D_{\Phi g}F(\omega) = \lim_{\delta \to 0} \frac{1}{\delta}\left[F\left(\omega + \delta \int_0^\cdot (\Phi g)(u) du\right) - F(\omega)\right]$$

if the limit exists in $L^p(\Omega, \mathcal{F}, P)$.

If there is a process $(D^\phi F_s, s \geq 0)$ such that

$$D_{\Phi g}F = \int_0^\infty D^\phi F_s g_s ds \quad a.s.$$

for all $g \in L^2_\phi$, then F is said to be ϕ-differentiable. Let $F : [0, T] \times \Omega \to \mathbb{R}$ be a stochastic process. The process is said to be ϕ-differentiable if for each $t \in [0, T]$, $F(t, \cdot)$ is ϕ-differentiable and $D^\phi_s F_t$ is jointly measurable.

Chain Rule: If $f : \mathbb{R} \to \mathbb{R}$ is smooth and $F : \Omega \to \mathbb{R}$ is ϕ-differentiable, then $f(F)$ is also ϕ-differentiable and

$$D_{\Phi g} f(F) = f'(F) D_{\phi g} F$$

and

$$D^\phi_s f(F) = f'(F) D^\phi_s(F).$$

(1) If $g \in L^2_\phi$, $F \in L^2(\Omega, \mathcal{F}, P)$ and $D_{\Phi g}F \in L^2(\Omega, \mathcal{F}, P)$, then

$$F \diamond \int_0^\infty g_s dW_s^H = F \int_0^\infty g_s dW_s^H - D_{\Phi g} F.$$

(2) If $g, h \in L^2_\phi$ and $F, G \in \mathcal{E}$, then

$$E\left(F \diamond \int_0^\infty g_s dW_s^H \quad G \diamond \int_0^\infty h_s dW_s^H\right) = E\left[D_{\Phi g} F D_{\Phi h} G + FG\langle g, h\rangle_\phi\right].$$

Let $\pi_n : 0 < t_1^{(n)} < t_2^{(n)} < \cdots < t_n^{(n)} = T$. Let $\mathcal{L}[0,T]$ be the family of stochastic processes on F on $[0,T]$ such that $E|F|_\phi^2 < \infty$, F is ϕ-differentiable, the trace of $(D_s^\phi F_t, 0 \le s \le T, 0 \le t \le T)$ exists and $E\int_0^T (D_s^\phi F_s)^2 ds < \infty$ and for each sequence of partitions $\{\pi_n, n \in \mathbb{N}\}$ such that as $|\pi_n| \to 0$

$$\sum_{i=0}^{n-1} E\left\{\int_{t_i^{(n)}}^{t_{i+1}^{(n)}} |D_s^\phi F_{t_i^{(n)}}^\pi - D_s^\phi F_s| ds\right\}^2 \to 0$$

and $E|F^\pi - F|_\phi^2 \to 0$ as $n \to \infty$.

For $F \in \mathcal{L}[0,T]$, define

$$\int_0^T F_s dW_s^H = l.i.m._{|\pi_n| \to 0} \sum_{i=0}^{n-1} F_{t_i} \diamond (W_{t_{i+1}}^H - W_{t_i}^H).$$

Proposition 1.1 Let $F, G, \in \mathcal{L}[0,T]$. Then
(i) $E\left(\int_0^T F_s dW_s^H\right) = 0$.
(ii) $E\left(\int_0^T F_s dW_s^H\right)^2 = E\left\{\left(D_s^\phi F_s ds\right)^2 + \left|I_{[0,T]} F\right|_\phi^2\right\}$.
(iii) $\int_o^t (aF_s + bG_s) dW_s^H = a\int_0^t F_s dW_s^H + b\int_0^t G_s dW_s^H$ a.s.
(iv) If $E\left[\sup_{0 \le s \le T} F_s\right]^2 < \infty$ and $\sup_{0 \le s \le T} E|D_s^\phi F_s|^2 < \infty$, then $\{\int_0^t F_s dW_s^H, 0 \le t \le T\}$ has a continuous version.

Here it is not assumed that $(F_s, s \in [0,T])$ is adapted to the fBm. Assume that $D_s^\phi F_s = 0, s \in [0,T]$. Then
(v) $E\left(\int_0^T F_s dW_s^H\right)^2 = \left|I_{[0,T]} F\right|_\phi^2 = E\int_0^T \int_0^T F_u F_v \phi(u,v) du dv$.

Fractional version of *Stratonovich Integral* is defined as

$$\int_0^t F_s \delta W_s^H := \int_0^t F_s dW_s^H + \int_0^t D_s^\phi F_s ds \text{ a.s.}$$

1.3 Fractional Itô Formula

If $f : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function with bounded derivatives of order two, then

$$f(W_T^H) - f(W_0^H) = \int_0^T f'(W_s^H) dW_s^H + H\int_0^T s^{2H-1} f''(W_s^H) ds \quad a.s.$$

For $H = \frac{1}{2}$, it gives the classical Itô formula for standard Brownian motion.

General Itô Formula

Let $\{F_u, 0 \leq u \leq T\}$ and $\{G_u, 0 \leq u \leq T\}$ be stochastic processes in $\mathcal{L}[0,T]$. Assume that there exists an $\alpha > 1 - H$ such that

$$E|F_u - F_v|^2 \leq C|u-v|^{2\alpha},$$

$$\lim_{|u-v|\to 0} E\{|D_u^\phi(F_u - F_v)|^2\} = 0$$

and

$$E \sup_{0\leq s\leq T} |G_s| < \infty.$$

Let

$$dX_t = G_t dt + F_t dW_t^H, X_0 = \xi \in \mathbb{R}, 0 \leq t \leq T,$$

i.e.,

$$X_t = \xi + \int_0^t G_s ds + \int_0^t F_s dW_s^H.$$

Let $f : \mathbb{R} \to \mathbb{R}$ be C_b^1 in the first variable and C_b^2 in the second variable and let $\left(\frac{\partial f}{\partial x}(s, X_s), s \in [0,T]\right) \in \mathcal{L}[0,T]$. Then

$$f(t, X_t) = f(0,\xi) + \int_0^t \frac{\partial f}{\partial s}(s, X_s) ds + \int_0^t \frac{\partial f}{\partial x}(s, X_s) G_s ds + \int_0^t \frac{\partial f}{\partial x}(s, X_s) F_s dW_s^H$$
$$+ \int_0^t \frac{\partial^2 f}{\partial x^2}(s, X_s) F_s D_s^\phi X_s ds.$$

Itô formula for Stratonovich Type integral:

Let $\{F_t, 0 \leq t \leq T\}$ satisfy the above assumptions. Let

$$\eta_t = \int_0^t F_s \delta W_s^H$$

be the Stratonovich integral. Let $g \in C_b^2$ and $\left(\frac{\partial g}{\partial x}(s, \eta_s) F_s, s \in [0,T]\right) \in \mathcal{L}[0,T]$.

Then for $t \in [0,T]$,

$$g(t, \eta_t) = g(0,0) + \int_0^t \frac{\partial g}{\partial s}(s, \eta_s) ds + \int_0^t \frac{\partial g}{\partial x}(s, \eta_s) F_s dW_s^H,$$

i.e.,

$$\delta g(t, \eta_t) = g_t(t, \eta_t) dt + g_x(t, \eta_t) d\eta_t.$$

1.4 Fractional Girsanov Theorem

Decreusefond and Ustunel (1999) gave a Girsanov formula using stochastic calculus of variation. Kleptsyna, Le Breton and Roubaud (1999) obtained the following Girsanov theorem.

Proposition 1.2 *Let h be a continuous function from $[0, T]$ to $\mathbb{R}$. Define for $0 < t \leq T$, the function $k_h^t = (k_h^t(s), 0 < s < t)$ by*

$$k_h^t(s) := -\rho_H^{-1} s^{\frac{1}{2}-H} \frac{d}{ds} \int_s^t d\omega \omega^{2H-1} (\omega - s)^{\frac{1}{2}-H} \frac{d}{d\omega} \int_0^\omega dz z^{\frac{1}{2}-H} (\omega - z)^{\frac{1}{2}-H} h(z)$$

where

$$\rho_H = \Gamma^2(3/2 - H)\Gamma(2H + 1) \sin \pi H.$$

Define for $0 \leq t \leq T$,

$$N_t^h := \int_0^t k_h^t(s) dW_s^H, \quad \langle N^h \rangle_t := \int_0^t h(s) k_h^t(s) ds.$$

Then the process $\{N_t^h, 0 \leq t \leq T\}$ is a Gaussian martingale with variance function $\{\langle N^h \rangle_t, 0 \leq t \leq T\}$.

For $h = 1$, the function k_h^t is $k_*^t(s) := \tau_H^{-1}(s(t-s))^{\frac{1}{2}-H}$ where $\tau_H := 2H\Gamma(3/2 - H)\Gamma(H + 1/2)$. Then the corresponding Gaussian martingale and its quadratic variation are

$$N_t^* = \int_0^t k_*^t(s) dW_s^H \text{ and } \langle N^* \rangle_t = \int_0^t k_*^t(s) ds = \lambda_H^{-1} t^{2-2H} \text{ where } \lambda_H = \frac{2H\Gamma(3 - 2H)\Gamma(H + 1/2)}{\Gamma(3/2 - H)}.$$

1.5 Anticipative Girsanov Theorem

The Carleman-Fredholm determinant is a complex-valued function which generalizes the determinant of a finite dimensional linear operator. We recall that the Carleman-Fredholm determinant of a Hilbert-Schmidt operator B from $L^2[0, T]$ into itself is defined by the product expansion

$$d_c(B) = \prod_{j=0}^{\infty} (1 - \lambda_j) \exp(\lambda_j)$$

where $\{\lambda_j, j \geq 0\}$ are the nonzero eigenvalues of B counted as many times as its multiplicities, see Simon (1979). In particular, if the operator B is nuclear, then

$$d_c(B) = \det(I - B) \exp\{\text{trace } B\}.$$

Thus if DU is nuclear, then

$$d_c(-DU) = \det(I + DU) \exp\{\text{trace } (-DU)\}.$$

The following is the nonadapted (anticipative) extension of the Girsanov theorem proved by Kusuoka (1982, Theorem 6.4). See also Theorem 4.1.2 in Nualart (1995).

Proposition 1.3 *Let $V : \Omega \to \Omega$ be a mapping of the form*

$$V(t, \omega) = \omega(t) + \int_0^t U(s, \omega) ds,$$

where U is a measurable mapping from Ω in to $H = L^2(0, T)$ and suppose that the following conditions are satisfied:

(i) V is bijective.

(ii) For all $\omega \in \Omega$, there exists a Hilbert-Schmidt operator $DU(\omega)$ from H into itself such that:

(a)

$$\|U(\omega + \int_0^\bullet h_s ds) - U(\omega) - DU(\omega)h\|_H = o(\|h\|_H)$$

for all $\omega \in \Omega$ as $\|h\|_H \to 0$,

(b) $h \to DU(\omega + \int_0^\bullet h_s ds)$ is continuous from H into $L^2([0,T]^2)$ the space of Hilbert-Schmidt operators for all ω,

(c) $I + DU(\omega) : H \to H$ is invertible.

Then if Q is the measure on $\Omega, \mathcal{F}$ such that $F = QV^{-1}$, Q is absolutely continuous with respect to P and

$$\frac{dQ}{dP} = |d_c(-DU)| \exp\left(-\int_0^T U(t)dW_t - \frac{1}{2}\int_0^T U_t^2 dt\right)$$

where $d_c(-DU)$ denotes the Carleman-Fredholm determinant of the Hilbert-Schmidt operator $-DU$ and $\int_0^T U(t)dW(t)$ is the Skorohod integral.

2. MLE Evolution Equation: MLE Dirichlet Process

2.1 Large Deviations

Consider the ordinary SDE

$$dX_t = f(\theta, X_t, t)dt + dW_t \tag{2.1}$$

where W is a standard Brownian motion.

We start with the uniform decay and equicontinuity results of parameter dependent stochastic integrals for unbounded parameter space, see Levanony *et al.* (1993).

Let the collection of continuous time martingales $\{F(\theta,t), \mathcal{F}_t, t \geq 0\}_{\theta\in\mathbb{R}}$ where for each (θ,t), $F(\theta,t) = \int_0^t f(\theta, X_s, s)dW_s$ is an Itô integral whose corresponding increasing process is $\langle F(\theta,t)\rangle_t = \int_0^t f^2(\theta, X_s, s)ds$.

Levanony *et al.* (1993) proved the following two results:

Proposition 2.1 *Suppose F and $\langle F\rangle_t$ are jointly continuous in mean square in (θ,t). Suppose there exists a $\gamma > 0$ such that for all $t_0 \geq 0$*

$$\lim_{|\theta|\to\infty} \inf_{t\geq t_0} \frac{\langle F(\theta,\cdot)\rangle_t}{(t|\theta|)^\gamma} = \infty.$$

Then

$$\lim_{|\theta|\to\infty} \sup_{t\geq t_0} \frac{|F(\theta,t)|}{\langle F(\theta,\cdot)\rangle_t} = 0.$$

Remark In unbounded paramter space sufficiency and Rao-Blackwellization of Vasicek model was studied in Bishwal (2011b). The optimal sampling problem was also solved.

Corollary 2.2 Suppose that there exists some $\delta > 0$ such that

$$\lim_{|\theta|\to\infty} \inf_{t\geq t_0} t^{-\delta}\langle F(\theta,\cdot)\rangle_t > 0 \text{ a.s. } t_0 \geq 0.$$

Then

$$\lim_{|\theta|\to\infty} \sup_{t\geq t_0} \frac{|F(\theta,t)|}{|\theta|^{\gamma}\langle F(\theta,t)\rangle_t} = 0 \ \forall\gamma > 0.$$

Now consider the fractional Ornstein-Uhlenbeck (fO-U)model satisfying the fractional SDE

$$dX_t = \theta X_t dt + dW_t^H, t \geq 0, \theta < 0.$$

We focus on the fundamental semimartingale behind the fO-U model. Define

$$\begin{aligned}
\kappa_H &:= 2H\Gamma(3/2-H)\Gamma(H+1/2),\\
k_H(t,s) &:= \kappa_H^{-1}(s(t-s))^{\frac{1}{2}-H},\\
\lambda_H &:= \frac{2H\Gamma(3-2H)\Gamma(H+\frac{1}{2})}{\Gamma(3/2-H)},\\
v_t \equiv v_t^H &:= \lambda_H^{-1}t^{2-2H}\\
\mathcal{M}_t^H &:= \int_0^t k_H(t,s)dW_s^H.
\end{aligned}$$

From Norros *et al.* (1999) it is well known that $\mathcal{M}_t^H$ is a Gaussian martingale, called the fundamental martingale whose variance function $\langle\mathcal{M}^H\rangle_t$ is v_t^H. Moreover, the natural filtration of the martingale $\mathcal{M}^H$ coincides with the natural filtration of the fBm W^H since

$$W_t^H := \int_0^t K(t,s)d\mathcal{M}_s^H$$

holds for $H \in (0.5, 1)$ where

$$K_H(t,s) := H(2H-1)\int_s^t r^{H-\frac{1}{2}}(r-s)^{H-\frac{3}{2}}dr, \quad 0 \leq s \leq t$$

and for $H = 0.5$, the convention $K_{1/2} \equiv 1$ is used.

Define

$$Q_t := \frac{d}{dv_t}\int_0^t k_H(t,s)X_s ds.$$

It is easy to see that

$$Q_t = \frac{\lambda_H}{2(2-2H)}\left\{t^{2H-1}Z_t + \int_0^t r^{2H-1}dZ_s\right\}.$$

Define the process $Z = (Z_t, t \in [0, T])$ by

$$Z_t := \int_0^t k_H(t,s)dX_s.$$

The following facts are known from Kleptsyna and Le Breton (2002):

(i) Z is the fundamental semimartingale associated with the process X.

(ii) Z is a $(\mathcal{F}_t)$ -semimartingale with the decomposition

$$Z_t = \theta \int_0^t Q_s dv_s + \mathcal{M}_t^H.$$

(iii) X admits the representation

$$X_t = \int_0^t K_H(t,s)dZ_s.$$

(iv) The natural filtration $(\mathcal{Z}_t)$ of Z and $(\mathcal{X}_t)$ of X coincide.

Now consider the fractional SDE

$$dX_t = f(\theta, X_t, t)dt + dW_t^H \tag{2.2}$$

where W^H is the fractional Brownian motion with Hurst parameter $H > 0.5$.

Let $\tilde{Z}$ is the fundamental semimartingale associated with the process X. Let the collection of continuous time martingales $\{G(\theta,t), \mathcal{G}_t, t \geq 0\}_{\theta\in\mathbb{R}}$ where for each (θ,t), $G(\theta,t) = \int_0^t f(\theta, \tilde{Z}_s, s)dW_s$ is an Itô integral whose corresponding increasing process is $\langle G(\theta,t)\rangle_t = \int_0^t f^2(\theta, \tilde{Z}_s, s)ds$.

From Theorem 3.4 in Buchman and Kluppelberg (2006), the fractional diffusion (2.2) can be represented as a monotone and differentiable functional of the fO-U process using the state space transform (SST) representation. Hence $\tilde{Z}$ can be represented as a SST of semimartingale in terms of Z.

Proposition 2.1 can be extended to the fractional SDE as follows:

Proposition 2.2 *Suppose G and $\langle G\rangle_t$ are jointly continuous in mean square in (θ, t). Suppose there exists a $\gamma > 0$ such that for all $t_0 \geq 0$*

$$\lim_{|\theta|\to\infty} \inf_{t\geq t_0} \frac{\langle G(\theta,\cdot)\rangle_t}{(t|\theta|)^\gamma} = \infty.$$

Then

$$\lim_{|\theta|\to\infty} \sup_{t\geq t_0} \frac{|G(\theta,t)|}{\langle G(\theta,\cdot)\rangle_t} = 0.$$

Corollary 2.2 can be extended to the fractional SDE as follows:

Corollary 2.3 Suppose that there exists some $\delta > 0$ such that

$$\lim_{|\theta|\to\infty} \inf \inf_{t\geq t_0} t^{-\delta}\langle G(\theta,\cdot)\rangle_t > 0 \; a.s. \; t_0 \geq 0.$$

Then

$$\lim_{|\theta|\to\infty} \sup_{t\geq t_0} \frac{|G(\theta,t)|}{|\theta|^{\gamma}\langle G(\theta,t)\rangle_t} = 0 \; \forall\gamma > 0.$$

Recall that by Girsanov theorem, the likelihood function of θ based on the observations $\{X_s, 0 \leq s \leq t\}$ is given by

$$L_t(\theta) = \exp\left\{\int_0^t f(\theta, X_s, s)dX_s - H(2H-1)\int_0^t f^2(\theta, X_s, s)(\int_0^s (s-r)^{2H-2}dr)ds\right\}. \quad (2.3)$$

Let

$$l_T(\theta) = \log L_t(\theta). \quad (2.4)$$

The MLE is defined as

$$\theta_t = \arg\sup_{\theta\in\mathbb{R}} l_t(\theta),$$

that is,

$$l_t(\theta_t) = \sup_{\theta\in\mathbb{R}} l_t(\theta).$$

Let

$$I_t = \int_0^t f_\theta^2(\theta_0, X_s, s)ds. \quad (2.5)$$

We have the strong consistency and asymptotic normality of the MLE:

Theorem 2.1

$$a) \quad \theta_t \to \theta_0 \quad a.s. \quad \text{as } t \to \infty,$$

$$b) \quad I_t^{1/2}(\theta_t - \theta_0) \to^{\mathcal{D}} \mathcal{N}(0, \; 1) \;\; \text{as } t \to \infty.$$

Proof: Due to the fundamental semimartingale representation $\widetilde{Z}$ of fractional diffusions along with state-space transform, main tools are Taylor expansion of the derivative of the log-likelihood $U_t(\theta)$ along with martingale SLLN and martingale CLT and delta method. We omit the details. □

Redefine the MLE as

$$\theta_t = \lim_{n\to\infty} \inf \arg\max_{|\theta|\leq n} l_t(\theta),$$

that is

$$l_t(\theta_t) = \sup_{\theta\in\mathbb{R}} l_t(\theta)$$

An $\mathcal{F}_t$-adapted MLE exists.

We derive the evolution equation for the trajectories of the MLE using the fractional Itô formula. Assume that our candidate for the MLE is a continuous Dirichlet process of the form

$$d\theta_t = a_t \; dt + b_t \; dX_t, \quad t \geq t_0. \quad (2.6)$$

The derivative (w.r.t. θ) of the log-likelihood $U_t(\theta)$ is a continuous Dirichlet process. Also $U_t(\cdot) \in C^2$ for all $t \geq 0$ a.s. and together with its derivatives is jointly (θ, t) continuous. Hence by fractional Itô formula

$$\begin{aligned} dU_t(\theta_t) &= f_\theta(\theta_t, X_t, t)[dX_t - f(\theta_t, X_t, t)dt] + R_t(\theta_t)d\theta_t \\ &\quad + H(2H-1)Q_t(\theta_t)b_t^2 dt + f_{\theta\theta}(\theta_t, X_t, t)b_t dt, \quad t \geq t_0 \end{aligned} \tag{2.7}$$

where R_t and Q_t are the second and the third derivative of the log-likelihood w.r.t. θ. Assuming that $R_t(\theta_t) < 0$ for all $t \geq t_0$, the MLE which solves $U_t(\theta) = 0 \ \ \forall t > 0$, is a solution of the equation

$$\begin{aligned} d\theta_t &= -R_t^{-1}(\theta_t)\{f_\theta(\theta_t, X_t, t)[dX_t - f(\theta_t, X_t, t)dt] \\ &\quad + [H(2H-1)Q_t(\theta_t)b_t^2 + f_{\theta\theta}(\theta_t, X_t, t)b_t]dt\}, \quad t \geq t_0 \end{aligned} \tag{2.8}$$

which after equating with (2.6) yields the MLE equation

$$\begin{aligned} d\theta_t &= -R_t^{-1}(\theta_t)\{f_\theta(\theta_t, X_t, t)[dX_t - f(\theta_t, X_t, t)dt] \\ &\quad + [H(2H-1)Q_t(\theta_t)R_t^{-2}(\theta_t)f_\theta^2(\theta_t, X_t, t) \\ &\quad - R_t^{-1}(\theta_t)f_\theta(\theta_t, X_t, t)f_{\theta\theta}(\theta_t, X_t, t)]dt\} \end{aligned} \tag{2.9}$$

with initial conditions: $|\theta_{t_0}| < \infty, \ U_t(\theta_{t_0}) = 0, \ R_t(\theta_{t_0}) < 0$.

The choice of the initial time $t_0 > 0$ is imposed by the fact that $R_0(\theta) = 0$ for all θ. Let $t_0 > 0$ be fixed. Define the stopping times

$$\tau := \inf\{t \geq t_0 : |\theta_t| = \infty\}, \qquad \sigma := \inf\{t \geq t_0 : |R_t(\theta_t)| = 0\}.$$

In fact, τ is the explosion time.

Existence and Uniqueness of the MLE Evolution Equation

Theorem 2.2 *The MLE equation (2.9) has a unique strong solution $\theta_t, \quad t_0 \leq t < \tau \wedge \sigma$.*

Proof. Write (2.9) in the form

$$d\theta_t = A(\theta_t, t)dt + B(\theta_t, t)dW_t^H \tag{2.10}$$

where the random functions A and B are obtained respectively by equating the drift and the diffusion term in (2.9). If A and B are jointly (θ, t) continuous and locally Lipschitz in θ (a.s.), then the proof follows from Kunita (1984, Theorem 3.4.5) and Mishura (2008). In our case since the term R^{-1} which appears in both A and B, may result in unbounded coefficients. Thus while in the classical local Lipschitz case, only truncation of θ_t is applied, here additional truncation argument is needed.

Fix $n < \infty$ and choose a C^∞ function ψ_n such that $\psi_n(\theta) = 1$ if $|\theta| \leq n, \ \psi_n(\theta) \in [0, 1]$ if $n \leq |\theta| \leq n+1, \ \psi_n(\theta) = 0, \ |\theta| > n+1$. Let $\phi_n(\theta, t) := \psi_n(R_t^{-1}(\theta))$. Define the truncated coefficients

$$\begin{aligned} A^n(\theta,t) &= A(\theta,t)\psi_n(\theta)\phi_n(\theta,t) \\ B^n(\theta,t) &= B(\theta,t)\psi_n(\theta)\phi_n(\theta,t) \end{aligned} \tag{2.11}$$

and consider the SDE

$$d\theta_t^n = A^n(\theta_t^n,t)dt + B^n(\theta_t^n,t)dW_t^H, \tag{2.12}$$

$$\theta_{t_0}^n = \theta_{t_0}\psi_n(\theta_{t_0})\phi_n(\theta_{t_0},t_0). \tag{2.13}$$

With this truncation and since A^n and B^n are jointly continuous and continuously differentiable w.r.t. θ (with jointly continuous derivatives), A^n and B^n are globally Lipschitz with globally linear growth a.s. Thus by Kunita (1984, Theorem 3.4.1), (2.12) possesses a unique strong solution $\{\theta_t^n, t_0 \le t < \infty.\}$

Define $S^n = \inf\{t \ge t_0 : |\theta_t^n| > n \text{ or } R_t(\theta_t^n) > -1/n\}$ and note that (2.12) coincides with (2.10) for all $t \in [t_0, S^n)$. Let $S^\infty = \lim_{n\to\infty} S^n$ and define $\{\theta_t,\ t_0 \le t < S^\infty\}$ by $\theta_t = \theta_t^n$ if $t < S^n$. With this definition, one has $S^\infty = \sigma \wedge \tau$ and $\{\theta_t,\ t_0 \le t < \sigma\wedge\tau\}$, a unique strong solution of (2.10).

□

Theorem 2.3 *If the MLE θ_t has a.s. continuous trajectories, then (2.9) holds for θ_t for sufficiently large t. If in addition, $P(R_t(\theta_t) < 0\ \forall t > 0) = 1$, the log-likelihood $l_t(\cdot)$ is strictly concave in some small neighborhood of the MLE for all $t > 0$, then (2.9) is the MLE evolution equation on $[t_0,\infty)$ a.s. for all $t_0 > 0$.*

Proof: Fix some $\epsilon > 0$, choose t_0 such that $P(t_0 > T) > 1-\epsilon$ and consider equation (2.9) on $[t_0, \tau\wedge\sigma)$ with initial condition $\theta_{t_0} = \tilde{\theta}_{t_0}$. Applying the fractional Itô Velntzell formula together with the initial condition $U_t(\theta_{t_0}) = U_t(\tilde{\theta}_{t_0}) = 0$ implies that $U_t(\tilde{\theta}_t) = 0$ for all $t \in [t_0, \tau\wedge\sigma)$. This and the fact that $R_t(\tilde{\theta}_t) < 0$ for all $t \in [t_0, \tau\wedge\sigma)$ indicate that $\tilde{\theta}_t$ is a strict maximum of $L_t(\cdot), t \in [t_0, \tau\wedge\sigma)$.

We now show that $\theta_t = \tilde{\theta}_t$ for all $t \in [t_0, \tau\wedge\sigma)$. Let $0 < \Delta_n \to 0$, define $t_n = t_{\Delta_n}$ and $\delta_{t_n} = |\theta_{t_n} - \tilde{\theta}_{t_n}|$. Then since $U_t(\theta_t) = U_t(\tilde{\theta}_t) = 0$ for all $t \in [t_0, \tau\wedge\sigma)$ it holds that $0 = |U_t(\theta_{t_n}) - U_t(\tilde{\theta}_{t_n})| = |R_{t_n}(\bar{\theta}_{t_n})|\delta_{t_n}$ for some $\bar{\theta}_t \in [\theta_t, \tilde{\theta}_t]$ which, because $\delta_{t_n} > 0$, for all n, results in $R_{t_n}(\bar{\theta}_{t_n}) = 0$ for all n.

Therefore, since by definition $t_n \to S$ and $\delta_{t_n}\delta_s$ (due to sample path continuity of θ_t and $\tilde{\theta}_t$), one may utilize the joint continuity of R to conclude that $R_{t_n}(\bar{\theta}_{t_n}) \to R_s(\theta_s) = R_s(\tilde{\theta}_s) = 0$ which by definition results in $S = \sigma$. Since this contradicts the underlying assumption (that bifurcation occurs before $\tau\wedge\sigma$), it confirms the validity of (2.9) for the MLE on $[t_0, \tau\wedge\sigma)$.

Now by the definition of T, $P(R(\theta_t) < 0\ \forall\ t \ge t_0) > 1-\epsilon$. This and the boundedness of θ_t imply that (2.9) holds for θ_t on $[0,\infty)$ w.p. $> 1-\epsilon$. The condition on T leads to second assertion (where $T = 0$ a.s.).

Remarks:

1. $R_t(\theta_t) \to -\infty$ a.s.

2. The sample paths of MLE are continuous and bounded. The MLE process is stable, i.e., it does not explode: $\sup_{t \geq t_0} |\theta_t| < \infty$ *a.s.* This along with $R_t(\theta_t) \to -\infty$ shows that $\tau \wedge \sigma = \infty$.

Newton-type Algorithm

Newton type algorithms are approximation of the MLE equation (2.9). However, (2.9) is not suitable for recursive estimation, it is valid for large t, and moreover, it requires the knowledge of exact MLE at the initial time.

Newton type algorithms are insensitive to initial conditions and implementable for all $t_0 > 0$. The algorithm makes the estimator θ_t follow the gradient when $U \neq 0$ until it enters the neighborhood of a local maximum and then keeps θ_t in this neighborhood as long as possible , i.e., as long as singularity does not arise (where afterwards the process repeats itself). This switching policy is needed in order to maintain the necessary flexibility which prevents the estimator for being 'trapped' in a no solution situation (i.e, when $R = 0$ in (2.9)).

Fix $\alpha > 0$ and some small ϵ, δ, define the set

$$A(t) = \{\theta : |U_t(\theta)| \leq \delta, R_t(\theta) \leq -\epsilon\}. \tag{2.14}$$

A *simplified version of the Newton Algorithm* is

$$\begin{aligned} d\theta_t \quad = \quad & -R_t^{-1}(\theta_t)\{f_\theta(\theta_t, X_t, t)[dX_t - f(\theta_t, X_t, t)dt] \\ & +[H(2H-1)Q_t(\theta_t)R_t^{-2}(\theta_t)f_\theta^2(\theta_t, X_t, t) \\ & -R_t^{-1}(\theta_t)f_\theta(\theta_t, X_t, t)f_{\theta\theta}(\theta_t, X_t, t) \\ & +\alpha \quad U_t(\theta_t)]dt\}I_{\{\theta_0^t \in A(t)\}} + t^{-\nu}U_t(\theta_t)dtI_{\{\theta_0^t \notin A(t)\}} \end{aligned} \tag{2.15}$$

with initial condition $\theta_{t_0}, t_0 > 0$.

When $\theta_t \in A(t)$, the algorithm follows the likelihood equation (with a decay term), where as when $\theta_t \in A^c(t)$, it follows the gradient towards a local maximum. The main problem with (2.14) is the fact that this scheme could result in infinitely many switchings in the bounded time intervals (or even uncountably many switchings). This prevents (2.14) from being an implementable algorithm.

Choose continuous $0 < \delta_t \downarrow 0$ and $0 < \epsilon_t \downarrow 0$ where δ_t satisfies

$$\int_{t_0}^{\infty} \delta_t dt = \infty, \quad (s/t)^\nu < \delta_t/\delta_s \; \forall t_0 \leq s < t. \tag{2.15}$$

For example $\delta = t^{-\beta}, 0 < \beta < 1 \wedge \nu$ will do.

Redefine the set $A(t)$,

$$A(t) := \{\theta : |U_t(\theta)| \leq \delta_t t^\nu, R_t(\theta) \leq -\epsilon_t\}. \tag{2.16}$$

Let

$$\mathcal{A}(t) := \{\phi_0^t \in C[0, t] : \exists s \leq t \text{ such that } R_s(\phi_s) \leq -2\epsilon_s \text{ and } \phi_r \in A(r) \forall r \in [s, t]\}. \tag{2.17}$$

$\mathcal{A}(t)$ sets for R the 'entrance level' $-2\epsilon_t$ into $A(t)$ and 'exit level' $-\epsilon_t$ (into and from $A(t)$ respectively).

The changes in (2.15) are in the definition of good event and the normalizing of the second term. The proposed *algorithm* is given by

$$\begin{aligned} d\theta_t &= -R_t^{-1}(\theta_t)\{f_\theta(\theta_t, X_t, t)[dX_t - f(\theta_t, X_t, t)dt] \\ &\quad +[H(2H-1)Q_t(\theta_t)R_t^{-2}(\theta_t)f_\theta^2(\theta_t, X_t, t) \\ &\quad -R_t^{-1}(\theta_t)f_\theta(\theta_t, X_t, t)f_{\theta\theta}(\theta_t, X_t, t) \\ &\quad +\alpha\; U_t(\theta_t)]dt\}\; I_{\{\theta_0^t\in\mathcal{A}(t)\}} + t^{-\nu}U_t(\theta_t)dt\; I_{\{\theta_0^t\notin\mathcal{A}(t)\}} \end{aligned} \tag{2.18}$$

which holds in $[t_0, \tau)$ (where τ is the explosion time), with any initial condition θ_{t_0}, $t_0 > 0$ (where $\theta_t = \theta_{t_0}\forall t \in [0, t_0]$).

Theorem 2.4 *The equation (2.18) possesses unique strong solution in* $[t_0, \tau)$.

Notice the difference between algorithm (2.18) and the conventional Newton type algorithm which is given by

$$d\widetilde{\theta}_t = -\bar{R}_t^{-1}(\theta_t)f_\theta(\widetilde{\theta}_t, X_t, t)\;[dX_t - f(\widetilde{\theta}_t, X_t, t)dt],\;\; \bar{R}_t(\widetilde{\theta}_{t_0}) < 0\;\; a.s. \tag{2.19}$$

where $\bar{R}$ is an approximation of R which is computed in a recursive way.

(2.19) is a first order approximation to the optimal algorithm where the drift terms are added in the Newton phase.

Using Corollary 2.3, we obtain

$$\lim_{|\theta|\to\infty}\sup_{t\geq t_0}\frac{|U_t(\theta)|}{I_t} = \infty \;\; a.s. \;\; \forall\; t_0 > 0.$$

This in turn gives the boundedness of the MLE:

Theorem 2.5

$$\sup_{t\geq t_0}|\theta_t| < \infty \;\; a.s.$$

Define the Fisher information process

$$I_t = I_t(\theta) = \int_0^t f_\theta^2(\theta, X_s, s)ds. \tag{2.20}$$

Define the empirical Fisher information process

$$\widehat{I_t(\theta)} = \int_0^t f_\theta^2(\theta_t, X_s, s)ds. \tag{2.21}$$

Theorem 2.6 *If* $\widetilde{\theta}_t$ *satisfies*

$$I_t^{-1/2}U_t(\widetilde{\theta}_t) \to 0 \;\; a.s. \text{ and } \sup_{t\geq t_0}|\widetilde{\theta}_t| < \infty \;\; a.s.,$$

then we have

$$a)\quad \widetilde{\theta}_t \to \theta_0 \quad a.s. \quad as\ t \to \infty,$$

$$b)\quad I_t^{1/2}(\widetilde{\theta}_t - \theta_0) \to^{\mathcal{D}} \mathcal{N}(0,\ 1)\quad as\ t \to \infty.$$

$$c)\quad I_t^{1/2}(\widetilde{\theta}_t - \theta_t) \to 0\quad a.s.\quad as\quad t \to \infty.$$

Proof. The consistency is based on the given conditions of the theorem. Asymptotic normality can be shown same way as in Theorem 2.1. Expanding $I_t^{1/2}(U_t(\widetilde{\theta}_t) - U_t(\theta_t))$ around θ_0 and using part of the theorem and Theorem 2.1, we obtain the result. We omit the details. □

Thus the Newton estimator and the MLE are asymptotically equivalent. In fact, one can obtain higher speed of convergence as follows:

Theorem 2.7 *For every* $0 < \alpha' < \alpha$ *(where* α *is from (2.15)),*

$$e^{\alpha' t} I_t(\widetilde{\theta}_t - \theta_t) \to 0 \qquad a.s.\quad as\quad t \to \infty.$$

Proof. Since $dU = -\alpha U dt$ there exists some large enough T such that

$$U_t(\theta_t) = U_T(\theta_T)e^{-\alpha(t-T)} = (\widetilde{\theta}_t - \theta_t)R_t(\bar{\theta}_t), \quad \bar{\theta}_t \in [\widetilde{\theta}_t, \theta_t]\ \ \forall t \geq T.$$

Since $\theta_t \to \tilde{\theta}_t \to \theta_0$ and $I_t^{-1}R_t(\theta_0) \to -1$ a.s., then due to the equicontinuity of $\{I_t^{-1}R_t(\cdot)\}_{t \geq t_0}$ we have $I_t^{-1}R_t(\bar{\theta}_t) \to -1$ a.s. This implies that $P(\sup_{t \geq T} R_t(\bar{\theta}_t) < 0) = 1$ which enables us to define $Y_t = -I_t^{-1}R_t^{-1}(\bar{\theta}_t), t \geq T$. Hence

$$U_T(\theta_T)Y_t e^{-\alpha(t-T)} = I_t(\widetilde{\theta}_t - \theta_t).$$

Choose some $0 < \alpha' < \alpha$, define $V = U_T(\theta_T)e^{\alpha T}$. Multiplying both sides by $e^{\alpha' t}$, we have

$$VY_t e^{-(\alpha-\alpha')t} = I_t(\widetilde{\theta}_t - \theta_t)e^{\alpha' t}.$$

Since $Y_t \to 1$, this leads to

$$|Ve^{-(\alpha-\alpha')t} - I_t(\widetilde{\theta}_t - \theta_t)e^{\alpha' t}| \to 0\ \ a.s.$$

Since $Ve^{-(\alpha-\alpha')t} \to 0$ (due to the almost sure finiteness of V), the theorem follows. □

Remarks

1. This shows much higher convergence speed than the classical result with rate $I_t^{1/2}$.
For $\alpha = 0$, $\ I_t^{\phi}|\widetilde{\theta}_t - \theta_t| \to 0\ \ a.s.\ $ as $\ t \to \infty\ \ \forall \phi < 1$.

3. Stochastic Gradient Descent in Continuous Time

In standard discrete time version of stochastic gradient descent, data is usually considered to be i.i.d. at every step. Thus it is natural to ask if one can discretize (2.1), for example by Euler-Maruyama method and apply traditional stochastic gradient descent. This can result in loss of accuracy, or may not even converge. For example, there is no guarantee that using a higher order discretization scheme, for example the second order Milstein scheme, to discretize the dynamics of the SDE (2.1) and then applying the traditional stochastic gradient descent will produce a statistical learning scheme which is higher-order accurate in time. Hence it makes sense to first develop the continuous-time statistical learning equation and then apply higher-order accurate numerical scheme.

Stochastic gradient descent in continuous time (SGDCT) provides a computationally efficient method for the statistical learning of continuous-time models. SGDCT algorithm follows a descent direction along a continuous stream of data. The parameter updates occur in continuous time and satisfy a stochastic differential equation (SDE). We analyze the asymptotic convergence rate by proving a central limit theorem. An L^p convergence rate is also proven.

Statistical estimation in SDEs have been studied using entire observed path of X, i.e. batch optimization, see Bishwal (2008). The vast majority of statistical learning, machine learning and stochastic gradient descent literature address discrete time algorithm. This paper analyzes statistical learning in continuous time (SGDCT). SGDCT can estimate unknown parameters and functions in SDE models. It is related to online maximum likelihood filtering and identification.

SGDCT can be used to solve continuous-time optimization problem such as American options. The value function is approximated by a parametric function and the parameter is estimated by SGDCT algorithm. Recall that Longstaff-Schwartz estimated the parameter by least squares method. One could discretize the dynamics and then use the Q-learning algorithm. The Q-learning algorithm is biased while SGDCT algorithm is unbiased.

The structure of the algorithm indicates that well known gradient and Newton type algorithm are first order approximations.

Consider the SDE

$$dX_t = f^*(X_t)dt + \sigma dW_t \tag{3.1}$$

where $f^*(x)$ is an unknown function. The goal is to estimate $f(x,\theta)$ from continuous observations of $(X_t)_{t\geq 0}$. The function may be convex or non-convex.

The SGD update in continuous time for the parameter $\theta \in \mathbb{R}$ satisfies the SDE

$$d\theta_t = \frac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)dX_t - \frac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t)dt \tag{3.2}$$

where α_t is the *learning rate*. For example, α_t could be $\frac{C_\alpha}{C_0+t}$. We assume that θ_0 is initialized according to some distribution with compact support.

The parameter update can be used both for statistical estimation given previously observed data as well as online learning, i.e, statistical estimation in real time as data becomes available.

Define the objective function

$$g(x,\theta) = \frac{1}{2}\|f(x,\theta) - f^*(x)\|^2_{\sigma^2} = \frac{1}{2\sigma^2}\langle(f(x,\theta) - f^*(x))\rangle^2 \tag{3.3}$$

which measures the distance between the model $f(x,\theta)$ and true dynamics $f^*(x)$ for a specific x. This is a minimum distance type estimator.

We assume that X_t is ergodic and it has some well behaved $\pi(dx)$ as its unique invariant distribution. Let the average over $\pi(dx)$ be denoted by

$$\bar{g}(\theta) = \int g(x,\theta)\pi(dx) \tag{3.4}$$

where $\pi(dx)$ is the invariant measure of X_t when it is ergodic, which is the natural objective function for our analysis of the asymptotic behavior of the algorithm θ_t. This is an weighted average of the distance between $f(x,\theta)$ and $f^*(x)$ where the weights are given by $\pi(dx)$, which is the distribution that X_t tends to when t become large. The distance $g(x,\theta)$ is decreased by moving θ in the descent direction $-g'(x,\theta)$ which motivates the algorithm

$$\begin{aligned}
d\theta_t &= -\alpha_t g'(X_t,\theta_t)dt \\
&= \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)(f^*(X_t) - f(X_t,\theta))dt \\
&= -\alpha_t g'(x,\theta_t)dt + \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t))\sigma dW_t \\
&= -\alpha_t \bar{g}'(x,\theta_t)dt - \alpha_t(g'(x,\theta_t) - \bar{g}'(x,\theta_t))dt + \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t))\sigma dW_t \\
&=: I_1 + I_2 + I_3
\end{aligned} \tag{3.5}$$

where I_1= Descent term, I_2 = fluctuation term, I_3= Noise term.

If α_t decays with time, e.g., $\alpha_t = \frac{C_\alpha}{C_0+t}$, the descent term I_1 will dominate the fluctuation term and the noise term for large t. Then θ_t, will converge to a local minimum of $\bar{g}(\theta)$. Sirignano and Spiliopoulos (2017) proved that θ_t converges to a critical point of the objective function $\bar{g}(\theta)$: $|\bar{g}'(\theta_t)| \to 0$ almost surely as $t \to \infty$.

Since $\lim_{t\to\infty} \alpha_t = 0$, the descent term $\alpha_t \bar{g}'(\theta_t) \to 0$ as $t \to \infty$. Descent term converges to zero as $t \to \infty$. Sirignano and Spiliopoulos (2018) proved the rate at which θ_t converges to zero. They obtained a central limit theorem (CLT) and an L^p convergence rate.

When $\bar{g}(\theta)$ has a single critical pint θ^*, let

$$J(\theta^*) := C_\alpha^2 \int_0^\infty e^{-2s(C_\alpha \bar{g}(\theta^*)-1)}\bar{h}(\theta^*)ds \tag{3.6}$$

Define

$$\Psi^{(p)}_{t,s} := \exp(-pC\int_s^t \alpha_u du),\ \ p \geq 1 \tag{3.7}$$

and let $\Phi^*_{t,s}$ be the solution to the ODE

$$d\Phi^*_{t,s} = -\alpha_t \bar{g}(\theta^*)\Phi^*_{t,s}dt, \quad \Phi^*_{s,s} = 1. \tag{3.8}$$

We assume that the general learning rate α_t, satisfies the following conditions:
(B1) $\int_0^\infty \alpha_t dt = \infty$,
(B2) $\int_0^\infty \alpha_t^2 dt < \infty$,
(B3) $\int_0^\infty |\alpha_t'| dt < \infty$,
(B4) There exists a $p > 0$ such that $\lim_{t\to\infty} \alpha_t^2 t^p = 0$,
(B5) $\int_0^t \alpha_s^{5/2} \Psi_{t,s}^{(2)} ds \le o(\alpha_t)$,
(B6) $\int_0^t \alpha_s^2 {\Phi^*}_{t,s}^2 ds = O(\alpha_t)$,
(B7) $\int_0^t \alpha_s^2 \Psi_{t,s}^{(1)} ds \le o(\alpha_t^{1/2})$,
(B8) For all $p \ge 2$, $\int_0^t (\alpha_s^2 + |\alpha_s'|) \Psi_{t,s}^{(p)} \alpha_s^{p/2-1} ds \le o(\alpha_t^{p/2})$,
(B9) For all $p \ge 2$, $\Psi_{t,0}^{(p)} \le O(\alpha_t^{p/2})$
(B10) $\Psi_{t,0}^{(1)} \le o(\alpha_t^{1/2})$.

The L^p convergence rate is given by

$$E|\theta_t - \theta^*|^p \le K \alpha_t^{p/2} \tag{3.9}$$

for $p \ge 1$. The CLT is given by

$$\alpha_t^{-1/2}(\theta_t - \theta^*) \to^D \mathcal{N}(0, J(\theta^*)). \tag{3.10}$$

A standard choice of the learning rate α_t which satisfies (B1)–(B10) is $\alpha_t = C_\alpha(C_0 + t)^{-1}$. Hence the L_p convergence rate is given by

$$E[|\theta_t - \theta^*|^p] \le \frac{K}{(C_0 + t)^{p/2}} \tag{3.11}$$

for $p \ge 1$ and the CLT for θ_t is given by

$$\sqrt{t}(\theta_t - \theta^*) \to^D \mathcal{N}(0, J(\theta^*)) \quad \text{as} \quad t \to \infty. \tag{3.12}$$

Sirignano and Spiliopoulos (2020) derived a stochastic integral to represent the $\sqrt{t}(\theta_t - \theta^*)$ using Duhamel's principle and the fundamental solution of the random ODE $d\Psi_{t,s} = -\alpha_t \bar{g}(\tilde{\theta}_t)\Psi_{t,s} dt$ where $\tilde{\theta}_t$ lies on a line connecting θ^* and θ_t. The integrand of this stochastic integral includes the fluctuation term and the noise term as well as Ψ_t and is *anticipative*. Hence standard approach such as Itô isometry cannot be applied directly. Also since $f(x, \theta)$ is allowed to grow with θ, hence the fluctuations as well as other terms can grow with θ. Hence they prove an a priori stability estimate for $|\theta_t|$. Proving central limit theorem for non-convex $\bar{g}(\theta)$ is not straightforward since the convergence speed of θ_t can arbitrarily slow in certain regions, and the gradient can even point away from the global minimum θ^*. To address this, we consider the stochastic integral after the time τ_δ which is defined as the final time θ_t enters a neighborhood of θ^*. However, τ_δ is anticipative, i.e., is not a stopping time, therefore careful analysis is required to study the behavior of the stochastic integral.

Let $Y_t := \theta_t - \theta^*$ be the error term. It satisfies

$$dY_t = -\alpha_t \Delta\bar{g}(\theta_t^1) Y_t dt + \alpha_t(\bar{g}'(\theta_t) - g'(X_t, \theta_t))dt + \alpha_t f'(X_t, \theta_t) dW_t. \tag{3.13}$$

$$E|Y_t|^2 \le Kt^{-1}, \quad E|Y_t|^p \le Kt^{-p/2}. \tag{3.14}$$

The stochastic integral $\sqrt{t}\int_1^t \alpha_s^2 \Phi_{t,s}\zeta(X_s,\theta_s)dW_s \to 0$ in probability as $t \to \infty$ where $\zeta(x,\theta)$ is a function that can grow at most polynomially in x and θ.

For the analysis of fluctuation term, the proofs use Poisson PDE for ergodicity, given below in Proposition 3.1. The central limit theorem for non-convex $\bar{g}(\theta)$ is challenging since the convergence speed of θ_t can become arbitrarily slow in certain regions and the gradient can even point away from the global minimum θ^*. *Interalia,* they prove convergence to zero of multidimensional stochastic integrals. The proof requires the analysis of stochastic integral with anticipative integrands, which is challenging since standard approaches like Itô isometry can not be directly applied.

Sirignano and Spiliopoulos (2020) remark that $t^{-1/2}$ is the fastest possible convergence rate given that the noise is Brownian motion. This is due to the quadratic variation of Brownian motion growing linearly in t. With other noises whose varianes grows sublinearly in time, one could allow for faster rate of convergence than $t^{-1/2}$. An example of a stochastic process whose variance grows sublinearly in time is fractional Brownian motion with appropriately chosen Hurst parameter.

Proposition 3.1 *(A Poisson Equation)*
Let $\mathcal{L}_x$ be the infinitesimal generator of the X process. Let $G(x,\theta) \in C^{\alpha,2}(\mathcal{X},\mathbb{R}^n)$ which satisfies

$$\int_{\mathcal{X}} G(x,\theta)\pi(dx) = 0.$$

and for some positive constants M and q, and

$$|G(x,\theta)| + \left|\frac{\partial}{\partial\theta}G(x,\theta)\right| + \left|\frac{\partial^2}{\partial\theta^2}G(x,\theta)\right| \le M(1+|x|^q).$$

Then the Poisson equation

$$\mathcal{L}_x u(x,\theta) = G(x,\theta), \quad \int_{\mathcal{X}} u(x,\theta)\pi(dx) = 0$$

has a unique solution that satisfies $u(x,\cdot) \in C^2$ for every $x \in \mathcal{X}$, $\partial_\theta^2 u \in C(\mathcal{X}\times\mathbb{R}^n)$ and there exist positive constants K and p such that

$$|u(x,\theta)| + \left|\frac{\partial}{\partial\theta}u(x,\theta)\right| + \left|\frac{\partial^2}{\partial\theta^2}u(x,\theta)\right| \le K(1+|x|^p).$$

4. Stochastic Gradient Descent Algorithm for American Option

Machine learing in finance has received recent attention, see Dixon *et al.* (2020). We study the SGDCT algorithm for American option. We compare it with Longstaff-Schwartz method. Longstaff-Schwartz developed an algorithm for the solution of a discrete time version of the a class of free boundary. Their algorithm, commonly called Longstaff-Schwartz Regression based method, uses dynamic programming and approximates the solution using a separate function approximator at each discrete time, typically a linear combination of basis functions.

Given a continuous stream of data, stochastic gradient descent in continuous time (SGDCT) can estimate unknown parameters or functions in stochastic differential equation (SDE) models for stocks, bonds, interest rates, and financial derivatives. High dimensional American option has been a long standing computational challenge in finance with traditional methods like the finite difference. SGDCT can accurately solve American options even in 100 dimensions.

Batch optimization for statistical estimation of continuous-time models can be impractical for large data sets where observations occur over a long time period. Batch optimization takes a sequence of descent steps for the model error for the entire observed path.

SGDCT provides a computationally efficient method for statistical learning over long time periods and for complex models. SGDCT continuously follows a descent direction along the path of the observation. Parameters are updated in continuous time, with the parameter updates θ_t satisfying an SDE.

Numerical analysis of SGDCT in model estimation of the drift and volatility functions of two common financial models like the O-U process and CIR process is studied. One has to simulate using Euler scheme, a single path of X_t for given θ^* and simultaneously solve for the path of θ_t.

Sirignano and Spiliopoulos (2018) studied deep learning algorithm or "Deep Galerkin Method" (DGM) which is Galerkin method with neural network. Neural network is trained on the batches of randomly sampled time and space points. Deep Galerkin method uses a deep neural network instead of basis functions. The deep neural network is trained to satisfy the differential operator, initial condition, and the boundary conditions using stochastic gradient descent at randomly sampled spatial points. By randomly sampling spatial points, we avoid the need to form a mesh (which is infeasible in higher dimensions) and instead convert the PDE problem to a machine learning problem. DGM is natural merger of Galerkin methods and machine learning.

Sirignano and Spiliopoulos (2017) obtained central limit theorem for the SGDCT estimator.

An American option is a financial derivative which the owner can choose to exercise at any time $t \in [0, T]$. If the owner exercises the option, they receive the payoff $g(X_t)$ where X_t is the prices of the underlying stocks. If the owner does not choose to exercise the option, they receive the payoff $g(X_T)$ at the final time T. The value (or price) of the American option at time t is $u(t, X_t)$ which satisfies a free boundary PDE:

$$\frac{\partial u}{\partial t}(t,x) + \mu(x)\frac{\partial u}{\partial x}(t,x) + \frac{1}{2}\sum_{i,j=1}^{n} \rho_{i,j}\sigma(x_j)\frac{\partial^2 u}{\partial x_i \partial x_j}(t,x) - ru(t,x) = 0, \tag{4.1}$$

$$\forall\{(t,x) : u(t,x) > g(x)\},$$

$$u(t,x) \geq g(x) \quad \forall (t,x),$$

$$u(t,x) \in C^1(\mathbb{R}_+ \times \mathbb{R}^d), \quad \forall\{(t,x) : u(t,x) = g(x)\},$$

$$u(T,x) = g(x), \quad \forall x.$$

The free boundary set is $F = \{(t,x) : u(t,x) = g(x)\}$. The value function $u(t,x)$ satisfies a PDE "above" the free boundary set F and $u(t,x)$ equals the function $g(x)$ "below" the free boundary set F. The free boundary set F is approximated using the current parameter estimate θ_n.

The SGDCT Algorithm

First, we recall the Q-learning algorithm: The Q-learning algorithm uses stochastic gradient descent to minimize an approximation to the discrete time HJB equation. Consider the Q-learning algorithm to estimate the value function

$$V(x) = E\left[\int_0^\infty e^{-\gamma t} r(X_t)dt \;\middle|\; X_0 = x\right], \quad X_t = x + W_t \tag{4.2}$$

where $\gamma > 0$ is a discount factor and $r(x)$ is a reward function. The function $Q(x,\theta)$ is an approximation for the value function $V(x)$. The traditional approach is to discretize the dynamics of $V(x)$ and apply a stochastic gradient descent update to the objective function:

$$E\left[\left(r(X_t)\Delta + e^{-\gamma\Delta}E[Q(X_{t+\Delta};\theta)|X_t] - Q(X_t;\theta)\right)^2\right]. \tag{4.3}$$

The result is the stochastic gradient descent algorithm:

$$\begin{aligned}\theta_{t+\Delta} = \theta_t - \frac{\alpha_t}{\Delta}&\left(e^{-\gamma\Delta}E[Q_\theta(X_{t+\Delta};\theta_t)|X_t] - Q_\theta(X_t;\theta_t))\right)\\ &\times\left(r(X_t)\Delta + e^{-\gamma\Delta}E[Q(X_{t+\Delta};\theta_t)|X_t] - Q(X_t;\theta_t)\right).\end{aligned} \tag{4.4}$$

The learning rate is Δ^{-1}. The Q-learning algorithm has a major computational issue. The expectation $E[Q_\theta(X_{t+\Delta};\theta_t)|X_t]$ is challenging to calculate if the process X_t is high dimensional. To circumvent this situation, Q-learning algorithm ignores the inner expectation leading to

$$\theta_{t+\Delta} = \theta_t - \frac{\alpha_t}{\Delta}(e^{-\gamma\Delta}Q_\theta(X_{t+\Delta};\theta_t) - Q_\theta(X_t;\theta_t)(r(X_t)\Delta + e^{-\gamma\Delta}Q(X_{t+\Delta};\theta_t) - Q(X_t;\theta_t)). \tag{4.4}$$

Although computationally efficient, the Q-learning algorithm is biased. The SGDCT algorithm can be directly derived by letting $\Delta \to 0$ and using Itô formula:

$$d\theta_t = -\alpha_t\left(\frac{1}{2}Q_{\theta xx}(X_t;\theta_t) - \gamma Q_\theta(X_t;\theta_t)\right)\left(r(X_t) + \frac{1}{2}Q_{xx}(X_t;\theta_t) - \gamma Q(X_t;\theta_t)\right)dt. \tag{4.5}$$

Furthermore, when $\Delta \to 0$, the Q-learning algorithm blows up.

SGDCT Algorithm for American Option

Let $X_t \in \mathbb{R}^d$ be the prices of d stocks. The maturity time is T and the payoff function is $g(x) : \mathbb{R}^d \to \mathbb{R}$. The stock price dynamics and the value functions are given by

$$dX_t^i = \mu(X_t^i)dt + \sigma(X_t^i)dW_t^i, \quad i = 1,2,\ldots,d \tag{4.6}$$

$$V_{t,x} = \sup_{\tau\geq t} E[e^{-r(\tau\wedge T)}g(X_{\tau\wedge T})|X_t = x] \tag{4.7}$$

where $W_t \in \mathbb{R}^d$ is a Brownian motion. The distribution of W_t is specified by $Var(W_t^i) = t, \quad i = 1, 2, \ldots, d$ and $Corr(W_t^i, W_t^j) = \rho_{i,j} dt$ for $i \neq j$. The price of the American option is $V_{0,x}$.

SGDCT for American option is given by

$$\theta_{t\wedge T}^{n+1} = \theta_0^n - \int_0^{\tau\wedge T} \alpha_t^{n+1} \left(\frac{\partial}{\partial t} Q_\theta(t, X_t; \theta_t^{n+1}) + \mathcal{L}_x Q_\theta(t, X_t; \theta_t^{n+1}) - r Q_\theta(t, X_t; \theta_n^{n+1}) \right)$$
$$\times \left(\frac{\partial}{\partial t} Q(t, X_t; \theta_t^{n+1}) + \mathcal{L}_x Q(t, X_t; \theta_t^{n+1}) - r Q(t, X_t; \theta_t^{n+1}) \right) dt$$
$$+ \alpha_{\tau\wedge T}^{n+1}\, Q_\theta\left(\tau \wedge T, X_{\tau\wedge T}; \theta_{\tau\wedge T}^{n+1}\right) \left(g(X_{\tau\wedge T}) - Q(\tau \wedge T, X_{\tau\wedge T}; \theta_{\tau\wedge T}^{n+1})\right), \tag{4.8}$$
$$\tau := \inf\{t \geq 0 : Q(t, X_t; \theta_t^{n+1}) < g(X_t)\}, \quad X_0 \sim \nu(dx). \tag{4.9}$$

The function $Q(x, \theta)$ is an approximation of the value function. The parameter θ must be estimated. Here $\mathcal{L}_x$ is the infinitesimal generator of the X process. The algorithm is run for many iterations $n = 0, 1, 2, \ldots$ until convergence.

The Longstaff-Schwarz algorithm works well in low dimension, but in high dimension the convergence is slow. In high dimension, SGD algorithm works very well.

Sirignano and Spiliopoulos (2017) implemented the American option in 100 dimensions and showed the accuracy of the SGD algorithm for Bachelier model and Black-Scholes model.

5. Berry-Esseen Inequality of Stochastic Gradient Descent Algorithm for American Option

We study the Berry-Esseen inequality for SGDCT algorithm for American option. We compare it with Longstaff-Schwartz method. We will use anticipative stochastic integral, Duhamel's principle for the stochastic gradient descent algorithm.

Stochastic gradient descent in continuous time (SGDCT) provides a computationally efficient method for the statistical learning of continuous-time models. SGDCT can estimate unknown parameters and functions in SDE models. SGDCT algorithm follows a descent direction along a continuous stream of data. The parameter updates occur in continuous time and satisfy a stochastic differential equation (SDE). The authors analyze the asymptotic convergence rate by proving a central limit theorem. An L^p convergence rate is also proven.

The vast majority of statistical learning, machine learning and stochastic gradient descent literature address discrete time algorithm. This section analyzes statistical learning in continuous time.

Statistical estimation in SDEs have been studied using entire observed path of X, i.e., batch optimization. MLE can be calculated via batch optimization. Maximum likelihood based on the entire observation path of X has been extensively studied, see Bishwal (2008).

SGDCT can be used to solve continuous-time optimization problem such as American options. The value function is approximated by a parametric function and the parameter is estimated by

SGDCT algorithm. Recall that Longstaff-Schwartz estimated the parameter by least squares method. One could discretize the dynamics and then use the Q-learning algorithm. The Q-learning algorithm is biased while SGDCT algorithm is unbiased.

Consider the SDE

$$dX_t = f^*(X_t)dt + \sigma dW_t, t \geq 0$$

where $f^*(x)$ is an unknown function. The goal is to estimate $f(x,\theta)$ from continuous observations of $(X_t)_{t\geq 0}$. The function may be convex or non-convex.

The SGD update satisfies

$$d\theta_t = \frac{\alpha_t}{\sigma^2} f'(X_t,\theta)dX_t - \frac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t)dt, t \geq 0$$

where α_t is the *learning rate*. For example, α_t could be $\frac{C_\alpha}{C_0+t}$. Assume that θ_0 is initialized according to some distribution with compact support.

The parameter update can be used both for statistical estimation given previously observed data as well as online learning, i.e, statistical estimation in real time as data becomes available.

Define the objective function

$$g(x,\theta) = \frac{1}{2}\|f(x,\theta) - f^*(x)\|^2 = \frac{1}{2\sigma^2}\langle(f(x,\theta) - f^*(x)\rangle)^2$$

which measures the distance between the model $f(x,\theta)$ and true dynamics $f^*(x)$ for a specific x. This is a minimum distance type estimator.

We assume that X_t is ergodic and it has some well behaved $\pi(dx)$ as its unique invariant distribution. Let the average be denoted by

$$\bar{g}(\theta) = \int g(x,\theta)\pi(dx)$$

where $\pi(dx)$ is the invariant measure of X_t when it is ergodic which is the natural objective function. This is an weighted average of the distance between $f(x,\theta)$ and $f^*(x)$ where the weights are given by $\pi(dx)$, which is the distribution that X_t tends to when t become large. The distance $g(x,\theta)$ is decreased by moving θ in the descent direction $-g'(x,\theta)$ which motivates the algorithm

$$\begin{aligned}
d\theta_t &= -\alpha_t g'(X_t,\theta_t)dt \\
&= \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)(f^*(X_t) - f(X_t,\theta))dt \\
&= -\alpha_t g'(x,\theta_t)dt + \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t))\sigma dW_t \\
&= -\alpha_t \bar{g}'(x,\theta_t)dt - \alpha_t(g'(x,\theta_t) - \bar{g}'(x,\theta_t))dt \\
&\quad + \tfrac{\alpha_t}{\sigma^2} f'(X_t,\theta_t)f(X_t,\theta_t))\sigma dW_t \\
&=: I_1 + I_2 + I_3
\end{aligned}$$

where I_1= Descent term, I_2 = fluctuation term, I_3= Noise term.

If α_t decays with time, e.g., $\alpha_t = \frac{C_\alpha}{C_0+t}$, the descent term I_1 will dominate the fluctuation term and the noise term for large t. Then θ_t, will converge to a local minimum of $\bar{g}(\theta)$. Sirignano and Spiliopoulos(2017) proved that θ_t converges to a critical point of the objective function $\bar{g}(\theta)$: $|\bar{g}'(\theta_t)| \to 0$ almost surely as $t \to \infty$.

Since $\lim_{t\to\infty}\alpha_t = 0$, the descent term $\alpha_t\bar{g}'(\theta_t) \to 0$. Descent term converges to zero as $t\to\infty$. Sirignano and Spiliopoulos (2018) proved the rate at which θ_t converges to zero. They obtained a central limit theorem (CLT) and an L^p convergence rate.

Sirignano and Spiliopoulos (2018) derived a stochastic integral to represent the $\sqrt{t}(\theta_t-\theta^*)$ using Duhamel's principle and the fundamental solution of the random ODE $d\Psi_{t,s} = -\alpha_t\bar{g}'(\theta_t)$ $\Psi_{t,s}dt$ where $\tilde{\theta}_t$ lies on a line connecting θ^* and θ_t. The integrand of this stochastic integral includes the fluctuation term and the noise term as well as Ψ_t and is *anticipative*. Also since $f(x,\theta)$ is allowed to grow with θ, hence the fluctuations as well as other terms can grow with θ. Hence they prove an a priori stability estimate for $|\theta_t|$. Proving central limit theorem for non-convex $\bar{g}(\theta)$ since the convergence speed of θ_t can arbitrarily slow in certain regions, and the gradient can even point away from the global minimum θ^*. To address this, we consider the stochastic integral after the time τ_δ which is defined as the final time θ_t enters a neighborhood of θ^*. However, τ_δ is anticipative, i.e., is not a stopping time, therefore careful analysis is required to study the behavior of the stochastic integral.

Let $Y_t := \theta_t - \theta^*$ be the error term. It satisfies

$$dY_t = -\alpha_t\Delta\bar{g}(\theta_t^1)Y_t dt + \alpha_t(\bar{g}'(\theta_t) - g'(X_t,\theta_t))dt + \alpha_t f'(X_t,\theta_t)dW_t.$$

$$E|Y_t|^2 \le Kt^{-1}, \quad E|Y_t|^p \le Kt^{-p/2}.$$

The stochastic integral

$$\sqrt{t}\int_1^t \alpha_s^2\Phi_{t,s}\zeta(X_s,\theta_s)dW_s \to 0$$

in probability as $t\to\infty$ where $\zeta(x,\theta)$ is a function that can grow at most polynomially in x and θ.

For the analysis of fluctuation term, the proofs use Poisson PDE for ergodicity, see Section 6. The central limit theorem for non-convex $\bar{g}(\theta)$ is challenging since the convergence speed of θ_t can become arbitrarily slow in certain regions and the gradient can even point away from the global minimum θ^*. *Interalia,* Sirignano and Spiliopoulos (2018) prove convergence to zero of multidimensional stochastic integrals. The proof requires the analysis of stochastic integral with anticipative integrands, which is challenging since standard approaches like Itô isometry can not be directly applied. It is related to online maximum likelihood filtering and identification.

We remark that $t^{-1/2}$ is the fastest possible convergence rate given that the noise is Brownian motion. This is due to the quadratic variation of Brownian motion growing linearly in t. With other noises whose variances grow sublinearly in time, one could allow for faster rate of convergence than $t^{-1/2}$. An example of a stochastic process whose variance grows sublinearly in time is fractional Brownian motion with appropriately chosen Hurst parameter discussed in section 1.

In this section we investigate the rate of weak convergence to normality of the update θ_t. We assume the following conditions:

(A1) The diffusion is nondegenerate and $\lim_{|x|\to\infty} f^*(x)\cdot x = -\infty$.

(A2) $g'(x,\cdot)\in C^2(\mathbb{R})$ for all x.

(A3) The function $f^*(x) \in C^{2+\alpha}(\mathcal{X})$, that is, it has two derivatives in x, with all partial derivatives being Hölder continuous, with exponent α, with respect to x.

(A4) SGD-SDE equation is well-posed.

(A5) There exists a constant $R < \infty$ and almost everywhere positive function $\kappa(x)$ such that $\langle -g'(x,\theta), \theta/|\theta|\rangle \leq -\kappa(x)|\theta|$ for $|\theta| \geq R$.

(A6) Consider the function $\tau(x,\theta) = \langle f'^2(x,\theta), \theta/|\theta|\rangle$. Then there exists a function $\lambda(x)$ growing not faster than polynomially in $|x|$ such that for any $x, \theta_1, \theta_2 \in \mathbb{R}$, $|\tau(x,\theta_1) - \tau(x,\theta_2)| \leq |\lambda(x)|\rho(|\theta_1 - \theta_2|)$ where $\rho(u)$ is an increasing function on $[0,\infty)$ with $\rho(0) = 0$ and $\int_{u>0} \rho^{-2}(u)du = \infty$.

(A7) The learning rate is $\frac{C_\alpha}{C_0+t}$ where $C_\alpha > 0$ and C_0 are constants.

(A8) $f_\theta^{(i)}(x,\theta) \leq K(1+|x|^q + |\theta|^{(2-i)\vee 0})$, $\quad i = 0,1,2$ for some finite constants $K, q < \infty$.

(A9) $\bar{g}(\theta)$ is strongly convex with constant C.

(A10) $CC_\alpha > 1$.

(A11) $\bar{g}(\theta) \in C^3$ and $|\bar{g}^{(i)}(\theta)| \leq K(1+|\theta|^{4-i})$ for $i = 0,1,2,3$ and some finite constant $K < \infty$.

The following theorem gives the rate of convergence to normal distribution of the SGDCT estimator:

Theorem 5.1 *Under (A1) – (A11) and (B1) – (B10), we have as $t \to \infty$*

$$\sup_{x\in\mathbb{R}} \left| P\left(\sqrt{\frac{t}{J(\theta^*)}}(\theta_t - \theta^*) \leq x\right) - \Phi(x)\right| \leq Ct^{-1/2}$$

where

$$J(\theta^*) = C_\alpha^2 \int_0^\infty e^{-2s(C_\alpha g'(\theta^*)-1)} \bar{h}(\theta^*) ds, \quad \bar{h}(\theta) = \int h(x,\theta)\pi(dx),$$

$$h(x,\theta) = \left(\frac{1}{\sigma^2} f'(x,\theta) - \dot{v}(x,\theta)\right)^2 \sigma^2$$

and $v(x,\theta)$ is the solution to the Poisson equation with

$$H(x,\theta) = g'(x,\theta) - \bar{g}'(\theta).$$

Proof: Using second order Taylor expansion

$$\bar{g}'(\theta_t) = \bar{g}'(\theta_t) + \bar{g}''(\theta^*)(\theta_t - \theta^*) + \frac{1}{2}\frac{\partial^3}{\partial\theta^3}\bar{g}(\theta_t^1)(\theta_t - \theta^*)^2.$$

The error term satisfies the SDE

$$d(\theta_t - \theta^*) = -\alpha_t \bar{g}'(\theta_t^1)dt - \frac{\alpha_t}{2}\frac{\partial^3}{\partial\theta^3}\bar{g}(\theta_t^1)dt + \alpha_t(\bar{g}'(\theta_t) - g'(X_t,\theta_t))dt + \alpha_t f'(X_t,\theta_t)dW_t.$$

Let $Y_t := \theta_t - \theta^*$. Then Y_t satisfies the SDE

$$dY_t = -\alpha_t \bar{g}'(\theta_t^1) Y_t dt - \frac{\alpha_t}{2}\frac{\partial^3}{\partial\theta^3}\bar{g}(\theta_t^1) Y_t^2 dt + \alpha_t(\bar{g}'(\theta_t) - g'(X_t,\theta_t))dt + \alpha_t f'(X_t,\theta_t)dW_t.$$

Let $\Phi_{t,s}$ be the fundamental solution satisfying

$$d\Phi_{t,s} = -\alpha_t \bar{g}'(\theta^*)\Phi_{t,s}dt, \Phi_{s,s} = 1$$

Y_t can be written in terms of $\Phi_{t,s}$:

$$Y_t = \Phi_{t,1}Y_1 - \frac{1}{2}\int_1^t \Phi_{t,s}\alpha_s \bar{g}^{(3)}(\theta_s)Y_s^2 ds + \int_1^t \Phi_{t,s}(\bar{g}'(\theta_s) - g'(X_s,\theta_s))ds + \int_1^t \Phi_{t,s}\alpha_s f'(X_s,\theta_s)dW_s$$
$$=: \Gamma_t^1 + \Gamma_t^2 + \Gamma_t^3 + \Gamma_t^4.$$

The problem is the weak convergence of *anticipative stochastic integral* which has not been studied much in the literature since standard approach such as Itô isometry can not be directly applied. We use Malliavin calculus approach as in Bishwal (2010b).

We show the rate at which the stochastic integral converges to the normal distribution $\mathcal{N}$,

$$\sqrt{t}\int_1^t \alpha_s \Phi_{t,s}\zeta(X_s,\theta_s)dW_s \to \mathcal{N}$$

in distribution as $t \to \infty$. By using large deviations, we show the rate at which $\sqrt{t}(\Gamma_t^1 + \Gamma_t^2 + \Gamma_t^3) \to 0$. By using large deviations, we show the rate at which

$$t\int_1^t \alpha_s^2 \Phi_{t,s}^2 \zeta^2(X_s,\theta_s)ds \to J(\theta^*)$$

in probability as $t \to \infty$. Combining all these in Y_t and the using squeezing technique in Chapter-1 in Bishwal (2008), we obtain the result.

Remark Bishwal (2011c) studied parameter estimation in interacting diffusions based on continuous and discrete sampling. The idea was used in Giesecke *et al.* (2020) for inference in large financial systems.

Next we focus on Monte Carlo method. Let $\hat{\theta}_{n,t}$ be the Monte Carlo estimate of θ_t based on n independent replications of the sample path, i.e.,

$$\hat{\theta}_{n,t} = \frac{1}{n}\sum_{i=1}^n \theta_{i,t}.$$

Theorem 5.2 *Under (A1) – (A11) and (B1) – (B10), we have as* $n \to \infty$

$$\sup_{x\in\mathbb{R}}\left| P\left(\sqrt{\frac{n}{J(\theta^*)}}(\hat{\theta}_{n,t} - \theta^*) \le x\right) - \Phi(x)\right| \le Cn^{-1/2}$$

where

$$J(\theta^*) = C_\alpha^2 \int_0^\infty e^{-2s(C_\alpha g'(\theta^*)-1)}\bar{h}(\theta^*)ds, \quad \bar{h}(\theta) = \int h(x,\theta)\pi(dx),$$

$$h(x,\theta) = \left(\frac{1}{\sigma^2}f'(x,\theta) - \dot{v}(x,\theta)\right)^2 \sigma^2$$

and $v(x,\theta)$ *is the solution to the Poisson equation with*

$$H(x,\theta) = g'(x,\theta) - \bar{g}'(\theta).$$

Proof: Berry-Esseen theorem for independent random variables (see Petov (1995)) along with anticipative Girsanov theorem (Proposition 1.3) gives the result. Details are omitted. □

References

[1] M. Arcones, On the law of logarithm for Gaussian processes, J. Theor. Prob. 8 (1995) 877-903.

[2] E. Alos, O. Mazet, D. Nualart, Stochastic calculus with respect to fractional Brownian motion with Hurst parameter less than $\frac{1}{2}$, Stoch. Proc. Appl. 86 (2000) 121-139.

[3] E. Alos, O. Mazet, D. Nualart, Stochastic calculus with respect to Gaussian processes, Ann. Prob. 29 (2001) 766-801.

[4] I.V. Basawa, D.J. Scott, Asymptotic Optimal Inference for Non-ergodic Models, Lecture Notes in Statistics 17, Springer, Berlin, (1983).

[5] M. El Machkouri, K. Es-Sebaiy, Y. Ouknine, Least squares estimator for non-ergodic Ornstein-Uhlebeck processes driven by Gaussian processes, J. Korean Stat. Soc. 45 (2016) 329-341.

[6] J.P.N. Bishwal, Parameter Estimation in Stochastic Differential Equations, Lecture Notes in Mathematics 1923, Springer-Verlag, (2008).

[7] J.P.N. Bishwal, Maximum likelihood estimation in Skorohod stochastic differential equations, Proc. Amer. Math. Soc. 138 (2010) 1471-1478.

[8] J.P.N. Bishwal, Maximum quasi likelihood estimation in fractional Levy stochastic volatility model, J. Math. Finance 1 (2011a) 58-62.

[9] J.P.N. Bishwal, Sufficiency and Rao-Blackwellization of Vasicek model, Theory Stoch. Proc. 17 (2011b) 12-15.

[10] J.P.N. Bishwal, Estimation in interacting diffusions: continuous and discrete sampling, Appl. Math. 2 (2011c) 1154-1158.

[11] B. Buchman, C. Kluppelberg, Fractional integral equations and state space transform, Bernoulli 12 (2006) 431-456.

[12] V. Borkar, A. Bagchi, Parameter estimation in continuous time stochastic processes, Stochastics 8 (1982) 193-212.

[13] P.E. Caines, D. Levanony, Performance monitored continuous time LQ stochastic adaptive control, IEEE Proceedings of the 33rd Conference on Decision and Control, San Antonio, TX, (1993).

[14] L. Coutin, L. Decreusefond, Stochastic differential equations driven by a fractional Brownian motion, Tech. Report 97C004, ENST Paris, (1997).

[15] L. Coutin, L. Decreusefond, Abstract nonlinear filtering in the presence of fractional Brownian motion, Ann. Appl. Prob. 9 (1999) 1058-1090.

[16] W. Dai, C.C. Heyde, Itô formula with respect to fractional Brownian motion and its applications, J. Appl. Math. Stoch. Anal. 9 (1996) 439-448.

[17] M. Dixon, I. Halperin, P. Bilokon, Machine Learning in Finance: From Theory to Practice, Springer Nature Switzerland AG, (2020).

[18] T.E. Duncan, Y. Hu, B. Pasik-Duncan, Stochastic calculus for fractional Brownian motion I, SIAM J. Control. Optim. 38 (2000) 582-612.

[19] P.D. Feigin, Some comments concerning curious singularity, J. Appl. Prob. 16 (1979) 440-444.

[20] L. Gerencser, V. Prokaj, Recursive identification of continuous time linear stochastic systems - An off-line approximation, Proceedings of the 19th International Symposium of Mathematical Theory of Network and Systems-MTNS 2010 Budapest, Hungary, (2010).

[21] K. Giesecke, G. Schwenkler, J.A. Sirignano, Inference for large financial systems, Math. Finance 30 (2020) 3-46.

[22] G. Gripenberg, I. Norros, On the prediction of fractional Brownian motion, J. Appl. Prob. 33 (1996) 400-410.

[23] E. Igloi, G. Terdik, Long-range dependence through Gamma mixed Ornstein-Uhlenbeck processes, Electron. J. Prob. 4 (1999) 1-33.

[24] M.L. Kleptsyna, P.E. Kloeden, V.V. Anh, Linear filtering with fractional Brownian motion, Stoch. Anal. Appl. 16 (1998a) 907-914.

[25] M.L. Kleptsyna, P.E. Kloeden, V.V. Anh, Nonlinear filtering with fractional Brownian motion, Probl. Inf. Transmission 34 (1998b) 65-76.

[26] M.L. Kleptsyna, P.E. Kloeden, V.V. Anh, Existence and uniqueness theorem for fBm stochastic differential equqtions, Probl. Inf. Transmission 34 (1998c) 51-61.

[27] M.L. Kleptsyna, A. Le Breton, M.C. Roubaud, An elementary approach to filtering in systems with fractional Brownian motion observation noise, In: Probability Theory and Mathematical Statistics, B. Grigelionis (Eds.), VSP/TEV, 373-392, (1998).

[28] M.L. Kleptsyna, A. Le Breton, Statistical analysis of the fractional Ornstein-Uhlenbeck type process, Stat. Inference Stoch. Proc. 5 (2002) 229-248.

[29] M.L. Kleptsyna, A. Le Breton, M.C. Roubaud, General approach to filtering with fractional Brownian noises- application to linear systems, Stoch. Stoch. Rep. 71 (2000) 119-140.

[30] A.N. Kolmogorov, Wiener skewline and other interesting curves in Hilbert space, Doklady Akad. Nauk 26 (1940) 115-118.

[31] H. Kunita, Asymptotic behavior of the nonlinear filtering errors of Markov processes, J. Multivar. Anal. 1 (1971) 365-393.

[32] H. Kunita, Stochastic Flows and Stochastic Differential Equations, Cambridge Univ. Press, New York, (1990).

[33] S. Kusuoka, The nonlinear transformation of Gaussian measures on Banach space and its absolute continuity. I J. Fac. Sci. Univ. Tokyo, Section IA, Math. 29 (1982) 567-597.

[34] Yu. A. Kutoyants, Parameter estimation for Stochastic Processes (Translated and edited by B.L.S. Prakasa Rao) Heldermann-Verlag, Berlin, (1984).

[35] Yu. A. Kutoyants, Identification of dynamical systems with small noise, Kluwer, Dordrecht, (1994).

[36] N. Lazrieva, T. Toronjadze, Recursive estimation procedures for one-dimensional parameter of statistical models associated with semimartingales, Trans. A. Razmadze Math. Inst. 171 (2017) 57-75.

[37] D. Levanony, Large deviations of consistent parameter estimates in diffusions, IEEE Proceedings of the 33rd Conference on Decision and Control, Lake Beuna Vista, FL, (1994).

[38] D. Levanony, Conditional tail probabilities in continuous-time martingale LLN with application to paramter estimation in diffusions, Stoch. Proc. Appl. 51 (1994) 117-134.

[39] D. Levanony, A. Shwartz, O. Zeitouni, Uniform decay and equicontinuity for normalized, parameter dependent, Ito integrals, Stoch. Stoch. Rep. 43 (1993) 9-28.

[40] D. Levanony, A. Shwartz, O. Zeitouni, Recursive identification in continuous-time stochastic processes, Stoch. Proc. Appl. 49 (1994) 245-275.

[41] P. Levy, Processus stochastiques et movement Brownien, Paris, (1948).

[42] S.J. Lin, Stochastic analysis of fractional Brownian motions, Stoch. Stoch. Rep. 55 (1995) 121-140.

[43] T.J. Lyons, Differential equations driven by rough signals, Rev. Mat. Iberoamericana 14 (1998) 215-310.

[44] T.J. Lyons, T.S. Zhang, Decomposition of Dirichlet processes and its application, Ann. Prob. 22 (1994) 494-524.

[45] T.J. Lyons, N. Victoir, An extension theorem to rough paths, Ann. Inst. Henri Poincare AN 24 (2007) 835-847.

[46] B. Mandelbrot, J.W. Van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Rev. 10 (1968) 422-437.

[47] Y. Mishura, Stochastic Calculus for Fractional Brownian Motion and Related Processes, Lecture Notes in Mathematics, 1929, Springer-Verlag, Berlin, (2008).

[48] I. Norros, E. Valkeila, J. Virtamo, An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motion, Bernoulli 5 (1999) 571-587.

[49] E. Pardoux, A.Y. Veretennikov, On Poisson equation and diffusion approximation I, Ann. Prob. 29 (2001) 1061-1085.

[50] E. Pardoux, A.Y. Veretennikov, On Poisson equation and diffusion approximation 3, Ann. Prob. 31 (2003) 1166-1192.

[51] R.F. Peltier, J. Levy Vehel, A new method for estimating the parameter of fractional Brownian motion, INRIA Res. Rep. 2396 (1994) 1-32.

[52] V.V. Petrov, Limit Theorems of Probability Theory, Oxford University Press, Oxford, (1995).

[53] V. Pipiras, M.S. Taqqu, Integration questions related to fractional Brownian motion, Prob. Theory Rel. Fields 118 (2000) 251-291.

[54] A.A. Ruzmaikina, Stochastic calculus for fractional Brownian motion, J. Stat. Phys. 100 (2000) 1049-1069.

[55] G. Samorodnitsky, M.S. Taqqu, Stable Non-Gaussian Randaom Processes-Stochastic Models with Infinite Variance, Chapman and Hall, London, (1994).

[56] B. Simon, Trace Ideals and Their Applications, London Mathematical Society, Lecture Note Series, 35, Cambridge University Press, Cambridge-New York, (1979).

[57] J. Sirignano, K. Spiliopoulos, Stochastic gradient descent in continuous time, SIAM J. Financial Math. 8 (2017) 933-961.

[58] J. Sirignano, K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, J. Comput. Phys. 375 (2018) 1339-1364.

[59] J. Sirignano, K. Spiliopoulos, Stochastic gradient descent in continuous time: A central limit theorem, Stoch. Syst. 10 (2020) 124-151.

[60] S. Surace, J. Pfister, Online maximum lilkelihood estimation of partially observed diffusion process, IEEE Trans. Autom. Control, 64 (2019) 2814-2829.

[61] M.S. Taqqu, Weak convergence to fractional Brownian motion and to the Rosenblatt process, Z. Wahr. verw. Gebiete 31 (1975) 287-302.

[62] C.A. Tudor, Analysis of Variations for Self-similar Processes: A Stochastic Calculus Approach, Springer, New York, (2013).

[63] L.C. Young, An inequality of the Hölder type connected with Stieltjes integration, Acta Math. 67 (1936) 251-282.

[64] M. Zahle, Integration with respect to fractal functions and stochastic calculus, Prob. Theory. Rel. Filelds 111 (1998) 333-374.

[65] M. Zahle, Integration with respect to fractal functions and stochastic calculus, Prob. Theory. Rel. Filelds 111 (1998) 333-374.