# PENALIZED NONPARAMETRIC MEAN SQUARE ESTIMATION OF THE COEFFICIENTS OF DIFFUSION PROCESSES.

F. COMTE[1*], V. GENON-CATALOT[1], AND Y. ROZENHOLC[1]

ABSTRACT. We consider a one-dimensional diffusion process $(X_t)$ which is observed at $n+1$ discrete times with regular sampling interval $\Delta$. Assuming that $(X_t)$ is strictly stationary, we propose nonparametric estimators of the drift and diffusion coefficients obtained by a penalized least square approach. Our estimators belong to a finite dimensional function space whose dimension is selected by a data-driven method. We provide non asymptotic risk bounds for the estimators. When the sampling interval tends to zero while the number of observations and the length of the observation time interval tend to infinity, we show that our estimators reach the minimax optimal rates of convergence. Numerical results based on exact simulations of diffusion processes are given for several examples of models and enlight the qualities of our estimation algorithms.

This version: November 2005

**Keywords.** Adaptive estimation. Diffusion processes. Discrete time observations. Drift and diffusion coefficients. Mean square estimator. Model selection. Penalized contrast. Retrospective simulation

Running title: Penalized estimation of drift and diffusion.

* *Corresponding author*

[1] Fabienne Comte, Valentine Genon-Catalot, Yves Rozenholc,
Université Paris V-René Descartes,
MAP5, UMR CNRS 8145.
45, rue des Saint-Pères,
75270 Paris Cedex 06, FRANCE.

fabienne.comte@univ-paris5.fr
Valentine.Genon-Catalot@math-info.univ-paris5.fr
yves.rozenholc@math-info.univ-paris5.fr

## 1. Introduction

In this paper, we consider the following problem. Let $(X_t)_{t\geq 0}$ be a one-dimensional diffusion process with dynamics described by the following stochastic differential equation:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad t \geq 0, \quad X_0 = \eta \tag{1}$$

where $(W_t)$ is a standard Brownian motion and $\eta$ is a random variable independent of $(W_t)$. Assuming that the process is strictly stationary (and ergodic), and that a discrete observation $(X_{k\Delta})_{1\leq k\leq n+1}$ of the sample path is available, we want to build nonparametric estimators of the drift function $b$ and the (square of the) diffusion coefficient $\sigma^2$.

Our aim is twofold: Construct estimators that have optimal asymptotic properties and that can be implemented through feasible algorithms. Our asymptotic framework is such that the sampling interval $\Delta = \Delta_n$ tends to zero while $n\Delta_n$ tends to infinity as $n$ tends to infinity. Nevertheless, the risk bounds obtained below are non asymptotic in the sense that they are explicitly given as functions of $\Delta$ or $1/(n\Delta)$ and fixed constants.

Nonparametric estimation of the coefficients of diffusion processes has been widely investigated in the last decades. The first estimators that have been proposed and studied are based on a continuous time observation of the sample path. Asymptotic results are given for ergodic models as the length of the observation time interval tends to infinity: See for instance the reference paper by Banon (1978), followed by more recent works of Prakasa Rao (1999), Spokoiny (2000), Kutoyants (2004) or Dalalyan (2005).

Then discrete sampling of observations has been considered, with different asymptotic frameworks, implying different statistical strategies. It is now classical to distinguish between low-frequency and high-frequency data. In the former case, observations are taken at regularly spaced instants with fixed sampling interval $\Delta$ and the asymptotic framework is that the number of observations tends to infinity. Then, only ergodic models are usually considered. Parametric estimation in this context has been studied by Bibby and Sørensen (1995), Kessler and Sørensen (1999), see also Bibby et al. (2002). A nonparametric approach using spectral methods is investigated in Gobet et al. (2004), where non standard nonparametric rates are exhibited.

In high-frequency data, the sampling interval $\Delta = \Delta_n$ between two successive observations is assumed to tend to zero as the number of observations $n$ tends to infinity. Taking $\Delta_n = 1/n$, so that the length of the observation time interval $n\Delta_n = 1$ is fixed, can only lead to estimating the diffusion coefficient. This is done by Hoffmann (1999) who generalizes results by Jacod (2000), Florens-Smirou (1993) and Genon-Catalot et al. (1992).

Now, estimating both drift and diffusion coefficients requires that the sampling interval $\Delta_n$ tends to zero while $n\Delta_n$ tends to infinity. For ergodic diffusion models, Hoffmann (1999) proposes nonparametric estimators using projections on wavelet bases together with adaptive procedures. He exhibits minimax rates and shows that his estimators automatically reach these optimal rates up to logarithmic factors. Hoffmann's estimators are based on computations of some random times which make them difficult to implement. Let us mention that Bandi and Phillips (2003) also consider the same asymptotic framework but with nonstationary diffusion processes: they study kernel estimators using local time estimations and random normalizations.

In this paper, we propose simple nonparametric estimators based on a penalized mean square approach. The method is investigated in full details in Comte and Rozenholc (2002,

2004) for regression models. We adapt it here to the case of discretized diffusion models. The estimators are chosen to belong to finite dimensional spaces that include trigonometric, wavelet generated and piecewise polynomials spaces. The space dimension is chosen by a data driven method using a penalization device. Due to the construction of our estimators, we measure the risk of an estimator $\hat{f}$ of $f$ (with $f = b, \sigma^2$) by $\mathbb{E}(\|\hat{f} - f\|_n^2)$ where $\|\hat{f} - f\|_n^2 = n^{-1} \sum_{k=1}^n (\hat{f}(X_{k\Delta}) - f(X_{k\Delta}))^2$. We give bounds for this risk (see Theorem 3.1 and Theorem 4.1). Looking at these bounds as $\Delta = \Delta_n \to 0$ and $n\Delta_n \to +\infty$ shows that our estimators achieve the optimal nonparametric asymptotic rates obtained in Hoffmann (1999) without logarithmic loss (when the unknown functions belong to Besov balls). Then we proceed to numerical implementation on simulated data for several examples of models. We emphasize that our simulation method for diffusion processes is not based on approximations (like Euler schemes). Instead, we use the exact retrospective simulation method described in Beskos et al. (2004) and Beskos and Roberts (2005). Then we apply the algorithms developed in Comte and Rozenholc (2002,2004) for nonparametric estimation using piecewise polynomials. The results are convincing even when some of the theoretical assumptions are not fulfilled.

The paper is organized as follows. In Section 2, we describe our framework (model, assumptions and spaces of approximation). Section 3 is devoted to drift estimation, Section 4 to diffusion coefficient estimation. In Section 5, we study examples and present numerical simulation results that illustrate the performances of estimators. Section 6 contains proofs. In the Appendix (Section 7), the retrospective exact simulation algorithm is briefly described.

## 2. Framework and assumptions

2.1. **Model assumptions.** Let $(X_t)_{t\geq 0}$ be a solution of (1) and assume that $n+1$ observations $X_{k\Delta}$, $k = 1, \dots, n+1$ with sampling interval $\Delta$ are available. Throughout the paper, we assume that $\Delta = \Delta_n$ tends to 0 and $n\Delta_n$ tends to infinity as $n$ tends to infinity. To simplify notations, we only write $\Delta$ without the subscript $n$. Nevertheless, when speaking of constants, we mean quantities that depend neither on $n$ nor on $\Delta$. We want to estimate the drift function $b$ and the diffusion coefficient $\sigma^2$ when $X$ is stationary and geometrically $\beta$-mixing. To this end, we consider the following assumptions:

[A1 ] $(i)$ $b \in C^1(\mathbb{R})$ and $\exists \gamma \geq 0, \forall x \in \mathbb{R}, |b'(x)| \leq \gamma(1 + |x|^\gamma)$,
$(ii)$ $\exists b_0,\ \forall x,\ \ |b(x)| \leq b_0(1 + |x|)$,
$(iii)$ $\exists d \geq 0, r > 0, \exists R > 0,\ \forall |x| \geq R,\ \mathrm{sgn}(x)b(x) \leq -r|x|^d$.

[A2 ] $(i)$ $\exists \sigma_0^2, \exists \sigma_1^2,\ \ \forall x,\ 0 < \sigma_0^2 \leq \sigma^2(x) \leq \sigma_1^2$ and $\exists L,\ \ \forall (x,y) \in \mathbb{R}^2,\ \ |\sigma(x) - \sigma(y)| \leq L|x-y|^{1/2}$.
$(ii)$ $\sigma \in C^2(\mathbb{R})$ and $\exists \gamma \geq 0, \forall x \in \mathbb{R}, |\sigma'(x)| + |\sigma"(x)| \leq \gamma(1 + |x|^\gamma)$.

Under [A1]-[A2], Equation (1) has a unique strong solution. Note that [A2]$(ii)$ is only used for the estimation of $\sigma^2$ and not for $b$.

Elementary computations show that the scale density

$$s(x) = \exp\left[-2\int_0^x \frac{b(u)}{\sigma^2(u)} du\right]$$

satisfies $\int_{-\infty} s(x)dx = +\infty = \int^{+\infty} s(x)dx$ and the speed density $m(x) = 1/(\sigma^2(x)s(x))$ satisfies $\int_{-\infty}^{+\infty} m(x)dx = M < +\infty$.

Hence, Model (1) admits a unique invariant probability $\pi(x)dx$ with $\pi(x) = M^{-1}m(x)$. Now we assume that

[A3 ] $X_0 = \eta$ has distribution $\pi$.

Under the additional Assumption [A3], $(X_t)$ is strictly stationary and ergodic.

Moreover, it follows from Proposition 1 in Pardoux and Veretennikov (2001) that there exist constants $K > 0$, $\nu > 0$ and $\theta > 0$ such that:

$$\mathbb{E}(\exp(\nu|X_0|)) < +\infty \text{ and } \beta_X(t) \leq Ke^{-\theta t}, \tag{2}$$

where $\beta_X(t)$ denotes the $\beta$-mixing coefficient of $(X_t)$ and is given by

$$\beta_X(t) = \int_{-\infty}^{+\infty} \pi(x)dx \|P_t(., dx') - \pi(x')dx'\|_{TV}.$$

The norm $\|.\|_{TV}$ is the total variation norm and $P_t$ denotes the transition probability. In particular, $X_0$ has moments of any (positive) order.

Now, [A1] $(i)$ ensures that for all $t \geq 0$, $h > 0$, and $k \geq 1$, there exists $c = c(k, \gamma)$ such that

$$\mathbb{E}( \sup_{s \in [t,t+h]} [|b(X_s) - b(X_t)|^k | \mathcal{F}_t) \leq ch^{k/2}(1 + |X_t|^c),$$

where $\mathcal{F}_t = \sigma(X_s, s \leq t)$ (see e.g. Gloter (2000), Proposition A). Thus, taking expectations, there exists $c'$ such that

$$\mathbb{E}( \sup_{s \in [t,t+h]} [|b(X_s) - b(X_t)|^k) \leq c'h^{k/2}. \tag{3}$$

Functions $b$ and $\sigma^2$ are estimated only on a compact set $A$. For simplicity and without loss of generality, we assume from now on that

$$A = [0, 1]. \tag{4}$$

It follows from [A1], [A2]$(i)$ and [A3] that the stationary density $\pi$ is bounded from below and above on any compact subset of $\mathbb{R}$ and we denote by $\pi_0$, $\pi_1$ two positive real numbers such that

$$0 < \pi_0 \leq \pi(x) \leq \pi_1, \ \ \forall x \in A = [0, 1]. \tag{5}$$

2.2. **Spaces of approximation: Piecewise polynomials.** We aim at estimating functions $b$ and $\sigma^2$ of Model (1) on $[0, 1]$ using a data driven procedure. For that purpose, we consider families of finite dimensional linear subspaces of $\mathbb{L}^2([0, 1])$ and compute for each space an associated least-squares estimator. Afterwards, an adaptive procedure chooses among the resulting collection of estimators the "best" one, in a sense that will be later specified, through a penalization device.

Several possible collections of spaces are available and discussed in Section 2.3. Now, to be consistent with the algorithm implemented in Section 5, we focus on a specific collection, namely the collection of dyadic regular piecewise polynomial spaces, denoted hereafter by [DP].

We fix an integer $r \geq 0$. Let $p \geq 0$ an integer. On each subinterval $I_j = [(j-1)/2^p, j/2^p]$, $j = 1, \ldots, 2^p$, consider $r + 1$ polynomials of degree $0, 1, \ldots, r$, $\varphi_{j,\ell}(x)$, $\ell = 0, 1, \ldots r$ and set $\varphi_{j,\ell}(x) = 0$ outside $I_j$. The space $S_m$, $m = (p, r)$, is defined as generated by the $D_m = 2^p(r + 1)$ functions $(\varphi_{j,\ell})$. A function $t$ in $S_m$ may be written as

$$t(x) = \sum_{j=1}^{2^p} \sum_{\ell=0}^{r} t_{j,\ell}\varphi_{j,\ell}(x).$$

The collection of spaces $(S_m, m \in \mathcal{M}_n)$ is such that

$$\mathcal{M}_n = \{m = (p, r), p \in \mathbb{N}, 2^p(r + 1) \leq N_n\}. \tag{6}$$

In other words, $D_m \leq N_n$ where $N_n \leq n$. The maximal dimension $N_n$ is subject to additional constraints given below.

More concretely, this is realized as follows. Consider the orthogonal collection in $\mathbb{L}^2([-1, 1])$ of Legendre polynomials $(Q_\ell, \ell \geq 0)$, where the degree of $Q_\ell$ is equal to $\ell$, generating $\mathbb{L}^2([-1, 1])$ (see Abramowitz and Stegun (1972), p.774). They satisfy $|Q_\ell(x)| \leq 1, \forall x \in [-1, 1]$, $Q_\ell(1) = 1$ and $\int_{-1}^{1} Q_\ell^2(u)du = 2/(2\ell+1)$. Then we set $P_\ell(x) = \sqrt{2\ell + 1}Q_\ell(2x-1)$, to get an orthonormal basis of $\mathbb{L}^2([0, 1])$. And finally,

$$\varphi_{j,\ell}(x) = 2^{p/2}P_\ell(2^px - j + 1)\mathbf{1}_{I_j}(x), \ \ j = 1, \dots, 2^p, \ \ell = 0, 1, \dots, r.$$

The space $S_m$ has dimension $D_m = 2^p(r + 1)$ and its orthonormal basis described above satisfies

$$\left\| \sum_{j=1}^{2^p} \sum_{\ell=0}^{r} \varphi_{j,\ell}^2 \right\|_\infty \leq D_m(r + 1).$$

Hence, for all $t \in S_m$, $\|t\|_\infty \leq \sqrt{r + 1}\sqrt{D_m}\|t\|$, where $\|t\|^2 = \int_0^1 t^2(x)dx$, for $t$ in $\mathbb{L}^2([0, 1])$, a property which is essential for the proofs.

In particular, the histogram basis corresponds to $r = 0$ and is simply defined by $\varphi_j(x) = \sqrt{2^p}\, \mathbf{1}_{[(j-1)/2^p, j/2^p[}(x)$.

2.3. **Other spaces of approximation.** From both theoretical and practical points of view, other spaces can be considered as, for example:
[T] *Trigonometric spaces*: $S_m$ is generated by $\{\ 1, \ \sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx)$ for $j = 1, \dots, m\ \}$, has dimension $D_m = 2m + 1$ and $m \in \mathcal{M}_n = \{1, \dots, [n/2] - 1\}$.
[W] *Dyadic wavelet generated spaces* with regularity $r$ and compact support, as described e.g. in Daubechies (1992), Donoho et al. (1996) or Hoffmann (1999).

The key properties that must be fulfilled to fit in our framework are the following:

$(\mathcal{H}_1)$ Norm connection: $(S_m)_{m\in\mathcal{M}_n}$ is a collection of finite-dimensional linear sub-spaces of $\mathbb{L}^2([0, 1])$, with dimension $\dim(S_m) = D_m$ such that $D_m \leq n$, $\forall m \in \mathcal{M}_n$ and satisfying:

$$\exists \Phi_0 > 0, \forall m \in \mathcal{M}_n, \forall t \in S_m, \|t\|_\infty \leq \Phi_0\sqrt{D_m}\|t\|. \tag{7}$$

An orthonormal basis of $S_m$ is denoted by $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ where $|\Lambda_m| = D_m$. It follows from Birgé and Massart (1997) that Property (7) in the context of $(\mathcal{H}_1)$ is equivalent to

$$\exists \Phi_0 > 0, \| \sum_{\lambda\in\Lambda_m} \varphi_\lambda^2 \|_\infty \leq \Phi_0^2 D_m. \tag{8}$$

Thus, for collection [DP], (8) holds with $\Phi_0^2 = r + 1$. Moreover, for results concerning adaptive estimators, we need an additional assumption:

$(\mathcal{H}_2)$ Nesting condition: $(S_m)_{m\in\mathcal{M}_n}$ is a collection of nested models, we denote by $\mathcal{S}_n$ the space belonging to the collection, such that $\forall m \in \mathcal{M}_n, S_m \subset \mathcal{S}_n$. We denote by $N_n$ the dimension of $\mathcal{S}_n$: $\dim(\mathcal{S}_n) = N_n$ $(\forall m \in \mathcal{M}_n, D_m \leq N_n)$.

As much as possible below, we keep general notations to allow extensions to other spaces of approximation than the collection [DP].

## 3. Drift estimation

### 3.1. **Drift estimators: non adaptive case.**

Let

$$Y_{k\Delta} = \frac{X_{(k+1)\Delta} - X_{k\Delta}}{\Delta} \quad \text{and} \quad Z_{k\Delta} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} \sigma(X_s)dW_s. \tag{9}$$

The following standard regression-type decomposition holds:

$$Y_{k\Delta} = b(X_{k\Delta}) + Z_{k\Delta} + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} (b(X_s) - b(X_{k\Delta}))ds$$

where $b(X_{k\Delta})$ is the main term, $Z_{k\Delta}$ the noise term and the last term is a negligible residual.

Now, for $S_m$ a space of the collection $\mathcal{M}_n$ and for $t \in S_m$, we consider the following regression contrast:

$$\gamma_n(t) = \frac{1}{n}\sum_{k=1}^{n}[Y_{k\Delta} - t(X_{k\Delta})]^2. \tag{10}$$

The estimator belonging to $S_m$ is defined as

$$\hat{b}_m = \arg\min_{t\in S_m} \gamma_n(t). \tag{11}$$

A minimizer of $\gamma_n$ in $S_m$, $\hat{b}_m$, always exists but may not be unique. Indeed in some common situations the minimization of $\gamma_n$ over $S_m$ leads to an affine space of solutions. Consequently, it becomes impossible to consider a classical $\mathbb{L}^2$-risk for "the least-squares estimator" of $b$ in $S_m$. In contrast, the random $\mathbb{R}^n$-vector $(\hat{b}_m(X_\Delta), \ldots, \hat{b}_m(X_{n\Delta}))'$ is always uniquely defined. Indeed, let us denote by $\Pi_m$ the orthogonal projection (with respect to the inner product of $\mathbb{R}^n$) onto the subspace $\{(t(X_\Delta), \ldots, t(X_{n\Delta}))', t \in S_m\}$ of $\mathbb{R}^n$. Then $(\hat{b}_m(X_\Delta), \ldots, b_m(X_{n\Delta}))' = \Pi_m Y$ where $Y = (Y_\Delta, \ldots, Y_{n\Delta})'$. This is the reason why we define the risk of $\hat{b}_m$ by

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}(\hat{b}_m(X_{k\Delta}) - b(X_{k\Delta}))^2\right] = \mathbb{E}(\|\hat{b}_m - b\|_n^2)$$

where

$$\|t\|_n^2 = \frac{1}{n}\sum_{k=1}^{n} t^2(X_{k\Delta}). \tag{12}$$

Thus, our risk is the expectation of an empirical norm. Note that, for a deterministic function $t$, $\mathbb{E}(\|t\|_n^2) = \|t\|_\pi^2 = \int t^2(x)d\pi(x)$ where $\pi$ denotes the stationary law. In view of (5), the $\mathbb{L}^2$-norm, $\|.\|$, and the $\mathbb{L}^2(\pi)$-norm, $\|.\|_\pi$, are equivalent for $A$-supported functions, a property that is used below.

3.2. **Risk of the non adaptive drift estimator.** Using (9)-(10)-(12), we have:

$$\begin{aligned}\gamma_n(t)-\gamma_n(b) &= \|t-b\|_n^2+\frac{2}{n}\sum_{k=1}^{n}(b-t)(X_{k\Delta})Z_{k\Delta}\\ &\quad+\frac{2}{n\Delta}\sum_{k=1}^{n}(b-t)(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\end{aligned}$$

In view of this decomposition, we define the centered empirical process:

$$\nu_n(t)=\frac{1}{n}\sum_{k=1}^{n}t(X_{k\Delta})Z_{k\Delta}. \tag{13}$$

Now denote by $b_m$ the orthogonal projection of $b$ on $S_m$. By definition of $\hat{b}_m$, $\gamma_n(\hat{b}_m)\leq\gamma_n(b_m)$. So, $\gamma_n(\hat{b}_m)-\gamma_n(b)\leq\gamma_n(b_m)-\gamma_n(b)$. This implies

$$\begin{aligned}\|\hat{b}_m-b\|_n^2 &\leq \|b_m-b\|_n^2+2\nu_n(\hat{b}_m-b_m)\\ &\quad+\frac{2}{n\Delta}\sum_{k=1}^{n}(\hat{b}_m-b_m)(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\end{aligned}$$

The functions $\hat{b}_m$ and $b_m$ being $A$-supported, we can cancel the terms $\|b\mathbf{1}_{A^c}\|_n^2$ that appears in both sides of the inequality. This yields

$$\begin{aligned}\|\hat{b}_m-b\mathbf{1}_A\|_n^2 &\leq \|b_m-b\mathbf{1}_A\|_n^2+2\nu_n(\hat{b}_m-b_m)\\ &\quad+\frac{2}{n\Delta}\sum_{k=1}^{n}(\hat{b}_m-b_m)(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\end{aligned} \tag{14}$$

On the basis of this inequality, we obtain the following result.

**Proposition 3.1.** *Let $\Delta=\Delta_n$ be such that $\Delta_n\to 0,\ n\Delta_n/\ln^2(n)\to+\infty$ when $n\to+\infty$. Assume that* [A1], [A2]*(i),* [A3] *hold and consider a space $S_m$ in the collection* [DP] *with $N_n=o(n\Delta/\ln^2(n))$ ($N_n$ is defined in ($\mathcal{H}_2$)). Then the estimator $\hat{b}_m$ of $b$ is such that*

$$\mathbb{E}(\|\hat{b}_m-b_A\|_n^2)\leq 7\pi_1\|b_m-b_A\|^2+K\frac{\mathbb{E}(\sigma^2(X_0))D_m}{n\Delta}+K'\Delta+\frac{K"}{n\Delta}, \tag{15}$$

*where $b_A=b\mathbf{1}_{[0,1]}$ and $K,K'$ and $K"$ are some positive constants.*

As a consequence, it is natural to select the dimension $D_m$ that leads to the best compromise between the squared bias term $\|b_m-b_A\|^2$ and the variance term of order $D_m/(n\Delta)$.

To compare the result of Proposition 3.1 with the optimal nonparametric rates exhibited by Hoffmann (1999), let us assume that $b_A$ belongs to a ball of some Besov space, $b_A\in\mathcal{B}_{\alpha,2,\infty}([0,1])$, and that $r+1\geq\alpha$. Then, for $\|b_A\|_{\alpha,2,\infty}\leq L$, we have $\|b_A-b_m\|^2\leq C(\alpha,L)D_m^{-2\alpha}$. Thus, choosing $D_m=(n\Delta)^{1/(2\alpha+1)}$, we obtain

$$\mathbb{E}(\|\hat{b}_m-b_A\|_n^2)\leq C(\alpha,L)(n\Delta)^{-2\alpha/(2\alpha+1)}+K'\Delta+\frac{K"}{n\Delta}. \tag{16}$$

The first term $(n\Delta)^{-2\alpha/(2\alpha+1)}$ is exactly the optimal nonparametric rate (see Hoffmann (1999)). Moreover, under the standard condition $\Delta=o(1/(n\Delta))$, the last two terms in (15) are

$O(1/(n\Delta))$ which is negligible with respect to $(n\Delta)^{-2\alpha/(2\alpha+1)}$.

Proposition 3.1 holds for the wavelet basis [W] under the same assumptions. For the trigonometric basis [T], the additional constraint $N_n \le O(\sqrt{n\Delta}/\ln(n))$ is necessary. Hence, when working with those bases, if $b_A \in \mathcal{B}_{\alpha,2,\infty}([0,1])$ as above, the optimal rate is reached for the same choice for $D_m$, under the additional constraint that $\alpha > 1/2$ for [T]. It is worth stressing that $\alpha > 1/2$ automatically holds under [A1].

### 3.3. **Adaptive drift estimator.**

As a second step, we must ensure an automatic selection of $D_m$, which does not use any knowledge on $b$, and in particular which does not require to know $\alpha$. This selection is standardly done by

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\left[\gamma_n(\hat{b}_m) + \mathrm{pen}(m)\right], \tag{17}$$

with pen$(m)$ a penalty to be properly chosen. We denote by $\hat{b}_{\hat{m}}$ the resulting estimator and we need to determine pen(.) such that, ideally,

$$\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_n^2) \le C \inf_{m\in\mathcal{M}_n}\left(\|b_A - b_m\|^2 + \frac{\mathbb{E}(\sigma^2(X_0))D_m}{n\Delta}\right) + K'\Delta + \frac{K"}{n\Delta},$$

with $C$ a constant which should not be too large. We almost reach this aim.

**Theorem 3.1.** *Let $\Delta = \Delta_n$ be such that $\Delta_n \to 0$, $n\Delta_n/\ln^2(n) \to +\infty$ when $n \to +\infty$. Assume that* [A1]-[A2]*(i),* [A3] *hold and consider the nested collection of models [DP] with maximal dimension $N_n = o(n\Delta/\ln^2(n))$. Let*

$$\mathrm{pen}(m) = \kappa\sigma_1^2\frac{D_m}{n\Delta}, \tag{18}$$

*where $\kappa$ is a universal constant. Then the estimator $\hat{b}_{\hat{m}}$ of $b$ with $\hat{m}$ defined in (17) is such that*

$$\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_n^2) \le C \inf_{m\in\mathcal{M}_n}\left(\|b_m - b_A\|^2 + \frac{\sigma_1^2 D_m}{n\Delta}\right) + K'\Delta + \frac{K"}{n\Delta}. \tag{19}$$

Some comments need to be made. The constant $\kappa$ in the penalty is numerical and must be calibrated for the problem. Its value is usually adapted by intensive simulation experiments. This point is discussed and clarified in Section 5.2. From (15), one would expect to obtain $\mathbb{E}(\sigma^2(X_0))$ instead of $\sigma_1^2$ in the penalty (18): We do not know if this is the consequence of technical problems or if this is a structural result. Another important point is that $\sigma_1^2$ is unknown. In practice, we just replace it by a rough estimator of $\mathbb{E}(\sigma^2(X_0))$ (see Section 5.2).

From (19), we deduce that the adaptive estimator automatically realizes the bias-variance compromise: Whenever $b_A$ belongs to some Besov ball (see (16)), if $r+1 \ge \alpha$ and $n\Delta^2 = o(1)$, $\hat{b}_{\hat{m}}$ achieves the optimal corresponding nonparametric rate, without logarithmic loss contrary to Hoffmann's adaptive estimator (see Hoffmann (1999, Theorem 5 p.159)). As mentioned above, Theorem 3.1 holds for the basis [W] and, if $N_n = o(\sqrt{n\Delta}/\ln(n))$, for [T] .

## 4. Adaptive estimation of the diffusion coefficient

4.1. **Diffusion coefficient estimator: non adaptive case.** To estimate $\sigma^2$ on $A = [0,1]$, we define

$$\hat{\sigma}^2_m = \arg\min_{t\in S_m} \breve{\gamma}_n(t), \text{ with } \breve{\gamma}_n(t) = \frac{1}{n}\sum_{k=1}^{n}\left[U_{k\Delta} - t(X_{k\Delta})\right]^2, \tag{20}$$

and

$$U_{k\Delta} = \frac{(X_{(k+1)\Delta} - X_{k\Delta})^2}{\Delta}. \tag{21}$$

For diffusion coefficient estimation under our asymptotic framework, it is now well known that rates of convergence are faster than for drift estimation. This is the reason why the regression-type equation has to be more precise than for $b$. Let us set

$$\psi = 2(\sigma'\sigma b) + [(\sigma')^2 + \sigma\sigma"]\sigma^2. \tag{22}$$

Some computations using Ito's formula and Fubini's theorem lead to

$$U_{k\Delta} = \sigma^2(X_{k\Delta}) + V_{k\Delta} + R_{k\Delta}$$

where $V_{k\Delta} = V^{(1)}_{k\Delta} + V^{(2)}_{k\Delta} + V^{(3)}_{k\Delta}$ with

$$\begin{aligned}
V^{(1)}_{k\Delta} &= \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)dW_s\right)^2 - \int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s)ds\right]\\
V^{(2)}_{k\Delta} &= \frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta}[(k+1)\Delta - s]\sigma'(X_s)\sigma^2(X_s)dW_s,\\
V^{(3)}_{k\Delta} &= 2b(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)dW_s,
\end{aligned}$$

and

$$\begin{aligned}
R_{k\Delta} &= \frac{1}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} b(X_s)ds\right)^2 + \frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(b(X_s) - b(X_{k\Delta}))ds\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)dW_s\\
&\quad + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}[(k+1)\Delta - s]\psi(X_s)ds,
\end{aligned}$$

Obviously, the main noise term in the above decomposition must be $V^{(1)}_{k\Delta}$ which will be proved below.

4.2. **Risk of the non adaptive estimator.** As for the drift, we write:

$$\breve{\gamma}_n(t) - \breve{\gamma}_n(\sigma^2) = \|\sigma^2 - t\|_n^2 + \frac{2}{n}\sum_{k=1}^{n}(\sigma^2 - t)(X_{k\Delta})V_{k\Delta} + \frac{2}{n}\sum_{k=1}^{n}(\sigma^2 - t)(X_{k\Delta})R_{k\Delta}.$$

We denote by $\sigma^2_m$ the orthogonal projection of $\sigma^2$ on $S_m$ and define

$$\breve{\nu}_n(t) = \frac{1}{n}\sum_{k=1}^{n} t(X_{k\Delta})V_{k\Delta}.$$

Again we use that $\breve{\gamma}_n(\hat{\sigma}^2_m) - \breve{\gamma}_n(\sigma^2) \leq \breve{\gamma}_n(\sigma^2_m) - \breve{\gamma}_n(\sigma^2)$ to obtain

$$\|\hat{\sigma}^2_m - \sigma^2\|^2_n \quad \leq \quad \|\sigma^2_m - \sigma^2\|^2_n + \frac{2}{n}\sum_{k=1}^{n}(\hat{\sigma}^2_m - \sigma^2_m)(X_{k\Delta})V_{k\Delta} + \frac{2}{n}\sum_{k=1}^{n}(\hat{\sigma}^2_m - \sigma^2_m)(X_{k\Delta})R_{k\Delta}.$$

Analogously, we can cancel on both sides the common term $\|\sigma_{A^c}\|^2_n$. This yields

$$\|\hat{\sigma}^2_m - \sigma^2_A\|^2_n \quad \leq \quad \|\sigma^2_m - \sigma^2_A\|^2_n + 2\breve{\nu}_n(\hat{\sigma}^2_m - \sigma^2_m) + \frac{2}{n}\sum_{k=1}^{n}(\hat{\sigma}^2_m - \sigma^2_m)(X_{k\Delta})R_{k\Delta}. \tag{23}$$

And, we obtain the result

**Proposition 4.1.** *Let $\Delta = \Delta_n$ be such that $\Delta_n \to 0$, $n\Delta_n/\ln^2(n) \to +\infty$ when $n \to +\infty$. Assume that* [A1]-[A3] *hold and consider a model $S_m$ in the collection* [DP] *with $N_n = o(n\Delta/\ln^2(n))$ where $N_n$ is defined in $(\mathcal{H}_2)$. Then the estimator $\hat{\sigma}^2_m$ of $\sigma^2$ defined by (20) is such that*

$$\mathbb{E}(\|\hat{\sigma}^2_m - \sigma^2_A\|^2_n) \leq 7\pi_1\|\sigma^2_m - \sigma^2_A\|^2 + K\frac{\mathbb{E}(\sigma^4(X_0))D_m}{n} + K'\Delta^2 + \frac{K"}{n}, \tag{24}$$

*where $\sigma^2_A = \sigma^2\mathbf{1}_{[0,1]}$, and $K$, $K'$, $K"$ are some positive constants.*

Let us make some comments on the rates of convergence. If $\sigma^2_A$ belongs to a ball of some Besov space, say $\sigma^2_A \in \mathcal{B}_{\alpha,2,\infty}([0,1])$, and $\|\sigma^2_A\|_{\alpha,2,\infty} \leq L$, with $r+1 \geq \alpha$, then $\|\sigma^2_A - \sigma^2_m\|^2 \leq C(\alpha, L)D_m^{-2\alpha}$. Therefore, if we choose $D_m = n^{1/(2\alpha+1)}$, we obtain

$$\mathbb{E}(\|\hat{\sigma}^2_m - \sigma^2_A\|^2_n) \leq C(\alpha, L)n^{-2\alpha/(2\alpha+1)} + K'\Delta^2 + \frac{K"}{n}. \tag{25}$$

The first term $n^{-2\alpha/(2\alpha+1)}$ is the optimal nonparametric rate proved by Hoffmann (1999). Moreover, under the standard condition $\Delta^2 = o(1/n)$, the last two terms are $O(1/n)$, *i.e.* negligible with respect to $n^{-2\alpha/(2\alpha+1)}$.

4.3. **Adaptive diffusion coefficient estimator.** As previously, the second step is to ensure an automatic selection of $D_m$, which does not use any knowledge on $\sigma^2$. This selection is done by

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\left[\breve{\gamma}_n(\hat{\sigma}^2_m) + \widetilde{\mathrm{pen}}(m)\right]. \tag{26}$$

We denote by $\hat{\sigma}^2_{\hat{m}}$ the resulting estimator and we need to determine the penalty $\widetilde{\mathrm{pen}}$ as for $b$. We can prove the following theorem.

**Theorem 4.1.** *Let $\Delta = \Delta_n$ be such that $\Delta_n \to 0$, $n\Delta_n/\ln^2(n) \to +\infty$ when $n \to +\infty$. Assume that* [A1]-[A3] *hold. Consider the nested collection of models [DP] with maximal dimension $N_n \leq n\Delta/\ln^2(n)$. Let*

$$\widetilde{\mathrm{pen}}(m) = \tilde{\kappa}\sigma_1^4\frac{D_m}{n}, \tag{27}$$

*where $\tilde{\kappa}$ is a universal constant. Then, the estimator $\hat{\sigma}^2_{\hat{m}}$ of $\sigma^2$ with $\hat{m}$ defined by (26) is such that*

$$\mathbb{E}(\|\hat{\sigma}^2_{\hat{m}} - \sigma^2_A\|^2_n) \leq C\inf_{m\in\mathcal{M}_n}\left(\|\sigma^2_m - \sigma^2_A\|^2 + \frac{\sigma_1^4 D_m}{n}\right) + K'\Delta^2 + \frac{K"}{n}. \tag{28}$$

As for the drift, it follows from (28) that the adaptive estimator automatically realizes the bias-variance compromise. Whenever $\sigma_A^2$ belongs to some Besov ball (see (25)), if $n\Delta^2 = o(1)$ and $r+1 \geq \alpha$, $\hat{\sigma}^2_{\hat{m}}$ achieves the optimal corresponding nonparametric rate $n^{-2\alpha/(2\alpha+1)}$, without logarithmic loss contrary to Hoffmann's adaptive estimator (see Hoffmann (1999, Theorem 6 p.160)). As mentioned for $b$, Proposition 4.1 and Theorem 4.1 hold for the basis [W] under the same assumptions on $N_n$. For [T], $N_n = o(\sqrt{n\Delta}/\ln(n))$ is needed.

## 5. Examples and numerical simulation results

In this section, we consider examples of diffusions and implement the estimation algorithms on simulated data. To simulate sample paths of diffusion, we use the retrospective exact simulation algorithms proposed by Beskos et al. (2004) and Beskos and Roberts (2005). Contrary to the Euler scheme, these algorithms produce exact simulation of diffusions under some assumptions on the drift and diffusion coefficient. Therefore, we choose our examples in order to fit in these conditions in addition with our set of assumptions. For sake of simplicity, we focus on models that can be simulated by the simplest algorithm of Beskos et al. (2004), which is called EA1. More precisely, consider a diffusion model given by the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t. \tag{29}$$

We assume that there is a $C^2$ one-to-one mapping $F$ on $\mathbb{R}$ such that $\xi_t = F(X_t)$ satisfies

$$d\xi_t = \alpha(\xi_t)dt + dW_t. \tag{30}$$

To produce an exact realization of the random variable $\xi_\Delta$, given that $\xi_0 = x$, the exact algorithm EA1 requires that $\alpha$ be $C^1$, $\alpha^2 + \alpha'$ be bounded from below and above. Moreover, setting $A(\xi) = \int^\xi \alpha(u)du$, the function

$$h(\xi) = \exp\left(A(\xi) - (\xi - x)^2/2\Delta\right) \tag{31}$$

must be integrable on $\mathbb{R}$ and an exact realization of a random variable with density proportional to $h$ must be possible. Provided that the process $(\xi_t)$ admits a stationary distribution that is also possibly simulatable, using the Markov property, the algorithm can therefore produce an exact realization of a discrete sample $(\xi_{k\Delta}, k = 0, 1, \ldots, n+1)$ in stationary regime. We deduce an exact realization of $(X_{k\Delta} = F^{-1}(\xi_{k\Delta}), k = 0, \ldots, n+1)$.

In all examples, we estimate the drift function $\alpha(\xi)$ and the constant 1 for models like (30) or both the drift $b(x)$ and the diffusion coefficient $\sigma^2(x)$ for models like (29). Let us note that Assumptions [A1]-[A2]-[A3] are fulfilled for all the models $(\xi_t)$ below. For the models $(X_t)$, the ergodicity and the exponential $\beta$-mixing property hold.

### 5.1. **Examples of diffusions.**

5.1.1. *Family 1.* First, we consider (29) with

$$b(x) = -\theta x, \quad \sigma(x) = c\sqrt{1+x^2}. \tag{32}$$

Standard computations of the scale and speed densities show that the model is positive recurrent for $\theta + c^2/2 > 0$. In this case, its stationary distribution has density

$$\pi(x) \propto \frac{1}{(1+x^2)^{1+\theta/c^2}}.$$

If $X_0 = \eta$ has distribution $\pi(x)dx$, then, setting $\nu = 1 + 2\theta/c^2$, $\sqrt{\nu}\,\eta$ has Student distribution $t(\nu)$ which can be easily simulated.

Now, we consider $F_1(x) = \int_0^x 1/(c\sqrt{1+x^2})dx = \arg\sinh(x)/c$. By the Ito formula, $\xi_t = F_1(X_t)$ satisfies (30) with

$$\alpha(\xi) = -(\theta/c + c/2)\tanh(c\xi). \tag{33}$$

Assumptions [A1]-[A3] hold for $(\xi_t)$ with $\xi_0 = F_1(X_0)$. Moreover,

$$\alpha^2(\xi) + \alpha'(\xi) = [(\theta/c + c/2)^2 + \theta + c^2/2]\tanh^2(c\xi) - (\theta + c^2/2)$$

is bounded from below and above. And

$$A(\xi) = \int_0^\xi \alpha(u)du = -(1/2 + \theta/c^2)\log(\cosh(c\xi)) \leq 0,$$

so that $\exp(A(\xi)) \leq 1$. Therefore, function (31) is integrable for all $x$ and by a simple rejection method, we can produce a realization of a random variable with density proportional to $h(\xi)$ using a random variable with density $\mathcal{N}(x, \Delta)$.

Note that model (29) satisfies Assumptions [A1]-[A3] except that $\sigma^2(x)$ is not bounded from above. Nevertheless, since $X_t = F_1^{-1}(\xi_t) = \sinh(c\,\xi_t)$, the process $(X_t)$ is exponentially $\beta$-mixing. The upper bound $\sigma_1^2$ that appears explicitly in the penalty function must be replaced by an estimated upper bound.

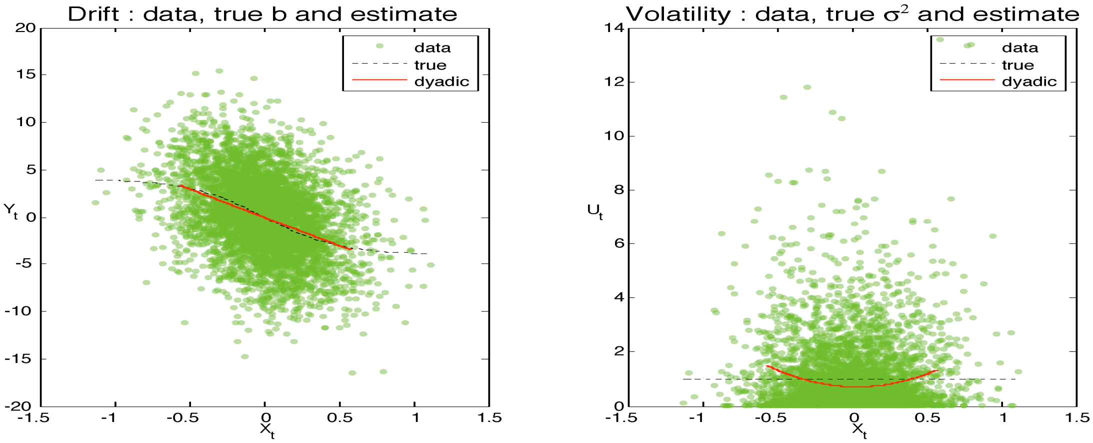

FIGURE 1. First example: $d\xi_t = -(\theta/c + c/2)\tanh(c\xi_t) + dW_t$, $n = 5000$, $\Delta = 1/20$, $\theta = 6$, $c = 2$, dotted line: true function, full line: estimated function.

5.1.2. *Family 2.* For the second family of models, we start with an equation of type (30) where the drift is now

$$\alpha(\xi) = -\theta\frac{\xi}{\sqrt{1+c^2\xi^2}}. \tag{34}$$

(Note that $\tilde{\xi}_t = c\xi_t$ satisfies an equation with drift $-\theta\tilde{\xi}/\sqrt{1+\tilde{\xi}^2}$ and constant diffusion coefficient equal to $c$).

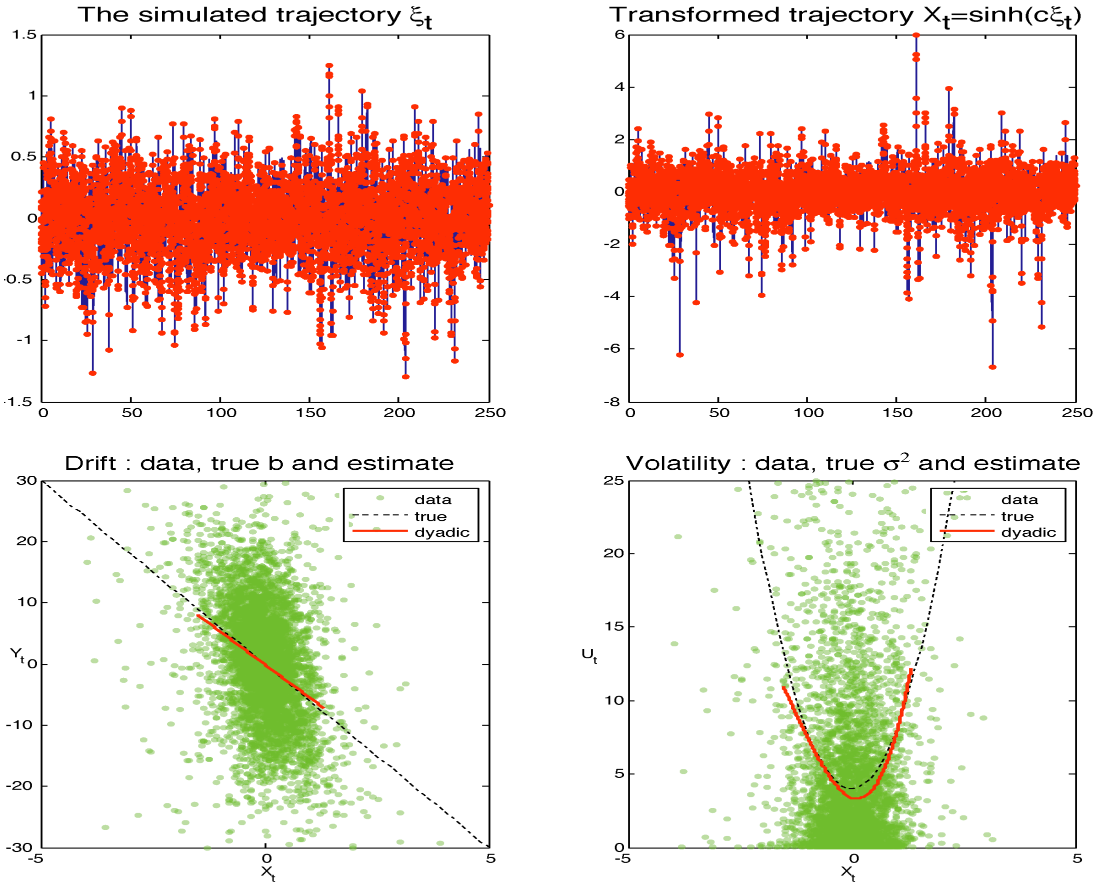

FIGURE 2. Second example: $dX_t = -\theta X_t dt +, c\sqrt{1+X_t^2} dW_t$, $n = 5000$, $\Delta = 1/20$, $\theta = 2$, $c = 1$, dotted line: true function, full line: estimated function.

The model for $(\xi_t)$ is positive recurrent on $\mathbb{R}$ for $\theta > 0$. Its stationary distribution is given by

$$\pi(\xi)d\xi \propto \exp(-2\frac{\theta}{c^2}\sqrt{1+c^2\xi^2}) = \exp(-2\theta|\xi|/c) \times \exp(\varphi(\xi)),$$

where $\exp\varphi(\xi) \leq 1$ so that a random variable with distribution $\pi(\xi)d\xi$ can be simulated by simple rejection method using a double exponential variable with distribution proportional to $\exp(-2\theta|\xi|/c)$. The conditions required to perform an exact simulation of $(\xi_t)$ hold. More precisely, $\alpha^2 + \alpha'$ is bounded from below and above and $A(\xi) = \int_0^\xi \alpha(u)du = -(\theta/c^2)\sqrt{1+c^2\xi^2}$. Hence $\exp(A(\xi)) \leq 1$, (31) is integrable and we can produce a realization of a random variable with density proportional to (31). Lastly, Assumptions [A1]-[A3] also hold for this model.

Now, we consider $X_t = F_2(\xi_t) = \arg\sinh(c\xi_t)$ which satisfies a stochastic differential equation with coefficients:

$$b(x) = -\left[\theta + \frac{c^2}{2\cosh(x)}\right]\frac{\sinh(x)}{\cosh^2(x)}, \quad \sigma(x) = \frac{c}{\cosh(x)}. \tag{35}$$

The process $(X_t)$ is exponentially $\beta$-mixing as $(\xi_t)$. The diffusion coefficient $\sigma(x)$ is not bounded from below but has an upper bound.

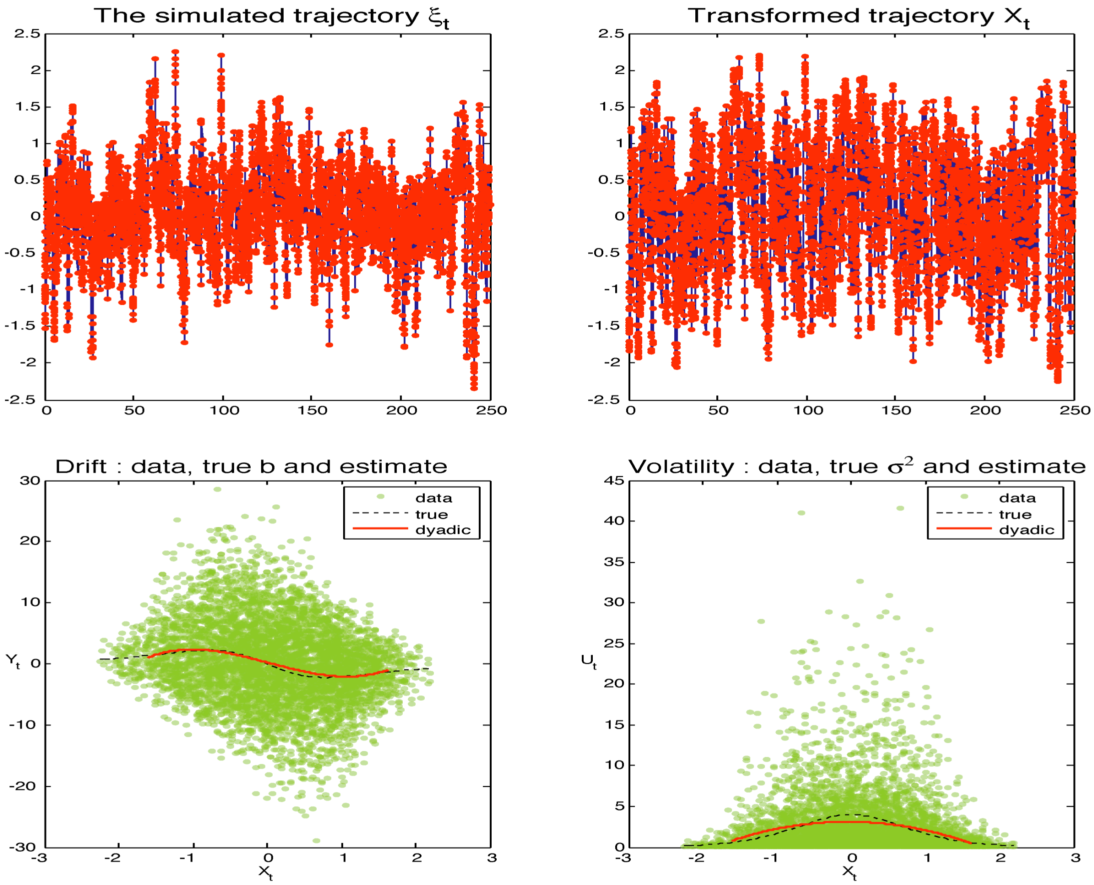

FIGURE 3. Third example, $dX_t = -\left[\theta + c^2/(2\cosh(X_t))\right](\sinh(X_t)/\cosh^2(X_t))dt + (c/\cosh(X_t))dW_t$, $n = 5000$, $\Delta = 1/20$, $\theta = 3$, $c = 2$, dotted line: true function, full line: estimated function.

To obtain a different shape for the diffusion coefficient, showing two bumps, we consider $X_t = G(\xi_t) = \arg\sinh(\xi_t - 5) + \arg\sinh(\xi_t + 5)$ where $(\xi_t)$ is as in (30)-(34). The function $G(.)$ is invertible and its inverse has the following explicit expression,

$$G^{-1}(x) = \frac{1}{\sqrt{2}\sinh(x)}\left[49\sinh^2(x) + 100 + \cosh(x)(\sinh^2(x) - 100)\right]^{1/2}.$$

The diffusion coefficient of $(X_t)$ is given by

$$\sigma(x) = \frac{1}{(1+(G^{-1}(x)-5)^2)^{1/2}} + \frac{1}{(1+(G^{-1}(x)+5)^2)^{1/2}}. \tag{36}$$

The drift is given by

$$b(x) = G'(G^{-1}(x))\alpha(G^{-1}(x)) + \frac{1}{2}G"(G^{-1}(x)).$$

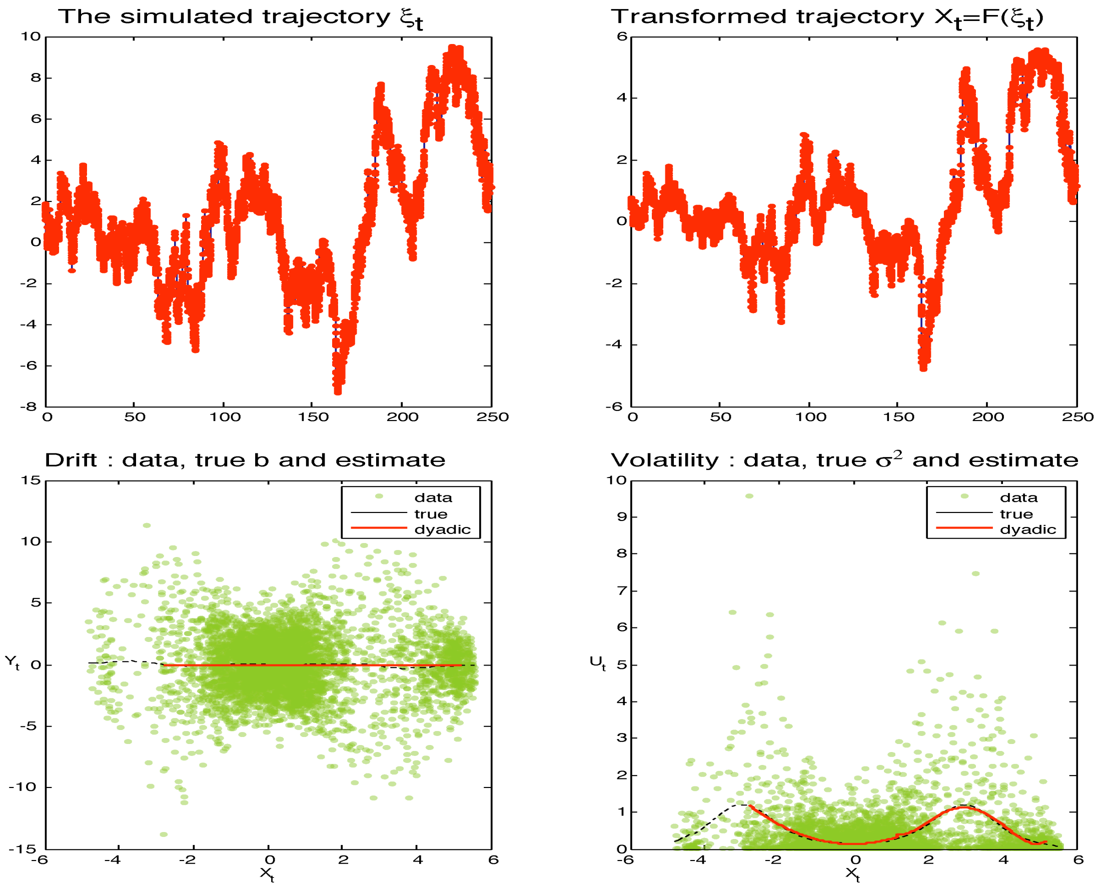

FIGURE 4. Fourth example, the two-bumps diffusion coefficient $X_t = G(\xi_t)$, $d\xi_t = -\theta\xi_t/\sqrt{1+c^2\xi_t^2}dt+dW_t$, $G(x) = \arg\sinh(x-5)+\arg\sinh(x+5)$, $n = 5000$, $\Delta = 1/20$, $\theta = 1$, $c = 10$, dotted line: true function, full line: estimated function.

5.2. **Estimation algorithms and numerical results.** We use the denoising algorithm described in full details in Comte and Rozenholc (2004). The setting here is simpler since we only consider regular dyadic piecewise polynomial spaces in the spirit of Comte and Rozenholc (2002). The algorithm minimizes the mean-square contrast and selects the space of approximation. Two versions of the algorithms are possible. The first one is the one investigated theoretically: the estimators are constructed as piecewise polynomials with fixed degree. The second version is the one used here: the algorithm also selects adaptively the degree of polynomials on each interval of the dyadic subdivision. Moreover, additive (but negligible) correcting terms are involved in the penalty (see Comte and Rozenholc (2004)).

The constant $\kappa$ in the drift penalty $\mathrm{pen}(m)$ has been set equal to 4, and in the diffusion coefficient penalty $\widetilde{\mathrm{pen}}(m)$, $\tilde{\kappa} = 2\kappa = 8$. The choice $\tilde{\kappa} = 2\kappa$ can be heuristically justified as follows: In the regression-type equation for the diffusion coefficient, the main noise term is rather of centered chi-square type. In the drift regression equation, the main noise term is rather of centered Gaussian type. Hence, a ratio $\tilde{\kappa}/\kappa = 2$.

We kept the idea that the adequate term in the penalty was $\mathbb{E}(\sigma^2(X_0))/\Delta$ for $b$ and $\mathbb{E}(\sigma^4(X_0))$ for $\sigma^2$, instead of those obtained ($\sigma_1^2/\Delta$ and $\sigma_1^4$ respectively). Therefore, in penalties, $\sigma_1^2/\Delta$ and $\sigma_1^4$ are replaced by empirical variances computed using initial estimators $\hat{b}$, $\hat{\sigma}^2$ chosen in the collection and corresponding to a space with medium dimension: $\sigma_1^2/\Delta$ for pen(.) is replaced $\hat{s}_1^2 = \gamma_n(\hat{b})$ (see (10)); and $\sigma_1^4$ for the other penalty is replaced by $\hat{s}_2^2 = \breve{\gamma}_n(\hat{\sigma}^2)$ (see (20)). Moreover, both penalties contain additional logarithmic terms which have been calibrated in other contexts by intensive simulation experiments (see Comte and Rozenholc (2002, 2004)).

More precisely, denoting by $r_j$ the degree of the polynomial on the interval $I_j = [(j-1)/2^p, j/2^p[$ for $j = 1$ to $2^p$, the function space $S_m$ is determined by $m = (p, r_1, \ldots, r_{2^p})$. The drift penalty is given by

$$\mathrm{pen}(m) = 4\frac{\hat{s}_1^2}{n}\left(\sum_{j=1}^{2^p}(r_j+1) + \sum_{j=1}^{2^p}\ln^{2.5}(r_j+1)\right)$$

and the diffusion coefficient penalty by

$$\widetilde{\mathrm{pen}}(m) = 8\frac{\hat{s}_2^2}{n}\left(\sum_{j=1}^{2^p}(r_j+1) + \sum_{j=1}^{2^p}\ln^{2.5}(r_j+1)\right).$$

Figures 1–4 illustrate our simulation results. We have plotted the sample paths $(\xi)$ (simulated) and $(X = F(\xi))$ (transformed), the data points $(X_{k\Delta}, Y_{k\Delta})$ (see (9)) and $(X_{k\Delta}, U_{k\Delta})$ (see (21)), the true functions $b$ and $\sigma^2$ and the estimated functions based on 95% of data points. Parameters have been chosen in the admissible range of ergodicity and so as to avoid too flat curves that would not allow to see the estimation performance. The sample size $n = 5000$ and the step $\Delta = 1/20$ are in accordance with the asymptotic context (great $n$'s and small $\Delta$'s) and may be relevant for applications in finance. It clearly appears that the estimated functions stick very well to the true ones.

The simulation algorithm of sample paths does not rely on Euler Schemes as in the estimation method. Therefore, the data simulation method is disconnected with the estimation procedures and cannot be suspected of being favorable to our estimation algorithm.

As a conclusion, the algorithms proposed here provide a complete tool for nonparametric estimation of drift and diffusion coefficients of diffusion processes. An analogous algorithm (but based on a projection contrast) may allow to estimate the stationary density $\pi$ (see e.g. Lacour (2005)).

## 6. Proofs

### 6.1. **Proof of Proposition 3.1.**

Starting from (13)-(14), we obtain

$$\begin{aligned}
\|\hat{b}_m - b_A\|_n^2 &\leq \|b_m - b_A\|_n^2 + 2\|\hat{b}_m - b_m\| \sup_{t\in S_m, \|t\|=1} |\nu_n(t)| \\
&\quad +2\|\hat{b}_m - b_m\|_n \sqrt{\frac{1}{n\Delta^2}\sum_{k=1}^{n}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2} \\
&\leq \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_m - b_m\|^2 + 8\sup_{t\in S_m, \|t\|=1}[\nu_n(t)]^2 \\
&\quad +\frac{1}{8}\|\hat{b}_m - b_m\|_n^2 + \frac{8}{n\Delta^2}\sum_{k=1}^{n}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2
\end{aligned}$$

Because the $\mathbb{L}^2$-norm, $\|.\|$, and the empirical norm (12) are not equivalent, we must introduce a set on which they are and afterwards, prove that this set has small probability. Let us define

$$\Omega_n = \left\{\omega / \left|\frac{\|t\|_n^2}{\|t\|^2} - 1\right| \leq \frac{1}{2}, \ \ \forall t \in \cup_{m,m'\in\mathcal{M}_n}(S_m + S_{m'})/\{0\}\right\}, \tag{37}$$

see (6). On $\Omega_n$, $\|\hat{b}_m - b_m\|^2 \leq 2\|\hat{b}_m - b_m\|_n^2$, and $\|\hat{b}_m - b_m\|_n^2 \leq 2(\|\hat{b}_m - b_A\|_n^2 + \|b_m - b_A\|_n^2)$. Hence, some elementary computations yield:

$$\frac{1}{4}\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Omega_n} \leq \frac{7}{4}\|b_m - b_A\|_n^2 + 8\sup_{t\in S_m, \|t\|=1}[\nu_n(t)]^2 + \frac{8}{n\Delta^2}\sum_{k=1}^{n}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2$$

Now,

$$\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2 \leq \Delta\int_{k\Delta}^{(k+1)\Delta}\mathbb{E}[(b(X_s)-b(X_{k\Delta}))^2]ds.$$

Then [A1]$(i)$ and (3) imply that

$$\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2 \ \leq \ \Delta\int_{k\Delta}^{(k+1)\Delta}c'\Delta ds \leq c'\Delta^3.$$

Consequently,

$$\mathbb{E}(\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Omega_n}) \leq 7\|b_m - b_A\|_\pi^2 + 32\,\mathbb{E}\left(\sup_{t\in S_m, \|t\|=1}[\nu_n(t)]^2\right) + 32c'\Delta. \tag{38}$$

Next, using (7)-(8)-(9)-(13), it is easy to see that

$$\begin{aligned}
\mathbb{E}\left(\sup_{t\in S_m,\|t\|=1}[\nu_n(t)]^2\right) &\leq \sum_{\lambda\in\Lambda_m}\mathbb{E}[\nu_n^2(\varphi_\lambda)]\\
&= \frac{1}{n^2\Delta^2}\sum_{k=1}^n\mathbb{E}\left[\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s)ds\right]\\
&\leq \frac{\Phi_0^2D_m}{n^2\Delta^2}\sum_{k=1}^n\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s)ds\right]\\
&= \frac{\Phi_0^2D_m}{n\Delta^2}\mathbb{E}\left(\int_0^\Delta\sigma^2(X_s)ds\right)=\frac{\Phi_0^2\mathbb{E}(\sigma^2(X_0))D_m}{n\Delta}.
\end{aligned}$$

Gathering bounds, and using the upper bound $\pi_1$ defined in (5), we get

$$\mathbb{E}(\|\hat{b}_m-b_A\|_n^2\mathbf{1}_{\Omega_n})\leq 7\pi_1\|b_m-b_A\|^2+32\frac{\Phi_0^2\mathbb{E}(\sigma^2(X_0))D_m}{n\Delta}+32c'\Delta.$$

Now, it remains to deal with $\Omega_n^c$. Since $\|\hat{b}_m-b_A\|_n^2\leq\|\hat{b}_m-b\|_n^2$, it is enough to check that $\mathbb{E}(\|\hat{b}_m-b\|_n^2\mathbf{1}_{\Omega_n^c})\leq c/n$. Write the regression model as $Y_{k\Delta}=b(X_{k\Delta})+\varepsilon_{k\Delta}$ with

$$\varepsilon_{k\Delta}=\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}[b(X_s)-b(X_{k\Delta})]ds+\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)dW_s.$$

Recall that $\Pi_m$ denotes the orthogonal projection (with respect to the inner product of $\mathbb{R}^n$) onto the subspace $\{(t(X_\Delta),\ldots,t(X_{n\Delta}))',t\in S_m\}$ of $\mathbb{R}^n$. We have $(\hat{b}_m(X_\Delta),\ldots,b_m(X_{n\Delta}))'=\Pi_mY$ where $Y=(Y_\Delta,\ldots,Y_{n\Delta})'$. Using the same notation for the function $t$ and the vector $(t(X_\Delta),\ldots,t(X_{n\Delta}))'$, we see that

$$\|b-\hat{b}_m\|_n^2=\|b-\Pi_mb\|_n^2+\|\Pi_m\varepsilon\|_n^2\leq\|b\|_n^2+n^{-1}\sum_{i=1}^n\varepsilon_{i\Delta}^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|b-\hat{b}_m\|_n^2\mathbf{1}_{\Omega_n^c}\right] &\leq \mathbb{E}\left(\|b\|_n^2\mathbf{1}_{\Omega_n^c}\right)+\frac{1}{n}\sum_{k=1}^n\mathbb{E}\left[\varepsilon_{k\Delta}^2\mathbf{1}_{\Omega_n^c}\right]\\
&\leq \left(\mathbb{E}^{1/2}(b^4(X_0))+\mathbb{E}^{1/2}(\varepsilon_\Delta^4)\right)\mathbb{P}^{1/2}(\Omega_n^c).
\end{aligned}$$

By [A1]$(ii)$, we have $\mathbb{E}(b^4(X_0))\leq c(1+\mathbb{E}(X_0^4))=K$. With the Burholder-Davis-Gundy inequality, we find

$$\mathbb{E}(\varepsilon_\Delta^4)\leq 2^3\left(\frac{1}{\Delta}\int_0^\Delta\mathbb{E}[(b(X_s)-b(X_\Delta))^4]ds+\frac{36}{\Delta^3}\mathbb{E}\left(\int_0^\Delta\sigma^4(X_s)ds\right)\right).$$

Under [A1]-[A2]$(i)$-[A3] and (3), we obtain $\mathbb{E}(\varepsilon_\Delta^4)\leq C(1+\sigma_1^4/\Delta^2):=C'/\Delta^2$. The next lemma enables us to complete the proof.

**Lemma 6.1.** *Let $\Omega_n$ be defined by (37) and assume that $n\Delta_n/\ln^2(n) \to +\infty$ when $n \to +\infty$. Then, if $N_n \le O(n\Delta_n/\ln^2(n))$ for collections* [DP] *and* [W]*, and if $N_n \le O(\sqrt{n\Delta_n}/\ln(n))$ for collection* [T],

$$\mathbb{P}(\Omega_n^c) \le \frac{c}{n^4}. \tag{39}$$

Now, we gather all terms and use (39) to get (15). □

Proof of Lemma 6.1. It is proved in Baraud et al. (2001a), that

$$\mathbb{P}(\Omega_n^c) \le 2n\beta_X(q_n\Delta) + 2n\exp(-C_0\frac{n}{q_nL_n(\phi)})$$

where $C_0$ is a constant depending on $\pi_0, \pi_1$, $q_n$ is an integer such that $q_n < n$ and $L_n(\phi)$ is a quantity depending on the basis of the largest nesting space $\mathcal{S}_n$ of the collection. This space has dimension denoted by $N_n = \dim(\mathcal{S}_n)$. For [T], $L_n(\phi) \le C_\phi N_n^2$. For [W] and [DP] (see Sections 2.2 and 2.3), $L_n(\phi) \le C'_\phi N_n$.
By assumption, the diffusion process $X$ is geometrically $\beta$-mixing. So, for some constant $\theta$,

$$\beta_X(q_n\Delta) \le e^{-\theta q_n\Delta}.$$

Provided that $\Delta = \Delta_n$ satisfies $\ln(n)/(n\Delta) \to 0$, it is possible to take $q_n = [5\ln(n)/(\theta\Delta)] + 1$. This yields

$$\mathbb{P}(\Omega_n^c) \le \frac{2}{n^4} + 2n\exp(-C'_0\frac{n\Delta}{\ln(n)N_n}).$$

The above constraint on $\Delta$ must be strengthened. Indeed, to ensure (39), we need that

$$\frac{n\Delta}{N_n} \ge \frac{5\ln^2(n)}{C'_0} \quad \textit{i.e.} \quad N_n \le \tilde{C}_0\frac{n\Delta}{\ln^2(n)}$$

for [W] and [DP] . This requires $n\Delta/\ln^2(n) \to +\infty$. The result for [T] follows analogously. □

6.2. **Proof of Theorem 3.1.** The proof relies on the following Bernstein-type inequality:

**Lemma 6.2.** *Under the assumptions of Theorem 3.1, for any positive numbers $\epsilon$ and $v$, we have*

$$\mathbb{P}\left[\sum_{k=1}^{n} t(X_{k\Delta})Z_{k\Delta} \ge n\epsilon, \|t\|_n^2 \le v^2\right] \le \exp\left(-\frac{n\Delta\epsilon^2}{2\sigma_1^2v^2}\right).$$

Proof of Lemma 6.2: We use that $\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta}$ can be written as a stochastic integral. Consider the process:

$$H_u^n = H_u = \sum_{k=1}^{n} \mathbb{1}_{[k\Delta,(k+1)\Delta[}(u)t(X_{k\Delta})\sigma(X_u)$$

which satisfies $H_u^2 \le \sigma_1^2\|t\|_\infty^2$ for all $u \ge 0$. Then, denoting by $M_s = \int_0^s H_u dW_u$, we get that

$$M_{(n+1)\Delta} = \sum_{k=1}^{n} t(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta} \sigma(X_s)dW_s, \quad \langle M\rangle_{(n+1)\Delta} = \sum_{k=1}^{n} t^2(X_{k\Delta})\int_{k\Delta}^{(k+1)\Delta} \sigma^2(X_s)ds.$$

Moreover, $\langle M\rangle_s = \int_0^s H_u^2 du \le n\sigma_1^2\Delta\|t\|_n^2$, $\forall s \ge 0$, so that $(M_s)$ and $\exp(\lambda M_s - \lambda^2\langle M\rangle_s/2)$ are martingales with respect to the filtration $\mathcal{F}_s = \sigma(X_u, u \le s)$. Therefore, for all $s \ge 0$, $c > 0$, $d > 0$, $\lambda > 0$,

$$\mathbb{P}(M_s \ge c, \langle M\rangle_s \le d) \le \mathbb{P}\left(e^{\lambda M_s - \frac{\lambda^2}{2}\langle M\rangle_s} \ge e^{\lambda c - \frac{\lambda^2}{2}d}\right) \le e^{-(\lambda c - \frac{\lambda^2}{2}d)}.$$

Therefore,

$$\mathbb{P}(M_s \ge c, \langle M\rangle_s \le d) \le \inf_{\lambda>0} e^{-(\lambda c - \frac{\lambda^2}{2}d)} = e^{-\frac{c^2}{2d}}.$$

Finally,

$$\begin{aligned}\mathbb{P}\left[\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta} \ge n\epsilon, \|t\|_n^2 \le v^2\right] &= \mathbb{P}(M_{(n+1)\Delta} \ge n\Delta\epsilon, \langle M\rangle_{(n+1)\Delta} \le nv^2\sigma_1^2\Delta) \\ &\le \exp\left(-\frac{(n\Delta\epsilon)^2}{2nv^2\sigma_1^2\Delta}\right) = \exp\left(-\frac{n\epsilon^2\Delta}{2v^2\sigma_1^2}\right). \quad \square\end{aligned}$$

Now we turn to the proof of Theorem 3.1. As in the proof of Proposition 3.1, we have to split $\|\hat{b}_{\hat{m}} - b_A\|_n^2 = \|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{I}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{I}_{\Omega_n^c}$. For the study on $\Omega_n^c$, the end of the proof of Proposition 3.1 can be used.

Now, we focus on what happens on $\Omega_n$. From the definition of $\hat{b}_{\hat{m}}$, we have, $\forall m \in \mathcal{M}_n$, $\gamma_n(\hat{b}_{\hat{m}}) + \text{pen}(\hat{m}) \le \gamma_n(b_m) + \text{pen}(m)$. We proceed as in the proof of Proposition 3.1 with the additional penalty terms (see (38)) and obtain

$$\begin{aligned}\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{I}_{\Omega_n}) &\le 7\pi_1\|b_m - b_A\|^2 + 4\text{pen}(m) + 32\mathbb{E}\left(\sup_{t\in S_m + S_{\hat{m}}, \|t\|=1} [\nu_n(t)]^2 \mathbf{I}_{\Omega_n}\right) \\ &\quad -4\mathbb{E}(\text{pen}(\hat{m})) + 32c'\Delta.\end{aligned}$$

The main problem here is to control the supremum of $\nu_n(t)$ on a random ball (which depends on the random $\hat{m}$). This is done by using the martingale property of $\nu_n(t)$.

Let us introduce the notation

$$G_m(m') = \sup_{t\in S_m + S_{m'}, \|t\|=1} \nu_n(t).$$

Now, we plug in a function $p(m, m')$, which will in turn fix the penalty:

$$\begin{aligned}G_m^2(\hat{m})\mathbf{I}_{\Omega_n} &\le [(G_m^2(\hat{m}) - p(m,\hat{m}))\mathbf{I}_{\Omega_n}]_+ + p(m,\hat{m}) \\ &\le \sum_{m'\in\mathcal{M}_n} [(G_m^2(m') - p(m,m'))\mathbf{I}_{\Omega_n}]_+ + p(m,\hat{m}).\end{aligned}$$

And pen is chosen such that $8p(m, m') \le \text{pen}(m) + \text{pen}(m')$. More precisely, the next proposition determines the choice of $p(m, m')$.

**Proposition 6.1.** *Under the assumptions of Theorem 3.1, there exists a numerical constant $\kappa_1$ such that, for $p(m, m') = \kappa_1\sigma_1^2(D_m + D_{m'})/(n\Delta)$, we have*

$$\mathbb{E}[(G_m^2(m') - p(m,m'))\mathbf{I}_{\Omega_n}]_+ \le c\sigma_1^2\frac{e^{-D_{m'}}}{n\Delta}.$$

Proof of Proposition 6.1. The result of Proposition 6.1 follows from the inequality of Lemma 6.2 by the $\mathbb{L}^2$-chaining technique used in Baraud et al. (2001b) (see Section 7 p.44-47, Lemma 7.1, with $s^2 = \sigma_1^2/\Delta$). □

It is easy to see that the result of Theorem 3.1 follows from Proposition 6.1 with $\mathrm{pen}(m) = \kappa\sigma_1^2 D_m/(n\Delta)$ and $\kappa = 8\kappa_1$. □

6.3. **Proof of Proposition 4.1.** First, we prove that

$$\mathbb{E}(\frac{1}{n}\sum_{k=1}^{n} R_{k\Delta}^2) \leq K\Delta^2. \tag{40}$$

Proof of (40). With obvious convention, let $R_{k\Delta} = R_{k\Delta}^{(1)} + R_{k\Delta}^{(2)} + R_{k\Delta}^{(3)}$ so that (40) holds if $\mathbb{E}[(R_{k\Delta}^{(i)})^2] \leq K_i\Delta^2$ for $i = 1, 2, 3$. Using [A1],

$$\begin{aligned}\mathbb{E}[(R_{k\Delta}^{(1)})^2] &\leq \mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta} b^2(X_s)ds\right)^2 \leq \Delta\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta} b^4(X_s)ds\right)\\ &\leq \Delta^2\mathbb{E}(b^4(X_0)) \leq c\Delta^2.\end{aligned}$$

$$\mathbb{E}[(R_{k\Delta}^{(2)})^2] \leq \frac{1}{\Delta^2}\left(\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s) - b(X_{k\Delta}))ds\right)^2 \mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)dW_s\right)^2\right)^{1/2}$$

Both terms have already been studied. Using (3), we get

$$\mathbb{E}[(R_{k\Delta}^{(2)})^2] \leq c'\Delta^2.$$

Lastly, using [A1]-[A2] and (22),

$$\mathbb{E}[(R_{k\Delta}^{(3)})^2] \leq \frac{1}{\Delta}\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}[(k+1)\Delta - s]^2\psi^2(X_s)ds\right) \leq \mathbb{E}(\psi^2(X_0))\frac{\Delta^2}{3} \leq c"\Delta^2.$$

Therefore (40) is proved. □

Now, we turn back to (23) and recall that $\Omega_n$ is defined by (37). The study is close to the one done for the drift estimator. On $\Omega_n$, $\|\hat{\sigma}_m^2 - \sigma_m^2\|^2 \leq 2\|\hat{\sigma}_m^2 - \sigma_m^2\|_n^2$,

$$\begin{aligned}\mathbb{E}(\|\hat{\sigma}_m^2 - \sigma_A^2\|_n^2 &\leq \|\sigma_m^2 - \sigma_A^2\|_n^2 + \frac{1}{8}\|\hat{\sigma}_m^2 - \sigma_m^2\|^2 + 8\sup_{t\in S_m, \|t\|=1}\breve{\nu}_n^2(t)\\ &\quad +\frac{1}{8}\|\hat{\sigma}_m^2 - \hat{\sigma}_m^2\|_n^2 + \frac{8}{n}\sum_{k=1}^{n} R_{k\Delta}^2\\ &\leq \|\sigma_m^2 - \sigma^2\|_n^2 + \frac{3}{8}\|\hat{\sigma}_m^2 - \sigma_m^2\|_n^2 + 8\sup_{t\in S_m, \|t\|=1}\breve{\nu}_n^2(t) + \frac{8}{n}\sum_{k=1}^{n} R_{k\Delta}^2.\end{aligned}$$

Setting $B_m(0,1) = \{t \in S_m, \|t\| = 1\}$, the following holds on $\Omega_n$:

$$\frac{1}{4}\|\hat{\sigma}_m^2 - \sigma_A^2\|_n^2 \leq \frac{7}{4}\|\sigma_m^2 - \sigma_A^2\|_n^2 + 8\sup_{t\in B_m(0,1)}\breve{\nu}_n^2(t) + \frac{8}{n}\sum_{k=1}^{n} R_{k\Delta}^2.$$

Moreover

$$\begin{aligned}\mathbb{E}(\sup_{t\in B_m(0,1)} \breve{\nu}_n^2(t)) &\leq \sum_{\lambda\in\Lambda_m} \mathbb{E}(\breve{\nu}_n^2(\varphi_\lambda)) = \frac{1}{n^2}\sum_{\lambda\in\Lambda_m} \mathbb{E}\left(\sum_{k=1}^n \varphi_\lambda^2(X_{k\Delta})V_{k\Delta}^2\right)\\ &\leq \frac{\Phi_0^2 D_m}{n}[12\mathbb{E}(\sigma^4(X_0)+4\Delta C_{b,\sigma})])\end{aligned}$$

where $C_{b,\sigma} = \mathbb{E}((\sigma'\sigma^2)^2(X_0)) + \sigma_1^2\mathbb{E}(b^2(X_0))$. Now using the condition on $N_n$, we have $\Delta D_m/n \leq \Delta N_n/n \leq \Delta^2/\ln^2(n)$. This yields the first three terms of the right-hand-side of (24).
The study on $\Omega_n^c$ is the same as for $b$ with the regression model $U_{k\Delta} = \sigma^2(X_{k\Delta}) + \eta_{k\Delta}$, where $\eta_{k\Delta} = V_{k\Delta} + R_{k\Delta}$. By standard inequalities, $\mathbb{E}(\eta_\Delta^4) \leq K[\Delta^4\mathbb{E}(b^8(X_0)) + \mathbb{E}(\sigma^8(X_0))]$. Hence, $\mathbb{E}(\eta_\Delta^4)$ is bounded. Moreover, using Lemma 6.1, $\mathbb{P}(\Omega_n^c) \leq c/n^2$.□

6.4. **Proof of Theorem 4.1.** This proof follows the same lines as the proof of Theorem 3.1. We start with a Bernstein-type inequality.

**Lemma 6.3.** *Under the assumptions of Theorem 4.1,*

$$\mathbb{P}\left(\sum_{k=1}^n t(X_{k\Delta})V_{k\Delta}^{(1)} \geq n\epsilon, \|t\|_n^2 \leq v^2\right) \leq \exp\left(-Cn\frac{\epsilon^2/2}{2\sigma_1^4 v^2 + \epsilon\|t\|_\infty \sigma_1^2 v}\right)$$

*and*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})V_{k\Delta}^{(1)} \geq v\sigma_1^2\sqrt{2x} + \sigma_1^2\|t\|_\infty x, \|t\|_n^2 \leq v^2\right) \leq \exp(-Cnx). \tag{41}$$

The non trivial link between the above two inequalities is enhanced by Birgé and Massart (1998) so that we just prove the first one.

Proof of Lemma 6.3. First we note that:

$$\begin{aligned}\mathbb{E}\left(e^{ut(X_{n\Delta})V_{n\Delta}^{(1)}}|\mathcal{F}_{n\Delta}\right) &= 1 + \sum_{p=2}^{+\infty}\frac{u^p}{p!}\mathbb{E}\left[(t(X_n\Delta)V_{n\Delta}^{(1)})^p|\mathcal{F}_{n\Delta}\right]\\ &\leq 1 + \sum_{p=2}^{+\infty}\frac{u^p}{p!}|t(X_{n\Delta})|^p\mathbb{E}\left[|V_{n\Delta}^{(1)}|^p|\mathcal{F}_{n\Delta}\right].\end{aligned}$$

Next we apply successively the Hölder inequality and the Burkholder-Davis-Gundy inequality with best constant (Proposition 4.2 of Barlow and Yor (1982)): For a continuous martingale $(M_t)$, with $M_0 = 0$, for $k \geq 2$, $M_t^* = \sup_{s\leq t}|M_s|$ satisfies $\|M^*\|_k \leq c\sqrt{k}\|\|\langle M\rangle^{1/2}\|_k$, with $c$ a universal constant. And we obtain:

$$\begin{aligned}\mathbb{E}(|V_{n\Delta}^{(1)}|^p|\mathcal{F}_{n\Delta}) &\leq \frac{2^{p-1}}{\Delta^p}\left\{\mathbb{E}\left(\left|\int_{n\Delta}^{(n+1)\Delta}\sigma(X_s)dW_s\right|^{2p}|\mathcal{F}_{n\Delta}\right) + \mathbb{E}\left(\left|\int_{n\Delta}^{(n+1)\Delta}\sigma^2(X_s)ds\right|^p|\mathcal{F}_{n\Delta}\right)\right\}\\ &\leq \frac{2^{p-1}}{\Delta^p}(c^{2p}(2p)^p\Delta^p\sigma_1^{2p} + \Delta^p\sigma_1^{2p}) \leq (2\sigma_1 c)^{2p}p^p.\end{aligned}$$

Therefore,

$$\mathbb{E}\left(e^{ut(X_{n\Delta})V_{n\Delta}^{(1)}}|\mathcal{F}_{n\Delta}\right) \leq 1+\sum_{k=2}^{\infty}\frac{p^p}{p!}(4u\sigma_1^2c^2)^p|t(X_{n\Delta})|^p.$$

Using $p^p/p! \leq e^{p-1}$, we find

$$\begin{aligned}\mathbb{E}\left(e^{ut(X_{n\Delta})V_{n\Delta}^{(1)}}|\mathcal{F}_{n\Delta}\right) &\leq 1+e^{-1}\sum_{k=2}^{\infty}(4u\sigma_1^2c^2e)^p|t(X_{n\Delta})|^p \\ &\leq 1+e^{-1}\frac{(4u\sigma_1^2c^2e)^2t^2(X_{n\Delta})}{1-(4u\sigma_1^2c^2e\|t\|_\infty)}.\end{aligned}$$

Now, let us set

$$a = e(4\sigma_1^2c^2)^2 \text{ and } b = 4\sigma_1^2c^2e\|t\|_\infty.$$

Since for $x \geq 0$, $1+x \leq e^x$, we get, for all $u$ such that $bu < 1$,

$$\mathbb{E}\left(e^{ut(X_{n\Delta})V_{n\Delta}^{(1)}}|\mathcal{F}_{n\Delta}\right) \leq 1+\frac{au^2t^2(X_{n\Delta})}{1-bu} \leq \exp\left(\frac{au^2t^2(X_{n\Delta})}{1-bu}\right).$$

This can also be written:

$$\mathbb{E}\left(\exp\left(ut(X_{n\Delta})V_{n\Delta}^{(1)}-\frac{au^2t^2(X_{n\Delta})}{1-bu}\right)|\mathcal{F}_{n\Delta}\right) \leq 1.$$

Therefore, iterating conditional expectations yields

$$\mathbb{E}\left\{\exp\left[\sum_{k=1}^{n}\left(ut(X_{k\Delta})V_{k\Delta}^{(1)}-\frac{au^2t^2(X_{k\Delta})}{1-bu}\right)\right]\right\} \leq 1.$$

Then, we deduce that

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{n}t(X_{k\Delta})V_{k\Delta}^{(1)} \geq n\epsilon, \|t\|_n^2 \leq v^2\right) &\leq e^{-nu\epsilon}\mathbb{E}\left[\mathbf{1}_{\|t\|_n^2\leq v^2}\exp\left(u\sum_{k=1}^{n}t(X_{k\Delta})V_{k\Delta}^{(1)}\right)\right] \\ \leq\ & e^{-nu\epsilon}\mathbb{E}\left[\mathbf{1}_{\|t\|_n^2\leq v^2}\exp\left(\sum_{k=1}^{n}(ut(X_{k\Delta})V_{k\Delta}^{(1)}-\frac{au^2t^2(X_{k\Delta})}{1-bu})\right)e^{(au^2)/(1-bu)\sum_{k=1}^{n}t^2(X_{k\Delta})}\right] \\ \leq\ & e^{-nu\epsilon}e^{(nau^2v^2)/(1-bu)}\mathbb{E}\left[\exp\left(\sum_{k=1}^{n}(ut(X_{k\Delta})V_{k\Delta}^{(1)}-\frac{au^2t^2(X_{k\Delta})}{1-bu})\right)\right] \\ \leq\ & e^{-nu\epsilon}e^{(nau^2v^2)/(1-bu)}.\end{aligned}$$

The inequality holds for any $u$ such that $bu < 1$. In particular, $u = \epsilon/(2av^2+\epsilon b)$ gives $-u\epsilon + av^2u^2/(1-bu) = -(1/2)(\epsilon^2/(2av^2+\epsilon b)$ and therefore

$$\mathbb{P}\left(\sum_{k=1}^{n}t(X_{k\Delta})V_{k\Delta}^{(1)} \geq n\epsilon, \|t\|_n^2 \leq v^2\right) \leq \exp\left(-n\frac{\epsilon^2/2}{2av^2+\epsilon b}\right). \ \square$$

As for $\hat{b}_{\hat{m}}$, we introduce the additional penalty terms and obtain that the risk satifies

$$\begin{aligned}\mathbb{E}(\|\hat{\sigma}_{\hat{m}}^2-\sigma^2\|_n^2\mathbf{1}_{\Omega_n}) &\leq 7\pi_1\|\sigma_m^2-\sigma^2\|^2+4\widetilde{\text{pen}}(m)+32\mathbb{E}\left(\sup_{t\in B_{m,\hat{m}}(0,1)}[\breve{\nu}_n(t)]^2\mathbf{1}_{\Omega_n}\right) \\ &\quad -4\widetilde{\text{pen}}(\hat{m})+K'\Delta^2 \qquad (42)\end{aligned}$$

where $B_{m,m'}(0,1)=\{t\in S_m+S_{m'}, \|t\|=1\}$. Let us denote by

$$\breve{G}_m(m')=\sup_{t\in S_m+S_{m'},\|t\|=1}\breve{\nu}_n^{(1)}(t)$$

the main quantity to be studied, where

$$\breve{\nu}_n^{(1)}(t)=\frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})V_{k\Delta}^{(1)},\ \ \breve{\nu}_n^{(2)}(t)=\frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})(V_{k\Delta}^{(2)}+V_{k\Delta}^{(3)}).$$

As for the drift, we write

$$\begin{aligned}\mathbb{E}(\breve{G}_m^2(\hat{m})) &\leq \mathbb{E}[(\breve{G}_m^2(\hat{m})-\tilde{p}(m,\hat{m}))\mathbf{I}_{\Omega_n}]_+ +\tilde{p}(m,\hat{m})\\ &\leq \sum_{m'\in\mathcal{M}_n}\mathbb{E}[(\breve{G}_m^2(m')-\tilde{p}(m,m'))\mathbf{I}_{\Omega_n}]_+ +\tilde{p}(m,\hat{m}).\end{aligned}$$

Then $\widetilde{\mathrm{pen}}$ is chosen such that $8\tilde{p}(m,m')\leq\widetilde{\mathrm{pen}}(m)+\widetilde{\mathrm{pen}}(m')$. More precisely, we can prove

**Proposition 6.2.** *Under the assumptions of Theorem 4.1, for $\tilde{p}(m,m')=\kappa^*\sigma_1^4(D_m+D_{m'})/n$, where $\kappa^*$ is a numerical constant, we have*

$$\mathbb{E}[(\breve{G}_m^2(m')-\tilde{p}(m,m'))\mathbf{I}_{\Omega_n}]_+\leq c\sigma_1^4\frac{e^{-D_{m'}}}{n}.$$

The result of Proposition 6.2 is obtained from inequality (41) of Lemma 6.3 by a $L^2-L^\infty$ chaining technique. For a description of this method, in a more general setting, we refer to Proposition 2-4 pp.282-287 in Comte (2001), to Theorem 5 in Birgé and Massart (1998) and to Proposition 7, Theorem 8 and Theorem 9 in Barron et al. (1999).
Choosing $\widetilde{\mathrm{pen}}(m)=\tilde{\kappa}\sigma_1^4D_m/n$ with $\tilde{\kappa}=8\kappa^*$, we deduce from (42) and Proposition 6.2 that,

$$\begin{aligned}\mathbb{E}(\|\hat{\sigma}_{\hat{m}}^2-\sigma^2\|_n^2) \ &\leq\ 7\pi_1\|\sigma_m^2-\sigma^2\|^2+8\tilde{\kappa}\sigma_1^4\frac{D_m}{n}+c\sigma_1^4\sum_{m'\in\mathcal{M}_n}\frac{e^{-D_{m'}}}{n}\\ &\quad +64\mathbb{E}\left(\sup_{t\in B_{m,\hat{m}}(0,1)}[\breve{\nu}_n^{(2)}(t)]^2\right)+K'\Delta^2+\mathbb{E}(\|\hat{\sigma}_{\hat{m}}^2-\sigma^2\|_n^2\mathbf{I}_{\Omega_n^c}).\end{aligned}$$

The bound for $\mathbb{E}(\|\hat{\sigma}_{\hat{m}}^2-\sigma^2\|_n^2\mathbf{I}_{\Omega_n^c})$ is the same as the one given in the end of the proof of Proposition 4.1. It is less than $c/n$ provided that $N_n\leq n\Delta/\ln^2(n)$ for [DP] and [W] and $N_n^2\leq n\Delta/\ln^2(n)$ for [T].
And since the spaces are nested and all contained in a space denoted by $\mathcal{S}_n$ with dimension $N_n$ bounded as right above, we have

$$\mathbb{E}\left(\sup_{t\in B_{m,\hat{m}}(0,1)}[\breve{\nu}_n^{(2)}(t)]^2\right)\leq\mathbb{E}\left(\sup_{t\in\mathcal{S}_n,\|t\|=1}[\breve{\nu}_n^{(2)}(t)]^2\right)\leq KC_{b,\sigma}\Phi_0^2\frac{\Delta N_n}{n}\leq K'\Delta^2.$$

The result of Theorem 4.1 follows. □

## 7. Appendix

In this section, we briefly recall the steps of the retrospective exact simulation algorithm proposed by Beskos et al (2004) in the simplest case. Consider the one-dimensional diffusion process given by

$$d\xi_t^x = \alpha(\xi_t^x)dt + dW_t, \xi_0^x = x. \tag{43}$$

where $\alpha(.)$ is a real function defined on $\mathbb{R}$ and $(W_t)$ is a standard Wiener process. It is possible to produce an exact realization of the random variable $\xi_\Delta^x$ for $\Delta > 0$ under the following assumptions :

- the function $\alpha(.)$ is $C^1$ and there exist constants $c_1, c_2$ such that $c_1 \leq \alpha^2 + \alpha' \leq c_2$.
- Let $A(\xi) = \int_0^\xi \alpha(u)du$. The function $h(\xi) = \exp\left(A(\xi) - (\xi - x)^2/2\Delta\right)$ is integrable on $\mathbb{R}$.
- It is possible to simulate an exact realization of a random variable with density proportional to $h$.

Set $\Phi(.) = \frac{1}{2}(\alpha^2(.) + \alpha'(.) - c_1)$ and $M = (c_2 - c_1)/2$ so that $0 \leq \Phi \leq M$. The following preliminaries are required.

Let $C_\Delta = \{\omega : [0, \Delta] \to \mathbb{R};\ \omega \text{ continuous}\}$ be the space on continuous real functions defined on $[0, \Delta]$, endowed with the Borel $\sigma$-field $\mathcal{C}_\Delta$ associated with the topology of uniform convergence on $[0, \Delta]$. Denote by $X = (X_t, 0 \leq t \leq \Delta)$ the canonical coordinate process of $C_\Delta$ defined by $X_t(\omega) = \omega(t)$. Let $P_\Delta^x$ denote the distribution on $(C_\Delta, \mathcal{C}_\Delta)$ of $(\xi_t^x, 0 \leq t \leq \Delta)$. And let $W_\Delta^x$ denote the distribution on $C_\Delta$ of $(x + W_t, 0 \leq t \leq \Delta)$. Now, the conditional distribution $W_\Delta^x(.|X_\Delta = y)$ on $C_\Delta$ is the distribution of a Brownian bridge started at $x$ ending at $y$, *i.e.* the distribution of $(x + \frac{t}{\Delta}(y - x) + W_t - \frac{t}{\Delta}W_\Delta, t \in [0, \Delta])$. Finally, define the probability $Z_\Delta^x$ on $C_\Delta$ as follows:

- Under $Z_\Delta^x$, $X_\Delta$ has density proportional to $h(.)$,
- $Z_\Delta^x(.|X_\Delta = y) = W_\Delta^x(.|X_\Delta = y)$.

Then, the following results hold.

**Proposition 7.1.** *The probability $P_\Delta^x$ is absolutely continuous w.r.t. $Z_\Delta^x$ with*

$$\frac{dP_\Delta^x}{dZ_\Delta^x} \propto \exp\left(-\int_0^\Delta \Phi(X_s)ds\right).$$

**Proposition 7.2.** *Assume that $\tau$ is a Poisson random variable with parameter $\Delta M$ and that, given $\tau = k$, $((T_i, V_i), i = 1, \ldots, k)$ are i.i.d. random variables uniformly distributed on $[0, \Delta] \times [0, M]$. Then*

$$N(ds, dv) = \sum_{i=1}^{\tau} \delta_{(T_i,V_i)}(ds, dv)\ 1_{\tau \geq 1}$$

*is a random Poisson measure with intensity $\lambda(ds, dv) = 1_{[0,\Delta]}(s)1_{[0,M]}(v)dsdv$ ($\delta_z$ denotes the Dirac measure at point $z$).*

*Let $B(X) = \{(s, v); 0 \leq s \leq \Delta, 0 \leq v \leq \Phi(X_s)\}$. Then,*

$$P(N(B(X)) = 0|X) = \exp\left(-\int_0^\Delta \Phi(X_s)ds\right).$$

We recall now the rejection method of simulation.

**Proposition 7.3.** *On the space $(S, \mathcal{S})$, let $\mu$ and $\nu$ be two probability measures (where we are able to simulate $\nu$). Assume that $\frac{d\mu}{d\nu} = \frac{1}{\varepsilon} f$ where $f \leq 1$. Let $((Y_n, I_n), n \geq 1)$ be a sequence of i.i.d. random variables with values in $S \times \{0,1\}$ and such that $Y_i$ has distribution $\nu$ and $P(I_i = 1 | Y_i = y) = f(y)$, $y \in S$. Let $\kappa = \min\{i \geq 1, I_i = 1\}$. Then, $P(\kappa < \infty) = 1$ and $Y_\kappa$ has distribution $\mu$.*

Finally, we describe the retrospective simulation algorithm for $\xi^x_\Delta$.

- (step 1) Simulate $X_\Delta = y$ according to a density proportional to $h$, (*e.g.* using Proposition 7.3)
- (step 2) Simulate $\tau = k$, and $((T_i, V_i) = (t_i, v_i), i = 1, \dots, k)$ according to Proposition 7.2,
- (step 3) Simulate $(X_{t_i} = x_i, i = 1, \dots, k)$ according to a Brownian bridge started at $x$ at time 0 ending at $y$ at time $\Delta$, at the time instants prescribed by step 2,
- (step 4) Compute the indicator

$$I = \prod_{i=1}^{k} 1_{(\Phi(x_i) \leq v_i)}.$$

If $I = 1$, i.e. $N(B(X)) = 0$, accept the sample $X$ (in application of Proposition 7.3). Else, go back to step 1.

In the end, an exact realization $(x, X_{t_1} = x_1, \dots, X_{t_k} = x_k, X_\Delta = y)$ according to $P^x_\Delta$ is obtained. And $X_\Delta = y$ is an exact realization of $\xi^x_\Delta$.

Once a discrete trajectory is accepted, it is possible to continue constructing it between the instants $t_i$ by simulating independent Brownian bridges starting at $x_i$ at time $t_i$ ending at $x_{i+1}$ at time $t_{i+1}$. In the present paper, we do not use this device and keep only the last value $y$ as our data for the statistical estimation procedure. Then, we iterate the algorithm starting at $y$ to produce an exact sample value for the time $2\Delta$ and so on. Nevertheless, when plotting the graph of the diffusion, we use interpolation between all values of the accepted trajectory.

The assumption that $\Phi(.)$ be bounded from above is a severe restriction that can be relaxed. If $\limsup_{u \to +\infty} \Phi(u) < \infty$ or $\limsup_{u \to -\infty} \Phi(u) < \infty$, then another algorithm described in Beskos et al (2004) is possible in a very simple way.

## References

[1] Abramowitz, M. and Stegun, I.A. (1972). Handbook of mathematical functions with formulas, graphs, and mathematical tables. Edited by Milton Abramowitz and Irene A. Stegun. John Wiley and Sons, Inc., New York.

[2] Bandi, F.M. and Phillips, P.C.B. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica*, **71**, 241-283.

[3] Banon, G. (1978). Nonparametric identification for diffusion processes. *SIAM J. Control Optim.* **16**, 380-395.

[4] Baraud, Y., Comte, F. and Viennet, G. (2001a). Adaptive estimation in an autoregression and a geometrical $\beta$-mixing regression framework. *Ann. Statist.* **39**, 839-875.

[5] Baraud, Y., Comte, F. and Viennet, G. (2001b). Model selection for (auto)-regression with dependent data. *ESAIM Probab. Statist.* **5**, 33-49.

[6] Barlow, M.T. and Yor, M. (1982). Semi-martingale inequalities via the Garsia-Rodemich-Rumsey Lemma and applications to local times. *Journal of Functional Analysis* **49**, 198-229.

[7] Barron, A.R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413.

[8] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2004) Retrospective exact simulation of diffusion sample paths with applications. *Working paper of Lancaster University*, available from http://www.maths.lancs.ac.uk/ papaspil/research.html.

[9] Beskos, A. and Roberts, G.O. (2005). Exact simulation of diffusions. *Ann. Appl. Probab.* **15**, to appear.

[10] Bibby, B.M., Jacobsen, M. and Sørensen, M. (2002). Estimating functions for discretely sampled diffusion-type models. In *Handbook of Financial Econometrics.* Amsterdam: North-Holland.

[11] Bibby, B.M. and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1**, 17–39.

[12] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. Festschrift for Lucien Le Cam, Springer, New York, 55-87.

[13] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

[14] Comte, F. (2001) Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli* **7**, 267-298.

[15] Comte, F. et Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.* **97**, 111-145.

[16] Comte, F. and Rozenholc, Y. (2004) A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.* **56** 449-473.

[17] Dalalyan, A. (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33**, to appear in vol.6.

[18] Florens-Smirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *J. Appl. Probab.* **30**, 790-804. **17**, 235-239.

[19] Donoho, D.L. Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

[20] Genon-Catalot, V., Larédo, C. and Picard, D. (1992). Nonparametric estimation of the diffusion coefficient by wavelet methods. *Scand. J. Statist.* **19**, 319-335.

[21] Gloter, A. (2000). Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM Probab. Statist.* **4**, 205-227.

[22] Gobet, E., Hoffmann, M. and Reiss, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32**, 2223–2253.

[23] Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic Process. Appl.* **79**, 135-163.

[24] Jacod, J. (2000). Non-parametric Kernel Estimation of the Coefficient of a Diffusion *Scand. J. Statist.* **27**, 83-96.

[25] Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* **5**, 299-314.

[26] Kutoyants, Y.A. (2004). Statistical inference for ergodic diffusion processes. Springer Series in Statistics. Springer-Verlag London, Ltd., London.

[27] Lacour, C. (2005). Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Preprint MAP5* 2005-8, available at http://www.math-info.univ-paris5.fr/map5/publis/titres05.html

[28] Pardoux, E. and Veretennikov, A. Yu. (2001). On the Poisson equation and diffusion approximation. I. *Ann. Probab.* **29**, 3, 1061-1085.

[29] Prakasa Rao, B.L.S. (1999). Statistical inference for diffusion type processes. Kendall's Library of Statistics, 8. Edward Arnold, London; Oxford University Press, New York.

[30] Spokoiny, V.G. (2000). Adaptive drift estimation for nonparametric diffusion model. *Ann. Statist.* **28**, 815–836.