ESAIM: PS 15 (2011) 197–216
DOI: 10.1051/ps/2009016

ESAIM: Probability and Statistics
www.esaim-ps.org

PENALIZED NONPARAMETRIC DRIFT ESTIMATION FOR A CONTINUOUSLY OBSERVED ONE-DIMENSIONAL DIFFUSION PROCESS

Eva Löcherbach[1], Dasha Loukianova[2] and Oleg Loukianov[3]

Abstract. Let X be a one dimensional positive recurrent diffusion continuously observed on $[0,t]$. We consider a non parametric estimator of the drift function on a given interval. Our estimator, obtained using a penalized least square approach, belongs to a finite dimensional functional space, whose dimension is selected according to the data. The non-asymptotic risk-bound reaches the minimax optimal rate of convergence when $t \to \infty$. The main point of our work is that we do not suppose the process to be in stationary regime neither to be exponentially β-mixing. This is possible thanks to the use of a new polynomial inequality in the ergodic theorem [E. Löcherbach, D. Loukianova and O. Loukianov, *Ann. Inst. H. Poincaré Probab. Statist.* **47** (2011) 425–449].

Mathematics Subject Classification. 60F99, 60J35, 60J55, 60J60, 62G99, 62M05.

Received March 25, 2009. Revised October 6, 2009.

1. Introduction

Let X_t be a one-dimensional diffusion process given by

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x,$$

where W is a standard Brownian motion. Assuming that the process is positive recurrent but not necessarily in the stationary regime (*i.e.* not starting from the invariant measure) and not necessarily exponentially β-mixing, we want to estimate the unknown drift function b on a fixed interval K from observations of X during the time interval $[0,t]$, for fixed t. We do not require any knowledge about smoothness of the drift function: b is not supposed to belong to some known Besov or Sobolev ball. Hence we aim at studying nonparametric adaptive estimators for the unknown drift b.

Keywords and phrases. Diffusion process, adaptive estimation, regeneration method, mean square estimator, model selection, deviation inequalities.

[1] Centre de Mathématiques, Faculté de Sciences et Technologie, Université Paris-Est Val-de-Marne, 61 avenue du Général de Gaulle, 94010 Créteil, France; locherbach@univ-paris12.fr

[2] Département de Mathématiques, Université d'Evry-Val d'Essonne, Bd François Mitterrand, 91025 Evry, France; dasha.loukianova@univ-evry.fr

[3] Département Informatique, IUT de Fontainebleau, Université Paris-Est, route Hurtault, 77300 Fontainebleau, France; oleg@iut-fbleau.fr

Article published by EDP Sciences

Nonparametric estimation in continuous time of the drift coefficient of diffusion processes has been widely studied over the last decades. To mention just a few, let us cite Banon [1], Prakasa Rao [18], Pham [17], Galtchouk and Pergamenschikov [9], Dalalyan and Kutoyants [7], Delattre, Hoffmann and Kessler [8], Loukianova and Loukianov [14], Löcherbach and Loukianova [15] and the extensive book of Kutoyants [11].

The adaptive estimation for the drift at a fixed point has been studied by Spokoiny [20], who uses Lepskii's method (see [12]) in order to construct an adaptive procedure. Dalalyan [6], uses kernel-type estimators and considers a weighted L^2-risk, where the weight is given by the invariant density. He has to work under quite strong ergodicity assumptions.

Our aim in this paper is twofold. Firstly, we aim at introducing a nonparametric estimation procedure based on model selection. Our estimator is obtained by minimizing a contrast function within a fixed finite-dimensional linear sub-space of $L^2(K, \mathrm{d}x)$ – quite in the spirit of mean square estimation and following ideas presented by Comte *et al.* [5], for discretely observed diffusions. These finite-dimensional sub-spaces include spaces such as piecewise polynomials or compactly supported wavelets. The risk we consider for a given estimator $\hat{b}$ of b is the expectation of an empirical L^2-norm defined by

$$\mathbb{E}_x\|\hat{b}-b\|_t^2, \text{ where } \|\hat{b}-b\|_t^2 = \frac{1}{t}\int_0^t (\hat{b}-b)^2(X_s)\mathrm{d}s.$$

The dimension of the space is chosen by a data-driven method using a penalization.

Secondly, we aim at working under the less restrictive assumptions on the ergodicity properties of the process that seem to be possible. We do not impose the diffusion to be exponentially β-mixing and do not assume the existence of exponential moments for the invariant measure, though we do have to impose the existence of a certain number of moments. Finally, note that we do not work in the stationary regime: the process starts from a fixed point $x \in K$, and is not yet in equilibrium. Note also that our approach is non-asymptotic in time. But we have to suppose that $t \geq t_0$ for some fixed explicitly given time horizon t_0 that is needed for theoretical reasons and defined precisely later in the text (see Prop. 3.4). A main ingredient of the proofs is a new polynomial inequality ensuring that empirical norm and theoretical L^2-norm are not too far away. This inequality is given in Loukianova *et al.* [16].

The paper is organized as follows. In Section 2 we describe our framework and give the main results: in Section 2.1 we give precise assumptions on the diffusion model, explain these assumptions and give some examples for models satisfying them. In Section 2.2 we introduce both the non-adaptive and adaptive estimator, Section 2.3 gives assumptions on the approximation spaces and Section 2.4 provides some examples of approximation spaces verifying these assumptions. The main results (rate of convergence of estimators) are given in Section 2.5. Section 3 presents probabilistic tools and auxiliary results necessary for the proof of the main results. Section 4 is devoted to the proofs of the main results: Section 4.1 deals with non-adaptive and Section 4.2 with adaptive drift estimation. Finally, Section 5 is an appendix, where we give the proof of one technical result (Lem. 4.2).

2. Framework, assumptions and main results

2.1. Assumptions on the diffusion

Let X_t be a one-dimensional diffusion process given by

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x. \tag{2.1}$$

We would like to estimate the drift function b on a fixed interval K, say $K = [0, 1]$. To insure the existence and the unicity of a strong non exploding solution of (2.1) we suppose

Assumption 2.1.

1. *b and σ are locally Lipschitz and b is at most of linear growth.*

2. *There exist* $0 < \sigma_0^2 \le \sigma_1^2 < \infty$ *such that for all* x, $\sigma_0^2 \le \sigma^2(x) \le \sigma_1^2$.

A more particular assumption is needed for the drift function to guarantee some "speed" of ergodicity of X.

Assumption 2.2.

1. *There are two known constants* M_0 *and* b_0 *such that* $K \subset [-M_0, M_0]$ *and for all* x *with* $|x| \le M_0$, $|b(x)| \le b_0$.
2. *We suppose that there is a positive constant* γ *such that for all* x *with* $|x| \ge M_0$,

$$xb(x) \le -\gamma.$$

3. *The constant* γ *satisfy* $2\gamma > 31\sigma_1^2$.

To clarify the meaning of Assumptions 2.2 let us recall some well-known facts about linear diffusions. We refer the reader to the book of Revuz and Yor [19]. The scale density of X is given by

$$s(x) = \exp\left(-2\int_0^x \frac{b(u)}{\sigma^2(u)}\mathrm{d}u\right),$$

and the scale function by $S(x) = \int_0^x s(t)\mathrm{d}t$. X is recurrent if and only if $\lim_{x\to\pm\infty} S(x) = \pm\infty$. In the case of recurrence the diffusion admits a unique up to a constant multiple invariant measure $m(\mathrm{d}x)$, given by $m(\mathrm{d}x) = 1/(s(x)\sigma^2(x))\mathrm{d}x$. Denote $M = \int_{-\infty}^{+\infty} m(\mathrm{d}x)$. The diffusion is positively recurrent if and only if $M < \infty$. In this case put

$$\mu(\mathrm{d}x) = p(x)\mathrm{d}x, \text{ where } p(x) = \frac{1}{Ms(x)\sigma^2(x)}.$$

The probability μ is called invariant or stationary probability of X.

Using Assumptions 2.1.2 and 2.2.1, 2.2.2 we see that for any x such that $|x| \le M_0$,

$$s^{-1}(x) \le \mathrm{e}^{\frac{2M_0b_0}{\sigma_0^2}},$$

and for $|x| \ge M_0$,

$$s^{-1}(x) \le \mathrm{e}^{\frac{2M_0b_0}{\sigma_0^2}} \left(\frac{M_0}{|x|}\right)^{\frac{2\gamma}{\sigma_1^2}}.$$

This shows that $S(x) \to \pm\infty$, when $x \to \pm\infty$. Hence X is recurrent. The same estimation gives $M < \infty$ (and X is positively recurrent) as soon as $2\gamma > \sigma_1^2$.

Actually Assumption 2.2.3: $2\gamma > 31\sigma_1^2$ guarantees more than positive recurrence. It is well known that the positive recurrence of X is equivalent to $\mathbb{E}_x T_a < \infty$ for all $a \in \mathbb{R}$, $x \in \mathbb{R}$, where T_a is the hitting time of level a. Under Assumptions 2.1.2 and 2.2.1, 2.2.2 the moments of hitting times of X satisfy $\mathbb{E}_x T_a^n < \infty$ for $n < \gamma/\sigma_1^2 + 1/2$, for all $x \in \mathbb{R}$, $a \in \mathbb{R}$, see Loukianova *et al.* [16], Theorem 5.5. Thus under Assumption 2.2.3 we have $\mathbb{E}_x T_a^n < \infty$ for $n \le 16$. This means that the "speed of recurrence" of X is polynomial of order 16 and will be used to bound the speed of convergence of our estimator. Though we do not use the mixing coefficient, note that Assumption 2.2 guarantees that the diffusion is polynomially β-mixing (see Veretennikov [21]).

It follows from the above assumptions that the invariant density p is continuous and hence bounded from above and below on any compact interval. So we have

$$0 < p_0 \le p(x) \le p_1 < \infty \text{ for all } x \in [0,1].$$

In the sequel we need to fix p_0. We get immediately that

$$M = \int_{-\infty}^{+\infty} (s(x)\sigma^2(x))^{-1}\mathrm{d}x \le \frac{2M_0}{\sigma_0^2}\mathrm{e}^{\frac{2M_0b_0}{\sigma_0^2}} \left[\frac{2\gamma}{2\gamma - \sigma_1^2}\right] =: M_+.$$

This yields the following lower bound for all $x \in [0,1]$,

$$p(x) \geq \frac{1}{M_+}\frac{1}{\sigma_1^2}\mathrm{e}^{-2b_0/\sigma_0^2} := p_0. \tag{2.2}$$

In conclusion of this subsection, let us give an example of a diffusion process which fulfills Assumptions 2.2. Consider the solution of

$$\mathrm{d}X_t = -\frac{\gamma X_t}{1+X_t^2}\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = x, \quad \gamma > \frac{31}{2}.$$

It is positive recurrent with stationary distribution

$$\mu(\mathrm{d}x) \sim \frac{\mathrm{d}x}{(1+x^2)^\gamma}$$

and satisfies all the assumptions of 2.2. Remark that there is no evidence whether this diffusion is exponentially β-mixing.

2.2. **Construction of the estimator**

In this section we introduce a nonparametric estimator of the unknown drift function b on an interval K. We use the penalized least-squares based approach, where an estimator is constructed as a "projection" on some finite dimensional approximation space. We firstly address the non-adaptive case, where the statistician chooses himself the dimension of the approximation space. This choice can be done in an optimal way for example if the smoothness of the unknown function b is known. Secondly we address the adaptive estimation procedure. In this case the dimension of the approximation space is chosen automatically using some penalization procedure, based on the data.

Consider a collection $\{\mathcal{S}_m; m \in \mathcal{M}_t\}$ of approximation spaces. Each of these spaces is a linear finite dimensional subspace of $L^2(K, \mathrm{d}x)$. Here $\mathcal{M}_t$ is a set of indices. We suppose that there exists a space denoted by $\mathcal{S}_t$, belonging to the collection, such that $\mathcal{S}_m \subseteq \mathcal{S}_t$ for all $m \in \mathcal{M}_t$. Denote by D_m the dimension of $\mathcal{S}_m$ and by D_t the dimension of $\mathcal{S}_t$.

Put

$$\|h\|_t^2 = \frac{1}{t}\int_0^t h^2(X_s)\,\mathrm{d}s$$

and denote the corresponding quadratic form by

$$T_X(h,f) = \frac{1}{t}\int_0^t h(X_s)f(X_s)\,\mathrm{d}s \text{ for all } f, h \in \mathcal{S}_t.$$

We firstly construct the non-adaptive estimator. To this end fix a linear subspace $\mathcal{S}_m \subset \mathcal{S}_t$. We shall write shortly $b_K(x) := b(x)1_K(x)$ for the restriction of the function b to the interval K. The estimator $\hat{b}_m$ of b_K will be defined as trajectorial minimizer on $\mathcal{S}_m$ of the following contrast function:

$$\gamma_t(h) = \|h\|_t^2 - \frac{2}{t}\int_0^t h(X_s)\,\mathrm{d}X_s.$$

To insure the existence of $\hat{b}_m$ we impose some condition under which T_X is a.s. positive-definite on $\mathcal{S}_t$ and hence on each $\mathcal{S}_m$, $m \in \mathcal{M}_t$. Denote by $\|h\|$ the $L^2(K, \mathrm{d}x)$-norm, and let

$$\rho_t(X) = \inf_{h \in \mathcal{S}_t; \|h\|=1} T_X(h,h).$$

Put

$$A_t = \left\{\rho_t(X) \geq t^{-1/2}\right\}. \tag{2.3}$$

Note that, since $\mathcal{S}_t$ is finite-dimensional, γ_t is almost surely defined for all $h \in \mathcal{S}_t$ (see Rem. 2.3 below). We finally put

$$\hat{b}_m = \arg\min_{h\in\mathcal{S}_m} \gamma_t(h) \text{ on } A_t \text{ and } \hat{b}_m = 0 \text{ on } A_t^c.$$

Clearly, for all $\omega \in A_t$, T_X is a strictly positive-definite quadratic form on $\mathcal{S}_m$, $m \in \mathcal{M}_t$, and γ_t is a difference between this strictly positive quadratic form and a linear form. Hence the minimizer of γ_t exists and is unique on $\mathcal{S}_m$, $m \in \mathcal{M}_t$. As it was explained, in the non-adaptive case the statistician chooses himself the approximation space.

In the adaptive case the dimension is chosen automatically using a model selection procedure. In order to describe this procedure, we have to define properly $\gamma_t(\hat{b}_m)$. Fix some basis $\{\varphi_1, \ldots, \varphi_{D_m}\}$ of $\mathcal{S}_m$. From the definition of γ_t it follows that on A_t,

$$\hat{b}_m = \sum_{i=1}^{D_m} \hat{\alpha}_i \varphi_i,$$

with random $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{D_m})^*$ (we denote by $*$ the usual matrix transposition) satisfying

$$T^{\varphi}\hat{\alpha} = \frac{1}{t}\int_0^t \varphi(X_s)\mathrm{d}X_s, \tag{2.4}$$

where T^{φ} is the $D_m \times D_m$ random matrix with elements

$$T^{\varphi}_{ij} = \frac{1}{t}\int_0^t \varphi_i(X_s)\varphi_j(X_s)\,\mathrm{d}s$$

and where

$$\int_0^t \varphi(X_s)\,\mathrm{d}X_s = \begin{pmatrix} \int_0^t \varphi_1(X_s)\mathrm{d}X_s \\ \vdots \\ \int_0^t \varphi_{D_m}(X_s)\mathrm{d}X_s \end{pmatrix}.$$

Define on A_t

$$\gamma_t(\hat{b}_m) := \|\hat{b}_m\|_t^2 - \frac{2}{t}\sum_{i=1}^{D_m}\hat{\alpha}_i \int_0^t \varphi_i(X_s)\,\mathrm{d}X_s. \tag{2.5}$$

Now we are able to introduce the adaptive estimator. Define

$$\hat{m} := \arg\min_{m\in\mathcal{M}_t}\left[\gamma_t(\hat{b}_m) + \mathrm{pen}(m)\right],$$

where the penalization term $pen(m)$ will be given later, see (2.7). Then the estimator that we propose is the following adaptive estimator

$$\hat{b}_{\hat{m}} := \begin{cases} \sum_n 1_{\{\hat{m}=n\}}\hat{b}_n & \text{on } A_t. \\ 0 & \text{on } A_t^c \end{cases}$$

Remark 2.3. The above considerations and in particular the definition of $\gamma_t(\hat{b}_m)$ of (2.5) do not depend on the special choice of bases.

Indeed, let $\{\varphi_1, \ldots, \varphi_n\}$ and $\{\psi_1, \ldots, \psi_n\}$ be two bases of $\mathcal{S}_t$ (or $\mathcal{S}_m$), with $n = D_t$ (resp. D_m), and let $A = (a_{ij})$ be the $n \times n$ matrix such that $\varphi_i = \sum_j a_{ij}\psi_j$, for any $1 \le i \le n$. We then have for a function h

$$h = \sum_{i=1}^n \alpha_i\varphi_i = \sum_{i=1}^n \beta_i\psi_i,$$

where $\beta = A^*\alpha$.

1. Hence, given a version of the stochastic integrals $\int \varphi_i(X_s)\mathrm{d}X_s$, $1 \le i \le D_t$, the equalities

$$\begin{aligned}\int_0^t h(X_s)\mathrm{d}X_s &= \sum_{i=1}^{D_t} \alpha_i \int_0^t \varphi_i(X_s)\mathrm{d}X_s = \alpha^* \int_0^t \varphi(X_s)\mathrm{d}X_s \\ &= \alpha^* \int_0^t A\psi(X_s)\mathrm{d}X_s = \alpha^* A \int_0^t \psi(X_s)\mathrm{d}X_s = \sum_{i=1}^{D_t} \beta_i \int_0^t \psi_i(X_s)\mathrm{d}X_s\end{aligned}$$

determine automatically a version of any stochastic integral $\int h(X_s)\mathrm{d}X_s$ on $\mathcal{S}_t$, that does not depend on the choice of the basis.

2. From the definition (2.5) of $\gamma_t(\hat{b}_m)$, we have

$$\begin{aligned}\gamma_t(\hat{b}_m) &= \|\hat{b}_m\|_t^2 - \frac{2}{t}\left(\hat{\alpha}^* \int_0^t \varphi(X_s)\mathrm{d}X_s\right) \\ &= \|\hat{b}_m\|_t^2 - \frac{2}{t}\left(\hat{\alpha}^* \int_0^t A\psi(X_s)\mathrm{d}X_s\right) \\ &= \|\hat{b}_m\|_t^2 - \frac{2}{t}\left(\hat{\alpha}^* A \int_0^t \psi(X_s)\mathrm{d}X_s\right) \\ &= \|\hat{b}_m\|_t^2 - \frac{2}{t}\left(\hat{\beta}^* \int_0^t \psi(X_s)\mathrm{d}X_s\right)\end{aligned}$$

where $\hat{\beta} = A^*\hat{\alpha}$. The equality (2.4) yields

$$T^{\psi}\hat{\beta} = A^{-1}T^{\varphi}(A^{-1})^* A^*\hat{\alpha} = \frac{1}{t}\int_0^t \psi(X_s)\,\mathrm{d}X_s,$$

hence $\hat{\beta}$ satisfies (2.4), when replacing all φ_i by ψ_i. This implies that the definition of $\hat{b}_m$ and of $\gamma_t(\hat{b}_m)$ does not depend on the choice of a basis in $\mathcal{S}_m$.

2.3. **Assumptions on linear subspaces of $L^2(K, \mathrm{d}x)$**

We assume that the approximation spaces satisfy the following conditions:

Assumption 2.4.

1. *We suppose that there exists $\Phi_0 > 0$ such that for all $m \in \mathcal{M}_t$, for all $h \in \mathcal{S}_m$,*

$$||h||_\infty \le \Phi_0 D_m^{1/2}||h||.$$

Recall that $\|h\|^2 = \int_K h^2(x)\mathrm{d}x$ is the usual $L^2(K, \mathrm{d}x)-$norm.

2. *We suppose that*

$$\sum_{m\in\mathcal{M}_t} \mathrm{e}^{-D_m} \le C,$$

where the constant C does not depend on t.

3. *Dimension condition.*

$$D_t \le t.$$

4. *We suppose that there exists an orthonormal basis $\{\varphi_1, \ldots, \varphi_{D_t}\}$ of $\mathcal{S}_t \subset L^2(K, dx)$ and a positive constant Φ_1 such that for all i,*

$$\mathrm{card}\,\{j : ||\varphi_i\varphi_j||_\infty \ne 0\} \le \Phi_1.$$

5. We suppose that the cardinality of $\mathcal{M}_t$ satisfies

$$\operatorname{card} \mathcal{M}_t \leq D_t.$$

2.4. **Example for approximation spaces**

We present a collection of models that can be used for estimation. We consider the space of piecewise polynomials, as introduced for example in Baraud *et al.* [2,3] and Comte *et al.* [5].

Take $K = [0,1]$ and fix an integer $r \geq 0$. For $p \in \mathbb{N}$, consider the dyadic subintervals $I_{j,p} = [(j-1)2^{-p}, j2^{-p}]$, for any $1 \leq j \leq 2^p$. On each subinterval $I_{j,p}$, we consider polynomials of degree less or equal to r, so we have polynomials $\varphi_{j,l}, 0 \leq l \leq r$ of degree l, such that $\varphi_{j,l}$ is zero outside $I_{j,p}$. Then the space $\mathcal{S}_m$, for $m = (r,p)$, is defined as the space of all functions that can be written as

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

Hence, $D_m = (r+1)2^p$. Then the collection of spaces $\{\mathcal{S}_m, m \in \mathcal{M}_t\}$ is such that

$$\mathcal{M}_t = \{m = (r,p), p \geq 0, r \in \{0, \ldots, r_{\max}\}, 2^p(r_{\max}+1) \leq D_t\}.$$

One possible choice of $\mathcal{S}_t$ and D_t is as follows: take

$$p_{\max} := \max\{p : 2^p(r_{\max}+1) \leq t\}, \ D_t = 2^{p_{\max}}(r_{\max}+1)$$

and let $\mathcal{S}_t$ be the space of piecewise polynomials associated to $m_{\max} := (r_{\max}, p_{\max})$. Then it is evident that any of the spaces $\mathcal{S}_m, m \in \mathcal{M}_t$, is contained in $\mathcal{S}_t$. Furthermore, $\operatorname{card} \mathcal{M}_t = (p_{\max}+1)(r_{\max}+1) \leq D_t \leq t$.

It is well known, see for instance Comte *et al.* [5], that for this model the assumption of norm connection 2.4.2.4 is satisfied. Note moreover that for a fixed $\varphi_{j,l} \in \mathcal{S}_t$,

$$\operatorname{card}\{(j',l') : \varphi_{j',l'} \ \varphi_{j,l} \neq 0\} = \operatorname{card}\{(j,l') : \varphi_{j,l'} \ \varphi_{j,l} \neq 0\} \leq r_{\max}+1,$$

which does not depend on t. Hence assumption 2.4.2.4 is satisfied. Finally, it is easy to check that also Assumption 2.4.2.4 holds:

$$\begin{aligned}
\sum_{m \in \mathcal{M}_t} \mathrm{e}^{-D_m} &= \sum_{r=0}^{r_{\max}} \sum_{p: 2^p(r_{\max}+1) \leq D_t} \mathrm{e}^{-(r+1)2^p} \\
&\leq \sum_{r=0}^{r_{\max}} \sum_{p: 2^p(r_{\max}+1) \leq D_t} \mathrm{e}^{-2^p} \\
&\leq (r_{\max}+1) \sum_{k \geq 0} \mathrm{e}^{-k} < +\infty,
\end{aligned}$$

where the last quantity does not depend on t.

Spaces generated by compactly supported wavelets, similar to those considered by Hoffmann [10] and Baraud *et al.* [2] or [3] are also covered by Assumption 2.4. On the other hand, spaces spanned by the trigonometric basis do not fulfill Assumption 2.4.2.4 and therefore do not fit to our set-up.

2.5. **Main results**

We have the following first result concerning the non-adaptive estimator. Recall that $b_K(x) = b(x)1_K(x)$ is the restriction of the function b to the interval K. We define the risk of the estimator $\hat{b}_m$ as

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 = \mathbb{E}_x\left(\frac{1}{t}\int_0^t (\hat{b}_m - b_K)^2(X_s)\mathrm{d}s\right).$$

Let b_m be the $L^2(K,\mathrm{d}x)$-projection of b_K onto $\mathcal{S}_m$. Then the following holds.

Theorem 2.5. *Suppose that $t \geq t_0 := 4/p_0^2$. Suppose that X satisfies Assumptions 2.1 and 2.2. Suppose that the collection of the approximation spaces satisfies Assumptions 2.4.2.4, 2.4.3–5. Then*

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 \quad \leq \quad 3\kappa\|b_m - b_K\|^2 + 16\sigma_1^2\frac{\kappa}{p_0}\frac{D_m}{t} + Ct^{-1}. \tag{2.6}$$

Here, $\kappa = \kappa(t) = \frac{2}{\sigma_0^2}(\frac{2\mathrm{diam}(K)}{t} + \frac{2\sigma_1}{\sqrt{t}} + 2b_0 + \frac{\sigma_1^2}{2})$ (see Prop. 3.1), and C is a positive constant depending on b_0, σ_1 and Φ_0.

Let us give some comments on (2.6). It is natural to choose the dimension D_m that balances the bias term $||b_m - b_K||^2$ and the variance term which is of order D_m/t. Assume that b_K belongs to some Besov space $B^\alpha_{2,\infty}([0,1])$ and consider the space of piecewise polynomials $\mathcal{S}_m$ such that $r > \alpha - 1$. Then it can be shown that $||b_m - b_K||^2 \leq CD_m^{-2\alpha}$, see for example Barron *et al.* [4], Lemma 12. Thus the best choice of D_m is to take

$$D_m = t^{\frac{1}{2\alpha+1}}$$

and then we obtain

$$\mathbb{E}_x(||\hat{b}_m - b_K||_t^2) \leq Ct^{-\frac{2\alpha}{2\alpha+1}} + C_1t^{-1},$$

and this yields exactly the classical nonparametric rate $t^{-\frac{2\alpha}{2\alpha+1}}$ (compare for example to Hoffmann [10]). This choice however supposes the knowledge of the regularity α of the unknown drift function, and that is why an adaptive estimation scheme has to be used, in order to choose automatically the best dimension D_m in the case when the regularity α is not known.

Concerning the adaptive drift estimator, we have the following theorem.

Theorem 2.6. *Suppose that X satisfies Assumptions 2.1 and 2.2. Suppose that the collection of the approximation spaces satisfies Assumption 2.4. Suppose that $t \geq t_0$, where $t_0 := 4/p_0^2$. Let*

$$\mathrm{pen}(m) = \chi\sigma_1^2\frac{D_m}{t}, \tag{2.7}$$

where χ is a universal constant that will be given explicitly in (4.11). Then we have

$$\mathbb{E}_x||\hat{b}_{\hat{m}} - b_K||_t^2 \leq 3\kappa\inf_{m\in\mathcal{M}_t}\left(||b_m - b_K||^2 + \mathrm{pen}(m)\right) + \frac{C}{t},$$

where $\kappa = \kappa(t) = \frac{2}{\sigma_0^2}(\frac{2\mathrm{diam}(K)}{t} + \frac{2\sigma_1}{\sqrt{t}} + 2b_0 + \frac{\sigma_1^2}{2})$ (compare to Prop. 3.1) and where C is a positive constant not depending on t.

3. Probabilistic tools and auxiliary results

In this section, we collect some probabilistic results and auxiliary lemmas that are needed for the proofs of the main results.

3.1. **Probabilistic tools**

In what follows we often need to compare empirical and theoretical norms. One way of doing this is given by the next proposition.

Proposition 3.1. *For any positive function f having support on a compact interval K, we have*

$$\frac{1}{t}\mathbb{E}_x \int_0^t f(X_s)\mathrm{d}s \leq \kappa(t) \int_K f(x)\mathrm{d}x,$$

where $\kappa(t) = \frac{2}{\sigma_0^2}(\frac{2\mathrm{diam}(K)}{t} + \frac{2\sigma_1}{\sqrt{t}} + 2b_0 + \frac{\sigma_1^2}{2})$.

Proof. By the occupation time formula and since f has support in K,

$$\mathbb{E}_x \int_0^t f(X_s)\mathrm{d}s = \int_K f(y)\frac{2}{\sigma^2(y)}\mathbb{E}_x L_t^y \mathrm{d}y.$$

We will derive a bound on $\mathbb{E}_x L_t^y$ for $y \in K$. Let y_0 be the leftmost point of K. We have

$$\mathbb{E}_x L_t^{y_0} - \mathbb{E}_x|L_t^y - L_t^{y_0}| \leq \mathbb{E}_x L_t^y \leq \mathbb{E}_x L_t^{y_0} + \mathbb{E}_x|L_t^y - L_t^{y_0}|$$

and

$$|L_t^y - L_t^{y_0}| \leq |y - y_0| + \left|\int_0^t \mathbf{1}_{\{y_0<X_s<y\}}\sigma(X_s)\mathrm{d}W_s\right| + \int_0^t \mathbf{1}_{\{X_s\in K\}}|b(X_s)|\mathrm{d}s.$$

Taking expectation we obtain

$$\mathbb{E}_x \int_0^t \mathbf{1}_{\{X_s\in K\}}|b(X_s)|\mathrm{d}s \leq t\, b_0,$$

and by norm inclusion and isometry,

$$\begin{aligned}\mathbb{E}_x|\int_0^t 1_{\{y_0<X_s<y\}}\sigma(X_s)\mathrm{d}W_s| &\leq \left(\mathbb{E}_x(\int_0^t \mathbf{1}_{\{y_0<X_s<y\}}\sigma(X_s)\mathrm{d}W_s)^2\right)^{1/2} \\ &\leq \left(\mathbb{E}_x(\int_0^t \mathbf{1}_{\{X_s\in K\}}\sigma^2(X_s)\mathrm{d}s)\right)^{1/2} \leq \sigma_1\sqrt{t}.\end{aligned}$$

In conclusion,

$$\mathbb{E}_x L_t^y \leq \mathbb{E}_x L_t^{y_0} + \mathrm{diam}(K) + \sigma_1\sqrt{t} + tb_0 = C_0 + L,$$

where $L := \mathrm{diam}(K) + \sigma_1\sqrt{t} + tb_0$ and $C_0 = \mathbb{E}_x L_t^{y_0}$. We also have $C_0 - L \leq \mathbb{E}_x L_t^y$, so

$$t \geq \mathbb{E}_x \int_0^t 1_K(X_s)\mathrm{d}s = \int_K \frac{2\mathbb{E}_x L_t^y}{\sigma^2(y)}\mathrm{d}y \geq \frac{2(C_0 - L)}{\sigma_1^2},$$

whence

$$C_0 \leq L + \sigma_1^2 t/2,$$

and thus finally,

$$\mathbb{E}_x L_t^y \leq 2L + \sigma_1^2 t/2 = 2(\mathrm{diam}(K) + \sigma_1\sqrt{t} + tb_0) + \sigma_1^2 t/2$$

This concludes the proof. □

Now we give a useful deviation inequality for the one-dimensional ergodic diffusion process X, which is an immediate consequence of deviation inequality obtained by Loukianova *et al.* [16]. For $f : \mathbb{R} \to \mathbb{R}$ denote as usually $\mu(f) = \int_{\mathbb{R}} f\mathrm{d}\mu$.

Theorem 3.2 (deviation inequality)**.** *Let f be a measurable bounded function with compact support such that $\mu(f) \neq 0$. Suppose that X satisfies Assumptions 2.1 and 2.2.1, 2.2.2. Then for all $n \in \mathbb{N}$ such that*

$$n < \frac{\gamma}{\sigma_1^2} + \frac{1}{2}$$

and any $0 < \varepsilon \leq 1$, we have the following polynomial bound

$$\mathbb{P}_x\left(\left|\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s - \mu(f)\right| \geq \varepsilon\right) \leq K(n)t^{-n/2}\varepsilon^{-n}\mu(|f|)^n,$$

where $K(n)$ is positive and finite, depending on the coefficients of the diffusion and on n but not depending on f, t, ε.

This theorem follows directly from Theorems 4.3 and 5.5 of [16].

Corollary 3.3. *Under Assumption 2.2.3 the previous theorem is satisfied for all $n \leq 16$.*

3.2. **Auxiliary results**

In what follows we also need to compare empirical and theoretical norms through the set

$$\Omega_t = \left\{\forall h \in \mathcal{S}_t, \quad \frac{1}{2}\mu(h^2) \leq \|h\|_t^2 \leq \frac{3}{2}\mu(h^2)\right\}, \tag{3.1}$$

where any $h \in \mathcal{S}_t$ is defined as 0 outside of K. Recall that A_t is given by (2.3) and p_0 by (2.2).

Proposition 3.4. *For all $t \geq 4/p_0^2$ it holds that $\Omega_t \subseteq A_t$.*

Proof. Note that by the definition of A_t and Ω_t, under the assumption $t \geq 4/p_0^2$, the inequality $\mu(h^2)/2 \leq \|h\|_t^2$ implies $\|h\|_t^2 \geq p_0\|h\|^2/2 \geq t^{-1/2}$, so $\Omega_t \subseteq A_t$. □

Proposition 3.5. *Suppose that X satisfies Assumptions 2.1, 2.2.1 and 2.2.2. Suppose that the collection of approximation spaces $\{\mathcal{S}_m,\ m \in \mathcal{M}_t\}$ satisfies Assumptions 2.4.2.4, 2.4.2.4. Then for all*

$$n < \frac{\gamma}{\sigma_1^2} + \frac{1}{2}$$

and for all $x \in \mathbb{R}$ we have that

$$\mathbb{P}_x(\Omega_t^c) \leq Ct^{-\frac{1}{2}(n-2)},$$

where C depends on n, the constant Φ_1 given in Assumption 2.4.2.4 and on the coefficients of X, but does not depend on t.

Proof. Recall that $\|f\|$ denotes the usual $L^2(K, \mathrm{d}x)-$norm. For any function f, write

$$Z_t(f) := \frac{1}{t}\int_0^t f(X_s)\mathrm{d}s - \mu(f).$$

Since for f supported by K, $\|f\|_\mu^2 = 1$ implies that $||f||^2 \leq p_0^{-1}$, we have that

$$\mathbb{P}_x(\Omega_t^c) \leq \mathbb{P}_x\left(\sup_{f\in\mathcal{S}_t, \|f\|\leq 1} |Z_t(f^2)| > 0,5p_0\right).$$

Let $\{\varphi_1, \ldots, \varphi_{D_t}\}$ be an orthonormal basis of $\mathcal{S}_t \subset L^2(K, dx)$, satisfying Assumption 2.4.2.4, and note that any function f with $||f|| \leq 1$ can be written as

$$f = \sum_{i=1}^{D_t} a_i \varphi_i \text{ with } \sum a_i^2 \leq 1.$$

Therefore,

$$\begin{aligned} \mathbb{P}_x(\Omega_t^c) &\leq \mathbb{P}_x(\sup_{||f||\leq 1} |Z_t(f^2)| > 0,5p_0) \\ &\leq \mathbb{P}_x\left(\sup_{\sum a_i^2 \leq 1} \sum_{i,j} a_i a_j |Z_t(\varphi_i\varphi_j)| > 0,5p_0\right). \end{aligned}$$

Write

$$C_{ij} := \mu(|\varphi_i\varphi_j|)$$

and fix some positive number ε. On the set

$$\{|Z_t(\varphi_i\varphi_j)| \leq C_{ij}\varepsilon,\ \forall i,j\},$$

we have that

$$\sup_{\sum a_i^2 \leq 1} \sum a_i a_j |Z_t(\varphi_i\varphi_j)| \leq \varepsilon \varrho(C),$$

where $\varrho(C)$ is the biggest eigenvalue of the matrix C. Then choosing $\varepsilon := p_0/(4\varrho(C))$, we conclude that

$$\mathbb{P}_x(\Omega_t^c) \leq \mathbb{P}_x\left(\exists i,j :\ |Z_t(\varphi_i\varphi_j)| > C_{ij}\varepsilon\right).$$

By Theorem 3.2, we have the upper bound

$$\mathbb{P}_x\left(|Z_t(\varphi_i\varphi_j)| > C_{ij}\varepsilon\right) \leq K(n)\varrho(C)^n\, t^{-n/2}.$$

Note that due to Assumption 2.4.2.4 and since $\mu(|\varphi_i\varphi_j|) \leq p_1$, we have that

$$\varrho(C) \leq \Phi_1 p_1$$

where the upper bound does not depend on t. Indeed, using that $2u_iu_j \leq u_i^2 + u_j^2$, we have that

$$\begin{aligned} \varrho(C) &= \sup_{u\in\mathbb{R}^{D_t}, ||u||\leq 1} < Cu, u > = \sup_{u\in\mathbb{R}^{D_t}, ||u||\leq 1} \sum_{i,j} C_{ij}u_iu_j \\ &\leq \sup_{u\in\mathbb{R}^{D_t}, ||u||\leq 1} \sum_{i,j} C_{ij}u_i^2 \\ &= \sup_{u\in\mathbb{R}^{D_t}, ||u||\leq 1} \sum_i u_i^2 \sum_{j:\varphi_i\varphi_j\neq 0} \mu(|\varphi_i\varphi_j|) \\ &\leq \sup_{u\in\mathbb{R}^{D_t}, ||u||\leq 1} \sum_i u_i^2 \Phi_1 p_1 \leq \Phi_1 p_1. \end{aligned}$$

Using once more that

$$\sum_i \sum_j 1_{\{\varphi_i\varphi_j \neq 0\}} \leq D_t \cdot \Phi_1,$$

due to Assumption 2.4.2.4 we conclude that

$$\mathbb{P}_x(\Omega_t^c) \le C\, D_t t^{-n/2} \le C t^{-(n/2-1)},$$

where $C = K(n)\Phi_1^{n+1} p_1^n$ depends on n and coefficients of X, but does not depend on t. □

4. Proofs of the main results

4.1. Proof of Theorem 2.5

The proof follows the lines of Comte *et al.* [5]. Recall that from the definition of γ_t it follows that on A_t,

$$\hat{b}_m = \sum_{i=1}^{D_m} \hat{\alpha}_i \varphi_i,$$

with random $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_{D_m})^*$ satisfying

$$T\hat{\alpha} = \frac{1}{t}\int_0^t \varphi(X_s)\mathrm{d}X_s,$$

where T is the $D_m \times D_m$ random matrix and $\int_0^t \varphi(X_s)\mathrm{d}X_s$ is the D_m-dimensional random vector with elements

$$T_{ij} = \frac{1}{t}\int_0^t \varphi_i(X_s)\varphi_j(X_s)\,\mathrm{d}s, \qquad \int_0^t \varphi(X_s)\,\mathrm{d}X_s = \begin{pmatrix} \int_0^t \varphi_1(X_s)\mathrm{d}X_s \\ \vdots \\ \int_0^t \varphi_{D_m}(X_s)\mathrm{d}X_s \end{pmatrix}.$$

Observe that $\hat{b}_m$ is a $\mathcal{F}_t$-measurable random variable with values in $\mathcal{S}_m$. If for such a random variable

$$h(\omega, x) = \sum_{i=1}^{D_m} \alpha_i(\omega)\varphi_i(x)$$

we put

$$\gamma_t(h) = \|h\|_t^2 - \frac{2}{t}\sum_{i=1}^{D_m} \alpha_i \int_0^t \varphi_i(X_s)\,\mathrm{d}X_s.$$

Then $\gamma_t(h) - \gamma_t(\hat{b}_m) \ge 0$ on A_t. This inequality is evidently valid for any basis of $\mathcal{S}_m$.

Finally, we define the risk of the estimator $\hat{b}_m$ as

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 = \mathbb{E}_x\left(\frac{1}{t}\int_0^t (\hat{b}_m - b_K)^2(X_s)\mathrm{d}s\right).$$

Let Ω_t be given by (3.1) and A_t given by (2.3). Recall that $\Omega_t \subseteq A_t$ (Prop. 3.4.)

Now write

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 = \mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 \mathbf{1}_{\Omega_t} + \mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 \mathbf{1}_{\Omega_t^c}.$$

We will treat separately the two terms on the right-hand side.

We start with the first one, recalling that $\Omega_t = \Omega_t \cap A_t$. In what follows it will be useful to use an orthonormal basis $\{\psi_1, \dots, \psi_{D_m}\}$ of $\mathcal{S}_m$ viewed as a subspace of $L^2(K, \mathrm{d}\mu)$. Hence, our estimator can be rewritten as

$$\hat{b}_m = \sum_{i=1}^{D_m} \hat{\beta}_i \psi_i, \text{ and } b_m = \sum_{i=1}^{D_m} \beta_i \psi_i.$$

Observe that a.s. on A_t

$$\begin{aligned}0 \le \gamma_t(\hat{b}_m) - \gamma_t(b_m) &= \|\hat{b}_m\|_t^2 - \|b_m\|_t^2 - \frac{2}{t}\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)\int_0^t \psi_i(X_s)\left(b(X_s)\,\mathrm{d}s + \sigma(X_s)\,\mathrm{d}W_s\right)\\ &= T_X(\hat{b}_m - b_m, \hat{b}_m + b_m) - 2T_X(\hat{b}_m - b_m, b_K) - \frac{2}{t}\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)\int_0^t \psi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s\\ &= \|\hat{b}_m - b_K\|_t^2 - \|b_m - b_K\|_t^2 - \frac{2}{t}\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)\int_0^t \psi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s,\end{aligned}$$

whence a.s. on A_t

$$\|\hat{b}_m - b_K\|_t^2 \le \|b_m - b_K\|_t^2 + 2\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)\left(\frac{1}{t}\int_0^t \psi_i(X_s)\sigma(X_s)\mathrm{d}W_s\right). \tag{4.1}$$

Remark that $\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)^2 = \|\hat{b}_m - b_m\|_\mu^2$. Using Cauchy-Schwartz inequality we have

$$\begin{aligned}\|\hat{b}_m - b_K\|_t^2\mathbf{1}_{\Omega_t} &\le \|b_m - b_K\|_t^2\mathbf{1}_{\Omega_t} + 2\sum_{i=1}^{D_m}(\hat{\beta}_i - \beta_i)\left(\frac{1}{t}\int_0^t \psi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s\right)\mathbf{1}_{\Omega_t}\\ &\le \|b_m - b_K\|_t^2 + \frac{1}{8}\|\hat{b}_m - b_m\|_\mu^2\mathbf{1}_{\Omega_t}\\ &\quad + 8\sum_{i=1}^{D_m}\left(\frac{1}{t}\int_0^t \psi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s\right)^2.\end{aligned} \tag{4.2}$$

Then on Ω_t

$$\frac{1}{8}\|\hat{b}_m - b_m\|_\mu^2\mathbf{1}_{\Omega_t} \le \frac{1}{2}(\|\hat{b}_m - b_K\|_t^2 + \|b_m - b_K\|_t^2)\mathbf{1}_{\Omega_t}.$$

Plugging this into (4.2) gives

$$\|\hat{b}_m - b_K\|_t^2\mathbf{1}_{\Omega_t} \le 3\|b_m - b_K\|_t^2 + 16\sum_{i=1}^{D_m}\left(\frac{1}{t}\int_0^t \psi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s\right)^2.$$

We have

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2\mathbf{1}_{\Omega_t} \le \frac{3}{t}\mathbb{E}_x\int_0^t (b_m - b_K)^2(X_s)\mathrm{d}s + \frac{16\sigma_1^2}{t^2}\sum_{i=1}^{D_m}\mathbb{E}_x\int_0^t \psi_i^2(X_s)\,\mathrm{d}s.$$

Using Proposition 3.1, we can write for any positive function f having support on K,

$$\mathbb{E}_x\int_0^t f(X_s)\mathrm{d}s \le \kappa t\int_K f\mathrm{d}x,$$

where the constant κ is explicitly given in Proposition 3.1 and does only depend on the model constants b_0, σ_0, σ_1. Using this estimation, we obtain the following bound for the integrated risk restricted on Ω_t:

$$\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2\mathbf{1}_{\Omega_t} \le 3\kappa\|b_m - b_K\|^2 + 16\sigma_1^2\frac{\kappa}{p_0}\frac{D_m}{t}.$$

We now consider the risk restricted on Ω_t^c. Recall that $A_t^c \subseteq \Omega_t^c$ and that $\hat{b}_m = 0$ on A_t^c, and write

$$\|b_K - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c} = \|b_K - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} + \|b_K\|_t^2 \mathbf{1}_{A_t^c} \tag{4.3}$$

Let $\tilde{b}_m$ be the almost surely defined on $\Omega_t^c \cap A_t$ orthogonal projection of b_K onto $\mathcal{S}_m$ w.r.t. $\|\cdot\|_t$. We have

$$\begin{aligned} \|b_K - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} &= \|b_K - \tilde{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} + \|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} \\ &\leq \|b_K\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} + \|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t}, \end{aligned}$$

which, combined with (4.3), implies

$$\|b_K - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c} \leq \|b_K\|_t^2 \mathbf{1}_{\Omega_t^c} + \|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t}. \tag{4.4}$$

Our Assumption 2.1.1 on $b(x)$ yields

$$\mathbb{E}_x \|b_K\|_t^2 \mathbf{1}_{\Omega_t^c} \leq b_0^2 \mathbb{P}_x(\Omega_t^c). \tag{4.5}$$

From the definition of $\tilde{b}_m$ it follows that $\tilde{b}_m = \sum_{i=1}^{D_m} \tilde{\alpha}_i \varphi_i$, with $\tilde{\alpha}$ satisfying

$$T\tilde{\alpha} = \frac{1}{t} \int_0^t \varphi(X_s) b(X_s) \mathrm{d}s.$$

Recall that on A_t, $\hat{b}_m = \sum_{i=1}^{D_m} \hat{\alpha}_i \varphi_i$, with $\hat{\alpha}$ given by (2.4). Hence on A_t, we can write $\hat{\alpha} - \tilde{\alpha} = T^{-1} M_t$, where

$$M_t = \frac{1}{t} \int_0^t \varphi(X_s) \sigma(X_s) \, \mathrm{d}W_s = \begin{pmatrix} \frac{1}{t} \int_0^t \varphi_1(X_s) \sigma(X_s) \mathrm{d}W_s \\ \vdots \\ \frac{1}{t} \int_0^t \varphi_{D_m}(X_s) \sigma(X_s) \mathrm{d}W_s \end{pmatrix}.$$

So on A_t we have $\hat{b}_m - \tilde{b}_m = \varphi^*(\hat{\alpha} - \tilde{\alpha}) = \varphi^* T^{-1} M_t$, where $\varphi^* = (\varphi_1, \dots \varphi_{D_m})$, and (we denote by $*$ the matrix-transposition operation),

$$(\hat{b}_m - \tilde{b}_m)^2(X_s) = M_t^* (T^*)^{-1} \varphi \varphi^*(X_s) T^{-1} M_t.$$

So,

$$\begin{aligned} \|\tilde{b}_m - \hat{b}_m\|_t^2 &= \frac{1}{t} \int_0^t (\tilde{b}_m - \hat{b}_m)^2(X_s) \mathrm{d}s \\ &= M_t^* (T^*)^{-1} T T^{-1} M_t = M_t^* (T^*)^{-1} M_t = \langle T^{-1} M_t, M_t \rangle, \end{aligned}$$

which gives, by the definition of A_t,

$$\|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} \leq \frac{1}{t^{-1/2}} \|M_t\|^2 \mathbf{1}_{\Omega_t^c} = t^{1/2} \sum_{i=1}^{D_m} \left(\frac{1}{t} \int_0^t \varphi_i(X_s) \sigma(X_s) \, \mathrm{d}W_s \right)^2 \mathbf{1}_{\Omega_t^c}. \tag{4.6}$$

Using Burkholder-Davis-Gundy inequalities and the hypothesis $\|\varphi_i^2\|_\infty \le \Phi_0^2 D_m$, it follows from (4.6),

$$\begin{aligned}\mathbb{E}_x\|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} &\le \frac{t^{1/2}}{t^2}\sum_{i=1}^{D_m}\mathbb{E}_x\left((\int_0^t \varphi_i(X_s)\sigma(X_s)\,\mathrm{d}W_s)^2 1_{\Omega_t^c}\right)\\ &\le t^{-3/2}\sum_{i=1}^{D_m}\sqrt{\mathbb{E}_x\left(\int_0^t \varphi_i(X_s)\sigma(X_s)\,\mathrm{d}Ws\right)^4 \mathbb{P}_x(\Omega_t^c)}\\ &\le t^{-3/2}\sum_{i=1}^{D_m}\sqrt{C(4)\mathbb{E}_x\left(\int_0^t \varphi_i^2(X_s)\sigma^2(X_s)\,\mathrm{d}s\right)^2 \mathbb{P}_x(\Omega_t^c)}.\end{aligned}$$

Here, $C(4)$ is a Burkholder-Davis-Gundy constant. But

$$\int_0^t \varphi_i^2(X_s)\sigma^2(X_s)\,\mathrm{d}s \le \Phi_0^2 D_m \sigma_1^2 t,$$

hence

$$\begin{aligned}\mathbb{E}_x\|\tilde{b}_m - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c \cap A_t} &\le \sqrt{C(4)}\, t^{-3/2}\sum_{i=1}^{D_m}\sqrt{\Phi_0^4 D_m^2 \sigma_1^4 t^2\, \mathbb{P}_x(\Omega_t^c)}\\ &\le \sqrt{C(4)}\sigma_1^2\Phi_0^2\, t^{-1/2} D_m^2 \sqrt{\mathbb{P}_x(\Omega_t^c)}.\end{aligned}$$

From (4.4) and (4.5) the integrated risk on Ω_t^c satisfies

$$\begin{aligned}\mathbb{E}_x\|b_K - \hat{b}_m\|_t^2 \mathbf{1}_{\Omega_t^c} &\le \left(b_0^2 + C\sigma_1^2\Phi_0^2\, t^{-1/2}\, D_m^2\right)\sqrt{\mathbb{P}_x(\Omega_t^c)}\\ &\le \left(b_0^2 + C\sigma_1^2\Phi_0^2\right)\, t^{-1/2}\, D_m^2\sqrt{\mathbb{P}_x(\Omega_t^c)}.\end{aligned} \tag{4.7}$$

As a consequence, since $D_m^2 \le t^2$, the full integrated risk satisfies

$$\begin{aligned}\mathbb{E}_x\|\hat{b}_m - b_K\|_t^2 &\le 3\kappa\|b_m - b_K\|^2 + 16\sigma_1^2\frac{\kappa}{p_0}\frac{D_m}{t}\\ &\quad + \left(b_0^2 + C\sigma_1^2\Phi_0^2\right)\, t^{3/2}\sqrt{\mathbb{P}_x(\Omega_t^c)}.\end{aligned}$$

Finally, Proposition 3.5, applied with $n = 12$, yields

$$\mathbb{P}_x(\Omega_t^c) \le \frac{C}{t^5}$$

for $t \ge t_0$. This finishes the proof. □

Remark 4.1. In the case when X is in the stationary regime, *i.e.* starting from the invariant measure μ, (2.6) can be improved to

$$\mathbb{E}_\mu\|\hat{b}_m - b_K\|_t^2 \le 3p_1\|b_m - b_K\|^2 + 16\sigma_1^2\frac{D_m}{t} + Ct^{-1}.$$

4.2. **Proof of Theorem 2.6**

Put

$$\nu_t(f) := \frac{1}{t}\int_0^t f(X_s)\sigma(X_s)\mathrm{d}W_s.$$

The same argument that yields (4.1) in the non-adaptive case gives for any $m \in \mathcal{M}_t$,

$$||\hat{b}_{\hat{m}} - b_K||_t^2 1_{A_t} \leq ||b_m - b_K||_t^2 1_{A_t} + 2\nu_t(\hat{b}_{\hat{m}} - b_m)1_{A_t} + (\mathrm{pen}(m) - \mathrm{pen}(\hat{m}))\, 1_{A_t}. \tag{4.8}$$

Here, a special attention has to be paid to the term $\nu_t(\hat{b}_{\hat{m}} - b_m)$, since it is not a priori clear that this stochastic integral is well-defined. On $\hat{m} = n$, $\hat{b}_{\hat{m}} - b_m$ is an element of $\mathcal{S}_n + \mathcal{S}_m$ viewed as linear subspace of $L^2(K, \mu)$. Put $k = \dim(\mathcal{S}_n + \mathcal{S}_m)$ and let $\{\psi_1, \ldots, \psi_k\}$ be an orthonormal basis of this subspace. Then $1_{\{\hat{m}=n\}}(\hat{b}_{\hat{m}} - b_m) = 1_{\{\hat{m}=n\}} \sum_{i=1}^k \hat{\beta}_i \psi_i$, and we define on $\hat{m} = n$,

$$\nu_t(\hat{b}_{\hat{m}} - b_m) := \sum_{i=1}^k \hat{\beta}_i \nu_t(\psi_i).$$

Hence, $\nu_t(\hat{b}_{\hat{m}} - b_m)$ is well-defined and linear. Thus we may write

$$\nu_t(\hat{b}_{\hat{m}} - b_m) \leq ||\hat{b}_{\hat{m}} - b_m||_\mu \cdot \nu_t\left(\frac{\hat{b}_{\hat{m}} - b_m}{||\hat{b}_{\hat{m}} - b_m||_\mu}\right) \leq ||\hat{b}_{\hat{m}} - b_m||_\mu \cdot \sup_{h \in \mathcal{S}_m + \mathcal{S}_{\hat{m}}, ||h||_\mu = 1} |\nu_t(h)|.$$

Write for short

$$G_m(m') := \sup_{h \in \mathcal{S}_m + \mathcal{S}_{m'}, ||h||_\mu = 1} |\nu_t(h)|.$$

We now investigate (4.8). First, on $A_t \cap \Omega_t$, using that $2ab \leq \frac{1}{8}a^2 + 8b^2$,

$$\begin{aligned}
||\hat{b}_{\hat{m}} - b_K||_t^2 &\leq ||b_m - b_K||_t^2 + 2||\hat{b}_{\hat{m}} - b_m||_\mu \quad G_m(\hat{m}) + [\mathrm{pen}(m) - \mathrm{pen}(\hat{m})] \\
&\leq ||b_m - b_K||_t^2 + \frac{1}{8}||\hat{b}_{\hat{m}} - b_m||_\mu^2 + 8\, G_m^2(\hat{m}) + [\mathrm{pen}(m) - \mathrm{pen}(\hat{m})] \\
&\leq ||b_m - b_K||_t^2 + \frac{1}{2}\left(||\hat{b}_{\hat{m}} - b_K||_t^2 + ||b_K - b_m||_t^2\right) \\
&\quad +8\, G_m^2(\hat{m}) + [\mathrm{pen}(m) - \mathrm{pen}(\hat{m})] \\
&\leq \frac{3}{2}||b_m - b_K||_t^2 + \frac{1}{2}||\hat{b}_{\hat{m}} - b_K||_t^2 + 8\, G_m^2(\hat{m}) + [\mathrm{pen}(m) - \mathrm{pen}(\hat{m})]\,.
\end{aligned}$$

This yields finally, on $A_t \cap \Omega_t = \Omega_t$,

$$||\hat{b}_{\hat{m}} - b_K||_t^2 \leq 3||b_m - b_K||_t^2 + 16\, G_m^2(\hat{m}) + 2\,[\mathrm{pen}(m) - \mathrm{pen}(\hat{m})]\,. \tag{4.9}$$

Now, as in Comte *et al.* [5], put $p(m, m') := p(m) + p(m')$, where

$$p(m) := \chi \sigma_1^2 \frac{D_m}{t}$$

and where χ is a universal constant. Then

$$\begin{aligned}
G_m^2(\hat{m}) 1_{\Omega_t} &\leq \left[(G_m^2(\hat{m}) - p(m, \hat{m}))1_{\Omega_t}\right]_+ + p(m, \hat{m}) \\
&\leq \sum_{n \in \mathcal{M}_t} \left[(G_m^2(n) - p(m, n))1_{\Omega_t}\right]_+ + p(m, \hat{m}).
\end{aligned}$$

Now we choose pen(m) such that $8p(m,m') \leq \text{pen}(m) + \text{pen}(m')$, *i.e.*

$$\text{pen}(m) := 8\chi\sigma_1^2 \frac{D_m}{t}\cdot$$

We have (see also Baraud *et al.* [3]):

Lemma 4.2. *Under the assumptions of Theorem 2.6,*

$$\mathbb{E}_x\left((G_m^2(m') - p(m,m'))1_{\Omega_t}\right)_+ \leq 1,6\chi\sigma_1^2\frac{1}{t}\mathrm{e}^{-D_{m'}}, \tag{4.10}$$

where χ is given by

$$\chi = 3c(\delta_0)(1+2\delta_0)(1+15\delta_0),\ c(\delta_0) = \max\left(2\ln 2 + 1, \ln\left(\frac{9}{2\delta_0^2}\right)\right)\cdot \tag{4.11}$$

Here, $0 < \delta_0 < 1$ is a free parameter. For the choice $\delta_0 = 0,0138$ this yields a constant $\chi = 38$.

The proof of Lemma 4.2 will be given in Section 4.2 below.

For any n, let $\{\varphi_1^n, \ldots, \varphi_{D_n}^n\}$ be an orthonormal basis of $\mathcal{S}_n$. On $A_t \cap \Omega_t^c$, using (4.4) and (4.6), we have

$$\begin{aligned}
||\hat{b}_{\hat{m}} - b_K||_t^2 1_{\{A_t\cap\Omega_t^c\}} &= \sum_{n\in\mathcal{M}_t} 1_{\{\hat{m}=n;A_t\cap\Omega_t^c\}}||\hat{b}_n - b_K||_t^2 \\
&\leq \|b_K\|_t^2\mathbf{1}_{\Omega_t^c} + \sum_{n\in\mathcal{M}_t} 1_{\{\hat{m}=n\}}\|\tilde{b}_n - \hat{b}_n\|_t^2\mathbf{1}_{\Omega_t^c\cap A_t} \\
&\leq \|b_K\|_t^2\mathbf{1}_{\Omega_t^c} + \sum_{n\in\mathcal{M}_t} 1_{\{\hat{m}=n\}}t^{1/2}\sum_{i=1}^{D_n}\left(\frac{1}{t}\int_0^t \varphi_i^n(X_s)\sigma(X_s)\,\mathrm{d}W_s\right)^2\mathbf{1}_{\Omega_t^c} \\
&\leq \|b_K\|_t^2\mathbf{1}_{\Omega_t^c} + \sum_{n\in\mathcal{M}_t} t^{1/2}\sum_{i=1}^{D_n}\left(\frac{1}{t}\int_0^t \varphi_i^n(X_s)\sigma(X_s)\,\mathrm{d}W_s\right)^2\mathbf{1}_{\Omega_t^c},
\end{aligned}$$

The same calculus that yields (4.7) now gives

$$\mathbb{E}_x||\hat{b}_{\hat{m}} - b_K||_t^2 1_{\{A_t\cap\Omega_t^c\}} \leq C\left(b_0^2 + \sigma_1^2\Phi_0^2\right)t^{-1/2}D_t^2|\mathcal{M}_t|\sqrt{\mathbb{P}_x(\Omega_t^c)}. \tag{4.12}$$

(4.9), (4.10) and (4.12) yield finally, for any m, using Assumption 2.4,

$$\begin{aligned}
\mathbb{E}_x||\hat{b}_{\hat{m}} - b_K||_t^2 1_{A_t} &\leq 3\mathbb{E}_x||b_m - b_K||_t^2 + 4\text{pen}(m) + 16\sum_{n\in\mathcal{M}_t} 1,6\chi\sigma_1^2\frac{1}{t}\mathrm{e}^{-D_n} \\
&\quad + C\left(b_0^2 + \sigma_1^2\Phi_0^2\right)t^{-1/2}D_t^2|\mathcal{M}_t|\sqrt{\mathbb{P}_x(\Omega_t^c)} \\
&\leq 3\kappa||b_m - b||^2 + 4\text{pen}(m) + C\chi\sigma_1^2\frac{1}{t} \\
&\quad + C(b_0^2,\sigma_1^2)\ t^{-1/2}\ D_t^3\mathbb{P}_x(\Omega_t^c)^{1/2}.
\end{aligned}$$

Now, recall that by Proposition 3.5, since $D_t^3 \leq t^3$, taking $n = 16$,

$$\mathbb{P}_x(\Omega_t^c)^{1/2} \leq Ct^{-7/2}$$

and by Proposition 3.4,

$$\mathbb{P}_x(A_t^c) \leq Ct^{-1}.$$

As a consequence,

$$\mathbb{E}_x||\hat{b}_{\hat{m}} - b_K||_t^2 \quad \leq \quad 3\kappa \inf_{m\in\mathcal{M}_t} \left(||b_m - b||^2 + \mathrm{pen}(m)\right) + C\chi\sigma_1^2\frac{1}{t} + C(b_0^2, \sigma_1^2)t^{-1}.$$

This finishes the proof. □

A. Appendix

In this appendix we give the proof of Lemma 4.2. Write

$$\nu_t(f) := \frac{1}{t}\int_0^t f(X_s)\sigma(X_s)\mathrm{d}W_s.$$

Using the classical Bernstein inequality for continuous martingales , we recall that for all $a > 0$, $v \neq 0$

$$\mathbb{P}_x(\nu_t(f) \geq a, ||f||_t^2 \leq v^2) \leq \exp\left(-\frac{ta^2}{2\sigma_1^2 v^2}\right). \tag{4.13}$$

Recall that $||f||_t^2 = \frac{1}{t}\int_0^t f^2(X_s)\mathrm{d}s$ and $||f||_\mu^2 = \mu(f^2)$.

The proof of Proposition 4.2 follows Baraud *et al.* [3], pages 45–47. The chaining arguments which permits to state exponential bounds on supremum of empirical processes can also be found in Barron *et al.* [4]. By Lorentz *et al.* [13], for any linear subspace S of $L^2([0,1],\mu)$ of dimension d, one can find a set $T_\delta \subset B$, where B is the unit ball of $S \subset L^2([0,1],\mu)$, such that

$$\mathrm{card}\,(T_\delta) \leq \left(\frac{3}{\delta}\right)^d, \text{ and } \forall f \in B\ \exists f_\delta \in T_\delta : ||f - f_\delta||_\mu \leq \delta.$$

Apply this to the linear space $\mathcal{S}_m + \mathcal{S}_{m'}$ of dimension $d(m') \leq D_m + D_{m'}$. Consider δ_k-sets $T_k = T_{\delta_k}$ where $\delta_k = \delta_0 2^{-k}$, where $\delta_0 < 1$ is to be chosen later. Set $H_k := \log \mathrm{card}\,(T_k)$. Write $B_{m'} := \{f \in \mathcal{S}_m + \mathcal{S}_{m'} : ||f||_\mu \leq 1\}$. Then for any $f \in B_{m'}$, one can find a sequence $(f_k)_k$ with $f_k \in T_k$ such that $||f - f_k||_\mu \leq \delta_k$. Hence we get

$$f = f_0 + \sum_{k\geq 1}(f_k - f_{k-1}).$$

Then as in Baraud *et al.* [3],

$$||f_0||_\mu \leq 1,\ ||f_k - f_{k-1}||_\mu^2 \leq 5\delta_{k-1}^2/2.$$

In the following, we shall work in restriction to the set Ω_t. Write $\mathbb{P}_t$ for the measure $\mathbb{P}_x(\cdot \cap \Omega_t)$. Put as in Baraud *et al.* [3],

$$\Delta := \sqrt{3}\sigma_1\left(\sqrt{x_0} + \sum_{k\geq 1}\delta_{k-1}\sqrt{5x_k/2}\right),$$

then we have that

$$\begin{aligned}\mathbb{P}_t\left(\sup_{f\in B_{m'}}\nu_t(f) \geq \Delta\right) &= \mathbb{P}_t\left(\exists (f_k)_k, f_k \in T_k : \nu_t(f_0) + \sum_{k\geq 1}\nu_t(f_k - f_{k-1}) \geq \Delta\right)\\ &\leq P_1 + P_2,\end{aligned}$$

where

$$P_1 = \sum_{f_0 \in T_0} \mathbb{P}_t \left(\nu_t(f_0) \geq \sqrt{3x_0}\sigma_1 \right),$$

and

$$P_2 = \sum_{k \geq 1} \sum_{f_k \in T_k, f_{k-1} \in T_{k-1}} \mathbb{P}_t \left(\nu_t(f_k - f_{k-1}) \geq \sigma_1 \delta_{k-1} \sqrt{15x_k/2} \right).$$

Recall (4.13): Since on Ω_t, $||f||_t^2 \leq \frac{3}{2}||f||_\mu^2$, we have for all $x > 0$,

$$\mathbb{P}_t \left(\nu_t(f) \geq \sqrt{3}\sigma_1 \sqrt{x} ||f||_\mu \right) \leq \exp(-tx).$$

We apply this inequality, remarking that $||f_0||_\mu \leq 1$, hence

$$P_1 \leq \operatorname{card}(T_0) \exp(-tx_0) = \exp(H_0 - tx_0)$$

and, since $||f_k - f_{k-1}||_\mu^2 \leq 5\delta_{k-1}^2/2$,

$$P_2 \leq \sum_{k \geq 1} \exp(H_{k-1} + H_k - tx_k).$$

Now, choose x_0 such that

$$tx_0 = H_0 + D_{m'} + \tau$$

and x_k such that

$$tx_k = H_{k-1} + H_k + D_{m'} + kd(m') + \tau.$$

Then, if $d(m') \geq 1$, we obtain as in Baraud *et al.* [3],

$$\mathbb{P}_t \left(\sup_{f \in B_{m'}} \nu_t(f) \geq \Delta \right) \leq 1,6 \mathrm{e}^{-\tau} \mathrm{e}^{-D_{m'}}. \tag{4.14}$$

Else, $d(m') = 0$, hence $\mathcal{S}_m + \mathcal{S}_{m'} = \{0\}$, and (4.14) holds trivially. Exactly as in Baraud *et al.* [3], it can be shown that

$$t\Delta^2 \leq \chi\sigma_1^2(D_{m'} + D_m + \tau),$$

where χ is the constant given in (4.11), and then we conclude as there

$$\mathbb{E}_x \left[\left(G_m^2(m') - \chi\sigma_1^2 \frac{D_{m'} + D_m}{t} \right)_+ 1_{\Omega_t} \right] \leq 1,6\chi\sigma_1^2 \frac{1}{t} \mathrm{e}^{-D_{m'}}.$$

Acknowledgements. The subject of this paper has been proposed to the authors by Fabienne Comte and Valentine Genon-Catalot during their research period at the University Paris-Descartes in spring 2008. The authors thank both of them for their kindness, their patience, and last, but not least for all the time spent on discussions and explanations on the paper. The authors are also grateful to the referees for comments that helped to significantly improve the paper. Eva Löcherbach has been partially supported by an ANR project: Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-BLAN-0220-01.

References

[1] G. Banon, Nonparametric identification for diffusion processes. *SIAM J. Control. Optim.* **16** (1978) 380–395.
[2] Y. Baraud, F. Comte and G. Viennet, Adaptive estimation in autoregression or β-mixing regression *via* model selection. *Ann. Stat.* **29** (2001) 839–875.
[3] Y. Baraud, F. Comte and G. Viennet, Model selection for (auto-)regression with dependent data. *ESAIM*: *P&S* **5** (2001) 33–49.
[4] A. Barron, L. Birgé and P. Massart, Risks bounds for model selection *via* penalization. *Prob. Th. Rel. Fields* **113** (1999) 301–413.
[5] F. Comte, V. Genon-Catalot and Y. Rozenholc, Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13** (2007) 514–543.
[6] A. Dalalyan, Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Stat.* **33** (2005) 507–2528.
[7] A.S. Dalalyan and Yu.A. Kutoyants, Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Math. Meth. Stat.* **11** (2002) 402–427.
[8] S. Delattre, M. Hoffmann and M. Kessler, *Dynamics adaptive estimation of a scalar diffusion*. Prpublication PMA-762, Univ. Paris 6 (2002). Available at www.proba.jussieu.fr/mathdoc/preprints/. Mathematical Reviews (MathSciNet): MR1895888 Project Euclid: euclid.bj/1078866865.
[9] L. Galtchouk and S. Pergamenschikov, Sequential nonparametric adaptive estimation of the drift coefficient in the diffusion processes. *Math. Meth. Stat.* **10** (2001) 316–330.
[10] M. Hoffmann, Adaptive estimation in diffusion processes. *Stochastic Processes Appl.* **79** (1999) 135–163.
[11] Yury A. Kutoyants, *Statistical inference for ergodic diffusion processes. Springer Series in Statistics*. London: Springer (2004).
[12] O. Lepskii, One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** (1999) 459–470.
[13] G.G. Lorentz, M. von Golitschek and Y. Makovoz, *Constructive approximation: advanced problems*. Grundlehren der Mathematischen Wissenschaften **304**. Berlin: Springer (1996).
[14] D. Loukianova and O. Loukianov, Uniform deterministic equivalent of additive functionals and non-parametric drift estimation for one-dimensional recurrent diffusions. *Ann. Inst. Henri Poincaré* **44** (2008) 771–786.
[15] E. Löcherbach and D. Loukianova, On Nummelin splitting for continuous time Harris recurrent Markov processes and application to kernel estimation for multi-dimensional diffusions. *Stoch. Proc. Appl.* **118** (2008) 301–1321.
[16] E. Löcherbach, D. Loukianova and O. Loukianov, Polynomial bounds in the Ergodic theorem for one-dimensional diffusions and integrability of hitting times. *Ann. Inst. H. Poincaré Probab. Statist.* **47** (2011) 425–449.
[17] T.D. Pham, Nonparametric estimation of the drift coefficient in the diffusion equation. *Math. Operationsforsch. Statist., Ser. Statistics* **1** (1981) 61–73.
[18] B.L.S. Prakasa Rao, *Statistical Inference for Diffusion Type Processes*. London: Edward Arnold. MR1717690 (1999)
[19] D. Revuz and M. Yor, Continuous martingales and Brownian motion. 3rd ed. Grundlehren der Mathematischen Wissenschaften **293**. Berlin: Springer (2005).
[20] V.G. Spokoiny, Adaptive drift estimation for nonparametric diffusion model. *Ann. Stat.* **28** (2000) 815–836.
[21] A.Yu. Veretennikov, On polynomial mixing bounds for stochastic differential equations. *Stoch. Proc. Appl.* **70** (1997) 115–127.