ON ESTIMATING A DYNAMIC FUNCTION OF STOCHASTIC SYSTEM WITH AVERAGING

LIPTSER, R. AND SPOKOINY, V.

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin, Germany

and

Dept. Electrical Engineering-Systems,
Tel Aviv University,
69978 Tel Aviv, Israel

1991 *Mathematics Subject Classification.* 62G05; Secondary 62M99.

Key words and phrases. fast and slow components, drift and diffusion coefficients, ergodic property, nonparametric estimation, bandwidth selection.

Abstract. We consider a two-scaled diffusion system, when drift and diffusion parameters of a "slow" component are contaminated by an unobservable " fast" one. The goal is to estimate the dynamic function which is defined by averaging the drift coefficient of the "slow" component w.r.t. the stationary distribution of the "fast" one. For estimation we use a locally linear smoother with a data-driven choice of bandwidth. A procedure proposed is fully adaptive and nearly optimal up to a *log log* factor.

1. Introduction

In this paper we consider an estimation problem for a two-scaled diffusion model describing by the following Itô stochastic differential equations (SDE) with respect to independent Wiener processes w_t, W_t

$$\begin{aligned} \mathrm{d}X_t^\varepsilon &= f(X_t^\varepsilon, Y_t^\varepsilon)\mathrm{d}t + g(X_t^\varepsilon, Y_t^\varepsilon)\mathrm{d}w_t, \qquad X_0^\varepsilon = x_0, && (1.1)\\ \varepsilon \mathrm{d}Y_t^\varepsilon &= F(Y_t^\varepsilon) + \sqrt{\varepsilon}G(Y_t^\varepsilon)\mathrm{d}W_t, \qquad Y_0^\varepsilon = y_0. && (1.2)\end{aligned}$$

Here ε is a small parameter and therefore the process Y_t^ε is fast-oscillating. Further X_t^ε (resp. Y_t^ε) is referred to as the slow (resp. the fast) component of the stochastic system (1.1), (1.2).

All the drift and diffusion coefficients f, g, F and G are unknown functions. Moreover, only the "slow" component X_t^ε is observed. With respect to the unobserved "fast" component Y_t^ε we only assume that the corresponding coefficients F, G satisfy some regularity conditions (see e.g. Veretennikov (1992) or assumption (A_5) in Section 3 below). These assumptions ensure that the fast component Y_t^ε forms an ergodic Markov process and its transition probabilities converge rapidly to the stationary density, say $p(y)$. Furthermore, under these conditions the slow component obeys the Bogolubov type averaging principle, that is, it can be approximated by a new diffusion with respect to another Wiener process $\overline{w}_t$ and the same initial condition x_0:

$$\mathrm{d}X_t = \overline{f}(X_t)\mathrm{d}t + \overline{g}(X_t)\mathrm{d}\overline{w}_t.$$

Here the coefficients $\overline{f}, \overline{g}$ are obtained by averaging the original ones with respect to the stationary distribution density $p(y)$ of the fast component Y_t^ε:

$$\overline{f}(x) = \int f(x,y)p(y)dy \quad \text{and} \quad \overline{g}(x) = \Big(\int g^2(x,y)p(y)dy\Big)^{1/2}$$

(see Khasminskii (1980), Freidlin and Wentzell (1984), Veretennikov (1991)).

In this paper we focus on the problem of statistical estimation of *a dynamic function* $\overline{f}(x)$ from observations X_t^ε, $0 \le t \le T$, where T is *the observation time.* Such a problem arises in many applications: tracking of moving objects, description of an incomplete market in financial mathematics, and the like provide examples. To simplify the presentation, we restrict ourselves to the scalar case when both slow and fast components X_t^ε and Y_t^ε are real-valued processes. However a vector case can be considered in the similar way.

The estimation theory for diffusion type processes is well developed under a parametric modeling when underlying functions (drift and diffusion coefficients) are specified up to a value of a finite dimensional parameter (cf. Kutoyants, 1984b). In contrast, a problem of nonparametric estimation is not studied in details. The existing in the literature results concern only statistical inference for ergodic diffusion processes with a small noise or for a large observation time T. The minimax rate of estimating the drift coefficient for such models was obtained by Kutoyants (1984a). Some pertinent results for discrete time models can be found in Doukhan and Ghindes (1980), Collomb and Doukhan (1983), Doukhan and Tsybakov (1993), Delyon and Juditsky (1997). In this paper we do not assume any ergodic property for the slow component. This makes the problem more complicated. Additional difficulties come from the fact that the coefficients of the slow process X^ε_t are contaminated by the unobserved fast component. To our knowledge, nonparametric statistical inference for diffusion models (1.1), (1.2) with averaging has not yet been considered.

We propose a locally linear nonparametric estimator of the dynamic function with a data-driven bandwidth choice and show that this method provides a nearly optimal rate of estimating up to a "*log log*" factor.

The paper is organized as follows. In the next section, we describe a locally linear estimator. Its properties are discussed in Section 3. A data-driven bandwidth selection and its properties are presented in Section 4. All proofs are gathered in Sections 5.

2. **Locally linear estimator**

For the fixed point x, we estimate the value $\overline{f}(x)$ by using locally linear smoothers (cf. Katkovnik (1985), Tsybakov (1986), Fan and Gijbels (1996)).

We begin with some heuristic explanation of the proposed procedure. First let us suppose that functions $f(u,y)$ and $g(u,y)$ are sufficiently smooth in u, at least in some small neighborhood $[x-h,x+h]$ of a point of the interest x. We can therefore approximate for all $u \in [x-h,x+h]$

$$\begin{aligned} f(u,y) &\approx f(x) + f_x(x,y)\,(u-x), \\ g(u,y) &\approx g(x,y) \end{aligned}$$

(here f_x stands for the derivative of $f(x,y)$ in x). Then for all t with $X^\varepsilon_t \in [x-h,x+h]$, the original model equation (1.1) can be approximated by the equation

$$\mathrm{d}X^\varepsilon_t = [f(x,Y^\varepsilon_t) + f_x(x,Y^\varepsilon_t)\,(X^\varepsilon_t - x)]\mathrm{d}t + g(x,Y^\varepsilon_t)\mathrm{d}w_t.$$

In turn, such defined process X^ε_t, due to the averaging principle, can be approximated by a diffusion one with the averaged coefficients

$$\overline{f}(x) = \int f(x,y)\,p(y)\,\mathrm{d}y \quad \text{and} \quad \overline{f}_x(x) = \int f_x(x,y)\,p(y)\,\mathrm{d}y$$

and we arrive at the linear SDE

$$\mathrm{d}X_t = \left[\overline{f}(x) + \overline{f}_x(x)\,(X_t - x)\right]\mathrm{d}t + \overline{g}(x)\,\mathrm{d}\overline{w}_t. \tag{2.1}$$

Now the parameters $\theta_0 = \overline{f}(x)$ and $\theta_1 = \overline{f}_x(x)$ can be estimated by applying the standard maximum likelihood method. The corresponding log-likelihood L_{θ_0,θ_1}

(for the case when only observations $X_t \in [x-h, x+h]$ are taken into account) reads as follows:

$$L_{\theta_0,\theta_1} = \frac{1}{|\overline{g}(x)|^2} \int_0^T [\theta_0 + \theta_1(X_t - x)]\, Q_t \,\mathrm{d}X_t - \frac{1}{2|\overline{g}(x)|^2} \int_0^T [\theta_0 + \theta_1(X_t - x)]^2\, Q_t \,\mathrm{d}t$$

where $Q_t = \mathbf{1}(X_t \in [x-h, x+h]) = \mathbf{1}\left(\frac{|X_t - x|}{h} \le 1\right)$. By maximizing this expression w.r.t. θ_0 and θ_1 we get the estimate $\widetilde{f}(x)$ of $\overline{f}(x)$ in the form

$$\widetilde{f}(x) = \frac{\int_0^T (X_t - x)\, Q_t \,\mathrm{d}X_t - \int_0^T Q_t \,\mathrm{d}X_t \int_0^T Q_t \,\mathrm{d}t}{\int_0^T (X_t - x)^2 Q_t \,\mathrm{d}t \int_0^T Q_t \,\mathrm{d}t - \left(\int_0^T (X_t - x)\, Q_t \,\mathrm{d}t\right)^2}.$$

To construct our estimate, we simply substitute in this expression X_t by the original observations X_t^ε as if they obey the linear approximating equation (2.1). For technical reasons, we also replace the indicator function $\mathbf{1}(|z| \le 1)$ by a smooth function $Q(z)$.

Now we present the formal description of our method. First we introduce a *kernel* Q, which is assumed to be a symmetric, non negative, bounded (for sake of simplicity by 1), compactly supported on $[-1, 1]$, and infinitely differentiable function. One possible example is as follows:

$$Q(u) = \begin{cases} \exp\left\{-\frac{z^2}{1-z^2}\right\}, & |z| \le 1, \\ 0, & |z| > 1. \end{cases}$$

Let us fix also a positive number h called a *bandwidth* and denote $Q_t^\varepsilon = Q\left(\frac{X_t^\varepsilon - x}{h}\right)$. We set

$$\begin{aligned} \mu_{0,T} &= \frac{1}{Th} \int_0^T Q_t^\varepsilon \,\mathrm{d}t, \\ \mu_{1,T} &= \frac{1}{Th^2} \int_0^T (X_t^\varepsilon - x)\, Q_t^\varepsilon \,\mathrm{d}t, \\ \mu_{2,T} &= \frac{1}{Th^3} \int_0^T (X_t^\varepsilon - x)^2 Q_t^\varepsilon \,\mathrm{d}t, \\ D_T &= \mu_{2,T}\mu_{0,T} - \mu_{1,T}^2 \end{aligned}$$

and

$$L_t^\varepsilon = \mu_{2,T} - \mu_{1,T} \frac{X_t^\varepsilon - x}{h}.$$

Then the estimate $\widehat{f}_T(x)$ of $\overline{f}(x)$ is defined by the following formula:

$$\begin{aligned} \widehat{f}_T(x) &= \frac{1}{D_T\, T\, h} \int_0^T L_t^\varepsilon\, Q_t^\varepsilon \,\mathrm{d}X_t^\varepsilon \\ &= \frac{1}{D_T\, T\, h} \int_0^T \left(\mu_{2,T} - \mu_{1,T} \frac{X_t^\varepsilon - x}{h}\right) Q_t^\varepsilon \,\mathrm{d}X_t^\varepsilon. \end{aligned} \tag{2.2}$$

The choice of the bandwidth h is very essential for the quality of estimation. We discuss the problem of the bandwidth selection in Sections 3 and 4.

3. Accuracy of locally linear estimate

Hereafter, the following conditions are assumed to be satisfied.

(A_1) Functions $f = f(x,y)$ and $g = g(x,y)$ ($F = F(y)$ and $G = G(y)$) are Lipschitz continuous in x,y (in y).

(A_2) For some positive constants $g_{\min} \le g_{\max}$

$$g_{\min} \le |g(x,y)| \le g_{\max}, \qquad g_{\min} \le |G(y)| \le g_{\max}.$$

(A_3) Function $f = f(x,y)$ is three times continuously differentiable in x ;

(A_4) Functions $F = F(y)$ and $G = G(y)$ are continuously differentiable (F once, G twice) and their derivatives are continuous and bounded.

(A_5) There exist constants $l > 0$ and $C > 1$ such that for $|y| > C$

$$\begin{aligned} &yF(y) \le -l|y|^2, \\ &F^2(y) - |F'(y)|G^2(y) \ge (1/l)G^2(y). \end{aligned}$$

Under (A$_1$), the Itô equations (1.1), (1.2) possess a unique strong solution, and under (A$_4$), (A$_5$) (see e.g. Khasminskii (1980)) the fast component Y^ε_t is an ergodic diffusion process with the invariant density

$$(3.1) \qquad p(y) = \text{Const.}\ \frac{\exp\left\{2\int\limits_0^y \frac{F(u)}{G^2(u)}\mathrm{d}u\right\}}{G^2(y)}.$$

In the sequel we use also the following notation

$$(3.2) \qquad \Delta_h(x) = \sup_{|u-x|\le h, y\in\mathbb{R}} |f(u,y) - f(x,y) - (u-x)f_x(x,y)|$$

and

$$(3.3) \qquad \sigma_T^2 = \frac{1}{D_T^2\,T^2\,h^2}\int_0^T |L^\varepsilon_t\,Q^\varepsilon_t\,g(X^\varepsilon_t, Y^\varepsilon_t)|^2\,\mathrm{d}t$$

Finally, with given positive constants $\mu_{\max} \ge \mu_{\min}$ and d_1, introduce random events

$$(3.4) \qquad \begin{aligned} A_{1,T} &= \{\mu_{0,T} \le \mu_{\max}\}, \\ A_{2,T} &= \left\{\frac{\mu_{1,T}^2}{\mu_{0,T}\mu_{2,T}} \le d_1\right\}, \\ A_{3,T} &= \left\{\frac{1}{Th}\int_0^T |L^\varepsilon_t\,Q^\varepsilon_t|^2\;dt \ge \mu_{\min}\right\}, \\ A_T &= A_{1,T} \cap A_{2,T} \cap A_{3,T}. \end{aligned}$$

Theorem 3.1. *Let conditions (A_1)–(A_5) be fulfilled and let the values εT and $\varepsilon T\,h^{-2}$ be small if ε is small enough. Then, for every $\lambda \ge 1$ and sufficiently small ε ,*

$$\begin{aligned} &\boldsymbol{P}\left(|\widehat{f}_T(x) - \overline{f}(x)| > (1-d_1)^{-1/2}\Delta_h(x) + \lambda\sigma_T(x),\ A_T\right) \\ &\quad \le C_1\lambda\exp\{-\lambda^2/2\} \end{aligned}$$

where

$$C_1 = e\left[2 + \log\left(\frac{g^2_{\max}\,\mu_{\max}}{g^2_{\min}\,\mu_{\min}}\right)\right].$$

Remark 3.1. Note that the properties of the estimate $\widehat{f}(x)$ is examined on the set A_T only. Such an analysis allows to eliminate irregular cases when, for instance, the trajectory X^ε_t, $0 \le t \le T$, does not pass through the interval $[x-h, x+h]$ (in this situation $\mu_{0,T} = \mu_{1,T} = \mu_{2,T} = D_T = 0$). Our approach admits also the following interpretation: comparing (1.1) with a standard regression model we see that the observable trajectory X^ε_t, $0 \le t \le T$, serves also as a regression design and values D_T, $\sigma_T(x)$ characterize the design regularity in the neighborhood of x. Therefore the constrains from the definition of A_T can be regarded as assumptions on the design which guarantee a nontrivial estimation. The constants $\mu_{\min}$, $\mu_{\max}$ and d_1 may be arbitrary positive numbers and they even might depend on ε and T. The result claims that if small $\mu_{\min}$ or large $q_{\max}$ are taken, the error probability degrades only by a log-factor.

It is also worth to mention that if the coefficients f and g of the slow component obey regularity conditions similar to (A_5), then the process X^ε_t is ergodic as well, and its transition probabilities converge to the stationary distribution as T tends to infinity, see e.g. Veretennikov (1991). It can be easily seen that in this situation the probability of A_T tends to one and we may therefore replace the risk on the set A_T by the unconditional risk on the whole probability space.

Remark 3.2. The result of the theorem claims also that the losses $\widehat{f}(x) - \overline{f}(x)$, being restricted to A_T, are bounded by the sum of two terms: $(1-d_1)^{-1/2}\Delta_h(x)$ and $\lambda\sigma_T(x)$. The first one reflects the error of approximation of $f(u,y)$ by a linear in u function in a small neighborhood $[x-h, x+h]$ of the point x. The second one is in proportion to $\sigma_T(x)$. This value is random (often $\sigma^2_T(x)$ is called the "stochastic variance") but it can be precisely evaluated from observations X_t, $0 \le t \le T$ (see Section 4).

3.1. Quality of estimation under smoothness assumptions

Due to assumption (A$_3$), the function f is twice continuously differentiable with respect to the first argument. Assume now that for u from a small vicinity of x and any y

$$\left|\frac{\partial^2 f(u,y)}{\partial u^2}\right| \le L. \tag{3.5}$$

Then the approximation error $\Delta_h(x)$ from (3.2) is bounded above by $Lh^2/2$. On the other hand, on the set A_T the stochastic variance satisfies the condition

$$\frac{s_{\min}}{\sqrt{Th}} \le \sigma_T \le \frac{s_{\max}}{\sqrt{Th}}$$

(see Lemma 5.1 below). Therefore, following to the standard approach in nonparametric estimation, the bandwidth h can be chosen by balancing the error of approximation and the stochastic error:

$$Lh^2 \asymp \frac{1}{\sqrt{T\,h}},$$

where the symbol "$\asymp$" means equivalence in the order. This leads to the choice $h \asymp (T\,L^2)^{-1/5}$ and hence to rate of the estimation $L^{1/5}T^{-2/5}$ which is optimal in the minimax sense under the smoothness assumptions (3.5), see e.g. Ibragimov and Khasmiskii (1981).

Unfortunately, such a bandwidth selection rule cannot be applied in practice since no information of (3.5) type is available. An adaptive (data-driven) choice of the bandwidth is discussed in the next section.

4. A data-driven bandwidth selection

In this section we consider the problem of the bandwidth selection for the locally linear estimator described in Section 2. It is assumed here that the method of estimation, i.e. the locally linear smoother with the kernel Q, is fixed and the bandwidth h has to be chosen only. We apply the method from Lepski, Mammen and Spokoiny (1997), see also Lepski and Spokoiny (1997).

Let some set $\mathcal{H}$ of all admissible values of bandwidth h be fixed. For technical reasons, we assume that this set is finite and denote by $\#\mathcal{H}$ the number of its elements. Usually $\mathcal{H}$ is taken as a geometric grid of the form

$$\mathcal{H} = \{h = h_{\min}a^k,\ k = 0, 1, 2, \ldots : h \le h_{\max}\},$$

where $h_{\min} \le h_{\max}$ and $a > 1$ are some prescribed constants. To emphasize the dependence of our estimator on h, we write hereafter $\widehat{f}_{h,T}(x)$, $A_{h,T}$, and $\sigma^2_{h,T}$ instead of $\widehat{f}_T(x)$, A_T, and σ^2_T respectively. As in Section 3, we restrict ourselves only to those h from $\mathcal{H}$ for which $A_{h,T}$ is fulfilled.

Our goal now is to select such a value $h \in \mathcal{H}$ which provides for $\widehat{f}_{h,T}(x)$ a minimal in some sense estimation error. To understand which bandwidth might be considered as a "good" one, we apply again the trade-off arguments between the accuracy of approximation and the stochastic error. Without loss of a generality one can assume that for every fixed h the value $\sigma_{h,T}(x)$ is known (it can be exactly evaluated from observations X_t, $0 \le t \le T$, see Subsection 4.2). Due to Lemma 5.1 below, it holds $\sigma^2_{h,T}(x)\,T\,h \in [s^2_{\min}, s^2_{\max}]$. We additionally suppose that $\sigma_{h,T}(x)$ decreases in h (otherwise each h with the property $\exists h' \in \mathcal{H},\ h' < h : \sigma_{h,T}(x) > \sigma_{h',T}(x)$ is excluded from $\mathcal{H}$). Furthermore, for a regular function f, the value $\Delta_h(x)$ is small when h is small, and it increases in h. We may therefore define a "good" bandwidth h^* as the largest possible h from $\mathcal{H}$ such that $\Delta_h(x)$ is still not larger in order than $\sigma_{h,T}(x)$.

Since the function $\Delta_h(x)$ is unknown, this bandwidth h^* is also unknown, and therefore it might be also called an "ideal" bandwidth. We present below an adaptive procedure and show that the corresponding accuracy of estimation is essentially the same as if we knew in advance the "ideal" bandwidth.

Our procedure involves two additional positive parameters λ_1 and λ_2 whose choice will be discussed a bit later. We define the data-driven bandwidth by

$$\begin{aligned}(4.1)\quad \widehat{h}(x) \;=\; &\max\Big\{h \in \mathcal{H} : |\widehat{f}_{h,T}(x) - \widehat{f}_{\eta,T}(x)| \\ &\le \lambda_1\big(\sigma_{h,T}(x) + \sigma_{\eta,T}(x)\big) + 2\lambda_2\sigma_{h,T}(x), \quad \forall \eta \in \mathcal{H}, \eta < h\Big\}.\end{aligned}$$

In other words, the procedure selects the largest value $h \in \mathcal{H}$ for which the corresponding estimate $\widehat{f}_{h,T}(x)$ does not differ essentially from every estimate $\widehat{f}_{\eta,T}(x)$ with smaller bandwidth values $\eta \in \mathcal{H}$.

Finally, we plug the data-driven bandwidth $\widehat{h}(x)$ in the estimate $\widehat{f}_{h,T}(x)$:

$$(4.2)\qquad \widehat{f}_T(x) = \widehat{f}_{\widehat{h}(x),T}(x).$$

In the next theorem we describe some properties of the adaptive estimate $\widehat{f}_T(x)$ restricting ourselves to the set

$$A_T^* = \bigcap_{h \in \mathcal{H}} A_{h,T}.$$

Theorem 4.1. *Let (A_1) through (A_5) be satisfied and let the values εT and $\varepsilon T h_{\min}^{-2}$ be small for sufficiently small ε. Assume also that there exists a bandwidth $h^* \in \mathcal{H}$ such that*

$$(4.3)\qquad \Delta_{h^*}(x) \le (1 - d_1)^{1/2}\lambda_2\sigma_{h^*,T}(x)$$

with the same λ_2 as in (4.1). Then for the estimate $\widehat{f}_T(x)$ defined in (4.1), (4.2) and every $\lambda \ge 1$

$$\begin{aligned}(4.4)\qquad &\boldsymbol{P}\Big(\Big\{|\widehat{f}_T(x) - \overline{f}(x)| > (\lambda + \lambda^*)\sigma_{h^*,T}(x)\Big\} \cap A_T^*\Big) \\ &\le C_1(\#\mathcal{H})^2\lambda_1\exp\{-\lambda_1^2/2\} + C_1\lambda\exp\{-\lambda^2/2\},\end{aligned}$$

where the constant C_1 is defined in Theorem 3.1 and

$$(4.5)\qquad \lambda^* = 2\lambda_1 + 3\lambda_2.$$

Remark 4.1. A choice of parameters λ_1, λ_2 from (4.1) plays an important role. The bound in (4.4) states that the probability for $|\widehat{f}_T(x) - \overline{f}(x)|$ of being large is small, provided that the value $(\#\mathcal{H})^2\lambda_1\exp\{-\lambda_1^2/2\}$ is sufficiently small as well. This leads to a choice

$$\lambda_1 \approx \sqrt{4\log(\#\mathcal{H}) + \lambda^2}$$

so that

$$(\#\mathcal{H})^2\lambda_1\exp\{-\lambda_1^2/2\} \approx \exp\{-\lambda^2/2\}.$$

If $\mathcal{H}$ is taken in the form of a geometric grid then we get $\#\mathcal{H} \approx \log_a(h_{\max}/h_{\min})$. Therefore taking $h_{\max} \approx T$ and $h_{\min} \approx 1$, we arrive at

$$\lambda_1 \approx \sqrt{4\log\log T + \lambda^2}.$$

We have much more degree of freedom in the choice of λ_2. This parameter controls the balance between the accuracy of approximation and the stochastic error in the definition of the “ideal” bandwidth h^* (see (4.3)). The results from Lepski and Spokoiny (1997) motivate the choice $\lambda_2 = \text{Const}\,\lambda_1$ (see also the next subsection).

At the same time, Lepski and Levit (1997) have showed that for a very smooth function f, a relevant choice is $\lambda_2 = o(\lambda_1)$.

4.1. The rate of adaptive estimation

Next, we compare the accuracy of the adaptive procedure (4.1) with the "optimal" one, designed for the case of known smoothness properties of the underlying function f (see Subsection 3.1).

Assume (3.5). Then $\Delta_h(x) \le Lh^2/2$ and the constraints $\Delta_h(x) \le \lambda_2\sigma_{h,T}(x)$ and $s_{\min}(hT)^{-1/2} \le \sigma_{h,T} \le s_{\max}(hT)^{-1/2}$ provide inequality (4.3) with

$$h^* \asymp (TL^2\lambda_2^2)^{-1/5}.$$

Hence, for the above indicated choice $\lambda_1 \asymp \sqrt{\log\log T}$ and for $\lambda_2 \approx \lambda_1$, by Theorem 4.1 we obtain the rate of adaptive estimation

$$(\lambda + 2\lambda_1 + 3\lambda_2)\sigma_{h^*,T}(x) \asymp L^{1/5}\left(\frac{\log\log T}{T}\right)^{2/5}. \tag{4.6}$$

At the same time the "optimal" choice of the bandwidth leads to the rate $L^{1/5}T^{-2/5}$, see Section 3.1. Thus, the adaptive rate is worse than the "optimal" one by some *log log*-factor only.

Note that due to Lepski (1990) and Brown and Low (1992) (see also Lepski and Spokoiny (1997)), for the problem of pointwise adaptive estimation, the optimal adaptive rate has to be worse than the optimal one by a *log*-factor. This is not in contradiction with our conclusions since the above-mentioned results have been obtained for the case of the loss function $w(x) = |x|^p$ with $p > 0$ whereas we consider a bounded loss function. It can be also shown that the obtained rate is optimal in the problem of poinwise adaptive estimation with a bounded loss function, cf. Spokoiny (1996).

4.2. Estimation of $\sigma_{h,T}$

We use a well known fact that a "diffusion parameter" of a continuous Itô process can be correctly recovered from observations of the process.

We use here this fact to find $\sigma_{h,T}$. Let us note that $(X_t^\varepsilon)_{t\ge 0}$ is a continuous semimartingale with the predictable quadratic variation $\langle X^\varepsilon\rangle_t = \int_0^t g^2(X_s^\varepsilon, Y_s^\varepsilon)\mathrm{d}s$. Introduce a new continuous semimartingale

$$Z_t = \int_0^t L_s^\varepsilon\, Q_s^\varepsilon\, \mathrm{d}X_s^\varepsilon.$$

It is well known that its predictable quadratic variation is defined by the formula $\langle Z\rangle_t = \int_0^t |L_s^\varepsilon\, Q_s^\varepsilon|^2\, \mathrm{d}\langle X^\varepsilon\rangle_s$ and therefore we have $\langle Z\rangle_T = \sigma_{h,T}^2\, D_T^2\, T^2\, h^2$, that is, since $\langle Z\rangle_T$ and $D_T^2 T^2$ are generated by $X_t^\varepsilon, t \le T$ only, the parameter $\sigma_{h,T}^2$ is reconstructed exactly via observation of the "slow" component. By the Itô formula

[1]

$$\langle Z\rangle_T = Z_T^2 - 2\int_0^T Z_s \mathrm{d}Z_s$$

which implies

$$\sigma_{h,T}^2 = \frac{Z_T^2 - 2\int_0^T Z_t\,\mathrm{d}Z_t}{D_T^2 T^2 h^2}.$$

5. Proofs

In this section we proof Theorems 3.1 and 4.1. For a generic positive constant, a notation 'ℓ' will be used hereafter.

5.1. Decomposition of $\widehat{f}_T(x)$

Due to (2.2) and (1.1), the estimate $\widehat{f}_T(x)$ can be represented in the form

$$\begin{aligned}
\widehat{f}_T(x) &= \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,\mathrm{d}X_t^\varepsilon \\
&= \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,f(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t \\
&\quad + \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,g(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}w_t
\end{aligned} \tag{5.1}$$

with $Q_t^\varepsilon = Q\left(\frac{X_t^\varepsilon - x}{h}\right)$ and $L_t^\varepsilon = \mu_{2,T} - \mu_{1,T}\frac{X_t^\varepsilon - x}{h}$. Next, we make use of the fact that $Q_t^\varepsilon \equiv 0$ for $|X_t^\varepsilon - x| > h$ and of a decomposition

$$f(X_t^\varepsilon, Y_t^\varepsilon) = f(x, Y_t^\varepsilon) + f_x(x, Y_t^\varepsilon)(X_t^\varepsilon - x) + \text{remainder}.$$

Substituting the right side of this decomposition in (5.1) we get

$$\begin{aligned}
&\widehat{f}_T(x) \\
&= \frac{1}{D_T\,T\,h}\int_0^T f(x, Y_t^\varepsilon)\,L_t^\varepsilon\,Q_t^\varepsilon\,\mathrm{d}t + \frac{1}{D_T\,T\,h}\int_0^T f_x(x, Y_t^\varepsilon)\,L_t^\varepsilon\,(X_t^\varepsilon - x)\,Q_t^\varepsilon\,\mathrm{d}t \\
&\quad + \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,g(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}w_t + r_T \\
&= f_T(x) + \zeta_T + \xi_T + r_T
\end{aligned} \tag{5.2}$$

[1] for more details see e.g. Liptser and Shiryaev (1989).

where

$$
\begin{aligned}
f_T(x) &= \frac{1}{D_T\,T\,h}\int_0^T f(x,Y_t^\varepsilon)\,L_t^\varepsilon\,Q_t^\varepsilon\,\mathrm{d}t, \\
\zeta_T &= \frac{1}{D_T\,T\,h}\int_0^T f_x(x,Y_t^\varepsilon)\,L_t^\varepsilon\,(X_t^\varepsilon - x)\,Q_t^\varepsilon\,\mathrm{d}t \\
\xi_T &= \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,g(X_t^\varepsilon,Y_t^\varepsilon)\,\mathrm{d}w_t, \\
r_T &= \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,\Big[f(X_t^\varepsilon,Y_t^\varepsilon) \\
&\qquad -f(x,Y_t^\varepsilon) - f_x(x,Y_t^\varepsilon)(X_t^\varepsilon - x)\Big]\,\mathrm{d}t.
\end{aligned}
\tag{5.3}
$$

5.2. Upper bound for r_T

Let $\Delta_h(x)$ and $A_{2,T}$ be defined in (3.2) and (3.4) respectively. We aim to show that on the set $A_{2,T}$ it holds $r_T \le (1-d_1)^{-1/2}\Delta_h(x)$. In fact, due to (3.2)

$$
r_T \le \frac{\Delta_h(x)}{D_T\,T\,h}\int\limits_0^T |L_t^\varepsilon|\,Q_t^\varepsilon\,\mathrm{d}t.
$$

Further, by the Cauchy-Schwarz inequality

$$
\int_0^T |L_t^\varepsilon|\,Q_t^\varepsilon\,\mathrm{d}t \le \left[\int_0^T |L_t^\varepsilon|^2\,Q_t^\varepsilon\,\mathrm{d}t \int_0^T Q_t^\varepsilon\,\mathrm{d}t\right]^{1/2}
$$

and so, by using the equality $\int_0^T Q_t^\varepsilon\,\mathrm{d}t = T\,h\,\mu_{0,T}$ we obtain

$$
\begin{aligned}
&\int_0^T |L_t^\varepsilon|^2\,Q_t^\varepsilon\,\mathrm{d}t \\
&= \int_0^T \left(\mu_{2,T} - \mu_{1,T}\frac{X_t^\varepsilon - x}{h}\right)^2 Q_t^\varepsilon\,\mathrm{d}t \\
&= \mu_{2,T}^2\int_0^T Q_t^\varepsilon\,\mathrm{d}t - \frac{2\mu_{2,T}\mu_{1,T}}{h}\int_0^T (X_t^\varepsilon - x)Q_t^\varepsilon\,\mathrm{d}t + \frac{\mu_{1,T}^2}{h^2}\int_0^T (X_t^\varepsilon - x)^2 Q_t^\varepsilon\,\mathrm{d}t \\
&= T\,h\,(\mu_{2,T}^2\mu_{0,T} - \mu_{2,T}\mu_{1,T}^2) \\
&= T\,h\,\mu_{2,T}\,D_T.
\end{aligned}
\tag{5.4}
$$

Therefore,

$$
\begin{aligned}
(5.5) \qquad r_T &\le \frac{\Delta_h(x)}{D_T\,T\,h}\left(T\,h\,\mu_{2,T}\,D_T\,T\,h\,\mu_{0,T}\right)^{1/2} \\
&= \Delta_h(x)\left(\frac{\mu_{2,T}\,\mu_{0,T}}{D_T}\right)^{1/2} \\
&= \Delta_h(x)\left(\frac{\mu_{2,T}\,\mu_{0,T}}{\mu_{2,T}\mu_{0,T}-\mu_{1,T}^2}\right)^{1/2} \\
&\le \Delta_h(x)\left(\frac{1}{1-d_1}\right)^{1/2}
\end{aligned}
$$

as required.

5.3. Upper bound for $\boldsymbol{\xi_T}$

We study here some properties of the "stochastic term"

$$\xi_T = \frac{1}{D_T\,T\,h}\int_0^T L_t^\varepsilon\,Q_t^\varepsilon\,g(X_t^\varepsilon,Y_t^\varepsilon)\,\mathrm{d}w_t.$$

Namely, we intend to show that the probability of the event $\{\xi_T > \lambda\sigma_T\}$ with σ_T from (3.3) is small provided that λ is large enough.

Denote by $M_t = \int_0^t L_s^\varepsilon\,Q_s^\varepsilon\,g(X_s^\varepsilon,Y_t^\varepsilon)\,\mathrm{d}w_s$. The Itô integral M_t forms a continuous local martingale with the predictable quadratic variation

$$(5.6) \qquad \langle M\rangle_t = \int_0^t |L_s^\varepsilon\,Q_s^\varepsilon\,g(X_s^\varepsilon,Y_t^\varepsilon)|^2\,\mathrm{d}s \;(:= V_T^2),$$

(see e.g. Liptser and Shiryayev (1989)). Note that $\{\xi_T > \lambda\sigma_T\} = \{M_T > \lambda V_T\}$. We give below the bound for $\boldsymbol{P}(M_T > \lambda V_T)$ having an independent interest.

Proposition 5.1. *Let $(M_t)_{t\ge0}$ be a continuous local martingale ($M_0 = 0$) with a predictable quadratic variation $(V_t^2)_{t\ge0}$. Then, with every positive $V_{\min} < V_{\max}$ and $\lambda \ge 1$, it holds*

$$\boldsymbol{P}\Big(M_T > \lambda V_T, V_{\min} \le V_T \le V_{\max}\Big) \le 2\lambda\Big[\log\frac{V_{\max}}{V_{\min}} + 1\Big]\exp\Big(-\frac{\lambda^2}{2}+\frac{1}{2}\Big).$$

Proof. With $\gamma \in \boldsymbol{R}$ let us set $R_t(\gamma) = \exp\Big(\gamma M_t - \gamma^2 V_t^2/2\Big)$. By the Itô formula we find $R_t(\gamma) = 1 + \int\limits_0^t \gamma R_s(\gamma)\mathrm{d}M_s$ and so, $R_t(\gamma)$ is a local martingale as well. Being positive, $R_t(\gamma)$ is a supermartingale (see Problem 1.4.4 in Liptser and Shiryayev (1986)). Hence, for every $T > 0$,

$$(5.7) \qquad \boldsymbol{E}R_T(\gamma) \le 1.$$

Given $a > 1$, introduce numbers $v_k = V_{\min}a^k$, and define random events $C_k = \{v_k \le V_T < v_{k+1}\}$, $k = 0,1,\dots$. Now for every $k \ge 0$, bound (5.7) with $\gamma = \frac{\lambda}{av_{k+1}}$

implies

$$\begin{aligned}
1 &\geq \boldsymbol{E} R_T(\gamma)\mathbf{1}(M_T > \lambda V_T, C_k) \\
&= \boldsymbol{E} \exp\Big(\gamma M_T - \gamma^2 V_T^2/2\Big)\mathbf{1}(M_T > \lambda V_T, C_k) \\
&\geq \boldsymbol{E} \exp\Big(\gamma \lambda V_T - \gamma^2 V_T^2/2\Big)\mathbf{1}(M_T > \lambda V_T, C_k) \\
&\geq \boldsymbol{E} \exp\Big(\inf_{v_k \leq v \leq v_{k+1}} \Big[\frac{\lambda^2 v}{v_{k+1}} - \frac{\lambda^2 v^2}{2v_{k+1}^2}\Big]\Big)\mathbf{1}(M_T > \lambda V_T, C_k) \\
&= \exp\Big(\lambda^2(a - a^2/2)\Big) P\Big(M_T > \lambda V_T, C_k\Big)
\end{aligned}$$

and therefore

$$\boldsymbol{P}\Big(M_T > \lambda V_T, C_k\Big) \leq \exp\Big(-\lambda^2(a - a^2/2)\Big).$$

This yields

$$\boldsymbol{P}(M_T > \lambda V_T, V_{\min} \leq V_T \leq V_{\max}) \leq \sum_{k=0}^{K} \boldsymbol{P}(M_T > \lambda V_T, C_k)$$

with $K = \log_a(V_{\max}/V_{\min})$, and we get

$$\boldsymbol{P}(M_T > \lambda V_T, V_{\min} \leq V_T \leq V_{\max}) \leq \Big[\log_a \frac{V_{\max}}{V_{\min}} + 1\Big] \exp\{-\lambda^2(a - a^2/2)\}.$$

We finally set $a = 1 + 1/\lambda$ so that $\lambda^2(a - a^2/2) = (\lambda^2 - 1)/2$ and use the obvious inequality $\log(1 + 1/\lambda) \geq 1/(2\lambda)$ for $\lambda \geq 1$. This gives

$$\log_a \frac{V_{\max}}{V_{\min}} \leq 2\lambda \log \frac{V_{\max}}{V_{\min}}$$

and the assertion follows.

□

To apply this proposition for estimating the probability of the event $\{\xi_T > \lambda\sigma_T\} = \{M_T > \lambda V_T\}$, we have to specify lower and upper bounds for V_T. This is done in the next lemma.

Lemma 5.1. *On the set* A_T *, it holds*

$$v_{\min}^2 \leq \frac{V_T^2}{Th} \leq v_{\max}^2$$
$$s_{\min}^2 \leq \sigma_T^2 Th \leq s_{\max}^2$$

where

$$\begin{aligned}
&v_{\min}^2 = g_{\min}^2 \, \mu_{\min}, && s_{\min}^2 = g_{\min}^2 \, \mu_{\min} \, \mu_{\max}^{-4}, \\
(5.8) \qquad & \\
&v_{\max}^2 = g_{\max}^2 \, \mu_{\max}, && s_{\max}^2 = g_{\max}^2 (1 - d_1)^{-2} \, \mu_{\max} \, \mu_{\min}^{-4}.
\end{aligned}$$

Proof. Recall that the kernel Q is compactly supported on $[-1, 1]$ and bounded above by 1. For all t this implies $|L_t^\varepsilon \, Q_t^\varepsilon|^2 \leq |L_t^\varepsilon|^2 \, Q_t^\varepsilon \leq \, Q_t^\varepsilon$ and hence

$$\frac{1}{Th} \int_0^T |L_t^\varepsilon \, Q_t^\varepsilon|^2 \, \mathrm{d}t \leq \mu_{2,T} \leq \mu_{0,T}.$$

Now, restricting ourselves to the set A_T, we have $\mu_{2,T} \le \mu_{0,T} \le \mu_{\max}$ and also $\mu_{0,T} \ge \mu_{2,T} \ge \frac{1}{Th}\int_0^T |L_t^\varepsilon\, Q_t^\varepsilon|^2\, \mathrm{d}t \ge \mu_{\min}$. Since $V_T^2 = \int_0^T |L_t^\varepsilon\, Q_t^\varepsilon g(X_t^\varepsilon, Y_t^\varepsilon)|^2\, \mathrm{d}t$, we get in view of assumption (A_2)

$$g_{\min}^2\, \mu_{\min} \le \frac{V_T^2}{Th} \le g_{\max}^2\, \mu_{\max}.$$

Similarly,

$$\begin{aligned} D_T &= \mu_{2,T}\mu_{0,T} - \mu_{1,T}^2 \le \mu_{2,T}\mu_{0,T} \le \mu_{\max}^2, \\ D_T &= \mu_{2,T}\mu_{0,T} - \mu_{1,T}^2 \ge (1-d_1)\mu_{2,T}\mu_{0,T} \ge (1-d_1)\mu_{\min}^2 \end{aligned}$$

and

$$\begin{aligned} \sigma_T^2(x) &\le g_{\max}^2 \frac{\int_0^T |L_t^\varepsilon\, Q_t^\varepsilon|^2\, \mathrm{d}t}{T^2h^2D_T^2} \le \frac{g_{\max}^2\mu_{\max}}{(1-d_1)^2\mu_{\min}^4 Th}, \\ \sigma_T^2(x) &\ge g_{\min}^2 \frac{\int_0^T |L_t^\varepsilon\, Q_t^\varepsilon|^2\, \mathrm{d}t}{T^2h^2D_T^2} \ge \frac{g_{\min}^2\mu_{\min}}{\mu_{\max}^4 Th}, \end{aligned}$$

and the assertion follows. □

Coupled these bounds with Proposition 5.1 we arrive at the main result of this subsection ($v_{\max}$ and $v_{\min}$ are defined in (5.8)):

$$\begin{aligned} &\boldsymbol{P}\left(\xi_T > \left(\lambda - \frac{1}{2\lambda}\right)\sigma_T, A_T\right) \\ &\quad\le 2\lambda\left[1 + \log\frac{v_{\max}}{v_{\min}}\right]\exp\left\{-\frac{1}{2}\left(\lambda - \frac{1}{2\lambda}\right)^2 + \frac{1}{2}\right\} \\ &\quad< e\lambda\left[\log\frac{v_{\max}^2}{v_{\min}^2} + 2\right]\exp\{-\lambda^2/2\}. \end{aligned} \tag{5.9}$$

5.4. Upper bound for $f_T(x) - \overline{f}(x)$

Recall that $f_T(x) = \frac{1}{D_T T h}\int_0^T f(x, Y_t^\varepsilon)\, L_t^\varepsilon\, Q_t^\varepsilon\, \mathrm{d}t$. Due to the definition of L_t^ε and Q_t^ε, we obtain

$$\begin{aligned} &f_T(x) \\ &= \frac{1}{D_T T h}\int_0^T f(x, Y_t^\varepsilon)\left(\mu_{2,T} - \mu_{1,T}\frac{X_t^\varepsilon - x}{h}\right) Q\left(\frac{X_t^\varepsilon - x}{h}\right)\mathrm{d}t \\ &= \frac{\mu_{2,T}}{D_T T h}\int_0^T f(x, Y_t^\varepsilon)\, Q\left(\frac{X_t^\varepsilon - x}{h}\right)\mathrm{d}t \\ &\quad - \frac{\mu_{1,T}}{D_T T h}\int_0^T f(x, Y_t^\varepsilon)\,\frac{X_t^\varepsilon - x}{h}\, Q\left(\frac{X_t^\varepsilon - x}{h}\right)\mathrm{d}t. \end{aligned} \tag{5.10}$$

We apply now a large deviation type estimation for the two scaled diffusion model (1.1), (1.2).

Proposition 5.2 (Liptser and Spokoiny, 1997). *Let (A_1)–(A_5) hold and let $\Psi^\varepsilon(u)$ be a twice continuously differentiable function such that for some constants C_0^ε, C_1^ε, depending on ε and for all u*

$$\begin{aligned} |\Psi^\varepsilon| &\le C_0^\varepsilon, \\ |\dot\Psi^\varepsilon(u)| + |\ddot\Psi^\varepsilon(u)| &\le C_1^\varepsilon, \end{aligned}$$

where $\dot\Psi^\varepsilon(u)$ and $\ddot\Psi^\varepsilon(u)$ stand for the first and second derivative of $\Psi^\varepsilon(u)$. Let also $a = a(y)$ be a continuously differentiable function with a bounded derivative. Denote $\overline{a} = \int_{\mathbb{R}} a(y)\, p(y)\,\mathrm{d}y$, where $p(y)$ is the invariant density of the fast component Y_t^ε, and define

$$U_{T^\varepsilon}^\varepsilon = \int_0^{T^\varepsilon} [a(Y_t^\varepsilon) - \overline{a}]\, \Psi^\varepsilon(X_t^\varepsilon)\,\mathrm{d}t.$$

If

$$\begin{aligned} \lim_{\varepsilon\to 0} T^\varepsilon \varepsilon &= 0, \\ \lim_{\varepsilon\to 0} (C_0^\varepsilon)^2 \sqrt{\varepsilon} &= 0, \\ \lim_{\varepsilon\to 0} (C_1^\varepsilon)^2 T^\varepsilon \varepsilon &= 0, \end{aligned}$$

then for every positive $z > 0$ and $0 < \kappa < 1/2$

$$\lim_{\varepsilon\to 0} (\varepsilon T^\varepsilon)^{1-2\kappa} \log \boldsymbol{P}\left((\varepsilon T^\varepsilon)^{-\kappa} |U_{T^\varepsilon}^\varepsilon| > z C_0^\varepsilon \right) \le -\frac{z^2}{2\gamma},$$

where

$$\gamma = \int_{\boldsymbol{R}} v^2(y)\, G^2(y)\, p(y)\,\mathrm{d}y,$$

and the function $v(y)$ is defined by

$$v(y) = \frac{2}{G^2(y)\, p(y)} \int_{\infty}^{y} [a(s) - \overline{a}]\, p(s)\,\mathrm{d}s.$$

Corollary 5.1. *For ε small enough and $\kappa_1 < 1 - 2\kappa$*

$$\boldsymbol{P}\Big(|U_{T^\varepsilon}^\varepsilon| > C_0^\varepsilon (\varepsilon T^\varepsilon)^\kappa \Big) < \exp\left\{ -\frac{1}{(\varepsilon T^\varepsilon)^{\kappa_1}} \right\}.$$

For fixed x, we apply now Corollary 5.1 for $a(y) = f(x,y)$ and $\Psi^\varepsilon(u) = Q\left(\frac{u-x}{h}\right)$. Under the assumptions of the theorem this function fulfills the conditions of Proposition 5.2 with $C_0^\varepsilon \equiv 1$ and $C_1^\varepsilon \equiv \ell h^{-2}$. Since $\overline{a} = \int_{\mathbb{R}} f(x,y)\, p(y)\,\mathrm{d}y = \overline{f}(x)$, by Corollary 5.1 we conclude that

$$\boldsymbol{P}\left(\left| \int_0^T \left[f(x, Y_t^\varepsilon) - \overline{f}(x) \right] Q\left(\frac{X_t^\varepsilon - x}{h} \right) \mathrm{d}t \right| \le (\varepsilon T)^\kappa \right) < \exp\left\{ -(\varepsilon T^\varepsilon)^{-\kappa_1} \right\}.$$

Next, it is easy to see that function $\Psi_1^\varepsilon(z) = \frac{z-x}{h} Q\left(\frac{z-x}{h}\right)$ fulfills the conditions of Proposition 5.2 with the same constants $C_0^\varepsilon \equiv 1$ and $C_1^\varepsilon \equiv \ell h^{-2}$. Hence again

$$\boldsymbol{P}\left(\left| \int_0^T [f(x, Y_t^\varepsilon) - \overline{f}(x)] \frac{X_t^\varepsilon - x}{h} Q\left(\frac{X_t^\varepsilon - x}{h} \right) \mathrm{d}t \right| \le (\varepsilon T)^\kappa \right) < \exp\left\{ -(\varepsilon T^\varepsilon)^{-\kappa_1} \right\}.$$

Coupling these estimates and using (5.10) and the equality

$$\begin{aligned} \overline{f}(x) &= \overline{f}(x)\,\frac{\mu_{2,T}\mu_{0,T}-\mu_{1,T}^2}{D_T} \\ &= \frac{\mu_{2,T}}{D_T\,T\,h}\int_0^T \overline{f}(x)\,Q\left(\frac{X_t^\varepsilon - x}{h}\right)\mathrm{d}t \\ &\qquad - \frac{\mu_{1,T}}{D_T\,T\,h}\int_0^T \overline{f}(x)\,\frac{X_t^\varepsilon - x}{h}\,Q\left(\frac{X_t^\varepsilon - x}{h}\right)\mathrm{d}t \end{aligned}$$

we obtain, provided that $h^{-2}(\varepsilon T)$ is small for small ε, that

$$\boldsymbol{P}\left(|f_T(x)-\overline{f}(x)| \le (\varepsilon T)^{\kappa}\frac{\mu_{2,T}+|\mu_{1,T}|}{D_T\,T\,h}\right) < 2\exp\left\{-(\varepsilon T^{\varepsilon})^{-\kappa_1}\right\}.$$

Next, by Lemma 5.1 on the set A_T the inequalities hold: $\mu_{2,T} \le \mu_{\max}$, $|\mu_{1,T}| \le \mu_{\max}$ and $D_T \ge (1-d_1)\mu_{\min}^2$. Therefore

$$\frac{\mu_{2,T}+|\mu_{1,T}|}{D_T} \le \frac{2\mu_{\max}}{(1-d_1)\,\mu_{\min}^2} \le \ell$$

and hence

$$\boldsymbol{P}\left(|f_T(x)-\overline{f}(x)| > \ell(Th)^{-1}(\varepsilon T)^{\kappa}, A_T\right) < 2\exp\left\{-(\varepsilon T)^{-\kappa_1}\right\}. \tag{5.11}$$

If ε is small enough, then we get in view of Lemma 5.1 on the set A_T

$$\frac{1}{4\lambda}\sigma_T(x) \ge \frac{s_{\min}}{4\lambda\sqrt{T\,h}} \ge \frac{\ell(\varepsilon T)^{\kappa}}{T\,h}.$$

Along with (5.11) this gives for sufficiently small ε

$$\boldsymbol{P}\left(|f_T(x)-\overline{f}(x)| > \frac{1}{4\lambda}\sigma_T(x), A_T\right) < 2\exp\left\{-(\varepsilon T)^{-\kappa_1}\right\}. \tag{5.12}$$

5.5. Upper bound for ζ_T

Recall that $\zeta_T = \frac{1}{D_T\,T}\int_0^T f_x(x, Y_t^\varepsilon)\,\frac{(X_t^\varepsilon - x)}{h}\,L_t^\varepsilon\,Q_t^\varepsilon\,\mathrm{d}t$. We again apply Corollary 5.1, now with $a(z) = f_x(x,z)$. Since

$$\begin{aligned} \int_0^T (X_t^\varepsilon - x)\,L_t^\varepsilon\,Q_t^\varepsilon\,\mathrm{d}t &= \int_0^T (X_t^\varepsilon - x)\left(\mu_{2,T}-\mu_{1,T}\frac{X_t^\varepsilon - x}{h}\right)Q_t^\varepsilon\,\mathrm{d}t \\ &= T\,h^2(\mu_{2,T}\mu_{1,T}-\mu_{1,T}\mu_{2,T}) = 0, \end{aligned}$$

we get in the same line as for evaluating $f_T(x)-\overline{f}(x)$

$$\boldsymbol{P}\left(|\zeta_T| > \frac{1}{4\lambda}\sigma_T(x), A_T\right) \le \exp\left\{-(\varepsilon T)^{-\kappa_1}\right\}.$$

We are now in the position to complete the proof of Theorem 3.1. Decomposition (5.2) along with (5.5), (5.9) and (5.12) implies

$$\begin{aligned} &\boldsymbol{P}\left(\left|\widehat{f}_T(x)-\overline{f}(x)\right| > (1-d_1)^{-1/2}\Delta_h(x)+\lambda\sigma_T(x),\ A_T\right) \\ &\qquad\qquad < C_1\lambda\exp\{-\lambda^2/2\}+3\exp\left\{-(\varepsilon T)^{-\kappa_1}\right\} \end{aligned}$$

and, since the second summand in the right side is exponentially small for small ε, the statement of the theorem holds. □

5.6. Proof of Theorem 4.1

Let h^* be shown in the theorem. Obviously

$$\begin{aligned}&\boldsymbol{P}\left(\left|\widehat{f}_T(x)-\overline{f}(x)\right|>(\lambda+\lambda^*)\sigma_{h^*,T}(x),\ A_T^*\right)\\&\quad\le\boldsymbol{P}\left(\left|\widehat{f}_T(x)-\overline{f}(x)\right|>(\lambda+\lambda^*)\sigma_{h^*,T}(x),\ \widehat{h}\ge h^*,\ A_T^*\right)+\boldsymbol{P}\left(\widehat{h}<h^*,\ A_T^*\right).\end{aligned}$$

Each summand in the right hand side of this inequality is evaluated separately. Since $\sigma_{h,T}(x)$ decreases in h, it holds on the set $\{\widehat{h}\ge h^*\}$ in view of the definition of $\widehat{h}$

$$|\widehat{f}_{\widehat{h},T}(x)-\widehat{f}_{h^*,T}(x)|\le\lambda_1\big(\sigma_{\widehat{h},T}(x)+\sigma_{h^*,T}(x)\big)+2\lambda_2\sigma_{\widehat{h},T}(x)\le 2(\lambda_1+\lambda_2)\sigma_{h^*,T}(x).$$

Further, using the inequality $(1-d_1)^{-1/2}\Delta_{h^*}(x)\le\lambda_2\sigma_{h^*,T}$ and applying Theorem 3.1, we get

$$\begin{aligned}&\boldsymbol{P}\left(|\widehat{f}_{h^*,T}(x)-\overline{f}(x)|>(\lambda_2+\lambda)\sigma_{h^*,T}(x),A_T^*\right)\\&\quad\le\boldsymbol{P}\left(|\widehat{f}_{h^*,T}(x)-\overline{f}(x)|>\lambda\sigma_{h^*,T}(x)+\frac{\Delta_{h^*}(x)}{\sqrt{1-d_1}},A_T^*\right)\\&\quad\le C_1\lambda\exp\left\{-\frac{\lambda^2}{2}\right\}.\end{aligned}$$

Coupled with the previous inequality this implies

$$\begin{aligned}\text{(5.13)}\qquad&\boldsymbol{P}\left(|\widehat{f}_T(x)-\overline{f}(x)|>(\lambda+\lambda^*)\sigma_{h^*,T}(x),A_T^*,\widehat{h}\ge h^*\right)\\&\quad\le C_1\lambda\exp\left\{-\frac{\lambda^2}{2}\right\}.\end{aligned}$$

Therefore, it remains to evaluate $\boldsymbol{P}(\widehat{h}<h^*)$ only. Due to the definition of $\widehat{h}$

$$\begin{aligned}&\{\widehat{h}<h^*\}=\\&\bigcup_{h\in\mathcal{H}:h<h^*}\ \bigcup_{\eta\in\mathcal{H}:\eta<h}\{|\widehat{f}_{h,T}(x)-\widehat{f}_{\eta,T}(x)|>\lambda_1\big(\sigma_{h,T}(x)+\sigma_{\eta,T}(x)\big)+2\lambda_2\sigma_{h,T}(x)\}.\end{aligned}$$

Note also that for every $\eta,h\in\mathcal{H}$ with $\eta<h<h^*$ it holds

$$\begin{aligned}(1-d_1)^{-1/2}\Delta_h(x)&\le(1-d_1)^{-1/2}\Delta_{h^*}(x)\le\lambda_2\sigma_{h^*,T}(x)\le\lambda_2\sigma_{h,T}(x)\\(1-d_1)^{-1/2}\Delta_\eta(x)&\le(1-d_1)^{-1/2}\Delta_{h^*}(x)\le\lambda_2\sigma_{h^*,T}(x)\le\lambda_2\sigma_{h,T}(x).\end{aligned}$$

Hence, by Theorem 3.1

$$\begin{aligned}
&\boldsymbol{P}\left(|\widehat{f}_{h,T}(x)-\widehat{f}_{\eta,T}(x)|>\lambda_1\big(\sigma_{h,T}(x)+\sigma_{\eta,T}(x)\big)+2\lambda_2\sigma_{h,T}(x),A_T^*\right)\\
&\qquad\le\boldsymbol{P}\left(|\widehat{f}_{h,T}(x)-\overline{f}(x)|>\lambda_1\sigma_{h,T}(x)+\frac{\Delta_h(x)}{\sqrt{1-d_1}},A_T^*\right)\\
&\qquad\quad+\boldsymbol{P}\left(|\widehat{f}_{\eta,T}(x)-\overline{f}(x)|>\lambda_1\sigma_{\eta,T}(x)+\frac{\Delta_\eta(x)}{\sqrt{1-d_1}},A_T^*\right)\\
&\qquad\le 2C_1\lambda_1\exp\left\{-\frac{\lambda_1^2}{2}\right\}.
\end{aligned}$$

Clearly an amount of pairs $\eta,h\in\mathcal{H}$ with above-mentioned property $\eta<h<h^*$ is at most $(\#\mathcal{H})^2/2$. Therefore

$$\boldsymbol{P}\left(\widehat{h}<h^*\right)\le(\#\mathcal{H})^2C_1\lambda_1\exp\left\{-\frac{\lambda_1^2}{2}\right\}$$

and the required assertion follows in view of (5.13). □

References

[1] Brown, L.D. and Low, M.G. (1992). Superefficiency and lack of adaptability in functional estimation. *Technical Report*, Cornell University.

[2] G. Collomb and P. Doukhan (1983). Estimation non parametrique de la fonction d'autoregression d'un processus stationnaire et phi melangeant: risques quadratiques pour la methode du noyau, *C. R. Acad. Sci.*, Paris, Ser. I 296, 859-862 .

[3] Delyon, B. and Juditsky, A. (1997). On minimax prediction for nonparametric autoregressive models. Unpublished manuscript.

[4] P. Doukhan and M. Ghindes (1980). Estimations dans le processus "$X_{n+1}=f(X_n)+$ *epsilon*$_n$", C. R. Acad. Sci., Paris, Ser. A 291, 61-64.

[5] Doukhan, P. and Tsybakov, A.B. (1993). Nonparametric recurrent estimation in nonlinear ARX models. *Problemy-Peredachi-Informatsii* **29**, no. 4, 24–34. Translation: *Problems Inform. Trans.* **29**, no. 4, 318–327.

[6] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall, London.

[7] Freidlin, M.I., Wentzell A.D. (1984). *Random Perturbations of Dynamical Systems.* N.Y. Springer.

[8] W. Häerdle and P. Vieu "Kernel regression smoothing of time series", *J. Time Ser. Anal.* 13, No.3, 209-232 (1992).

[9] Ibragimov,I.A. and Khasminskii,R.Z. (1981). *Statistical Estimation: Asymptotic Theory* Springer, New York.

[10] Katkovnik, V. Ja. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach.* Nauka, Moscow (in Russian).

[11] Khasminskii, R.Z. (1980). *Stochastic stability of differential equations.* Sijthoff & Noordhoff.

[12] Kutoyants, Yu.A. (1984a). On nonparametric estimation of trend coefficients in a diffusion process. Collection: Statistics and control of stochastic processes, Moscow, 230–250.

[13] Kutoyants, Yu.A. (1984b). Parameter estimation for stochastic processes. Translated from the Russian and edited by B. L. S. Prakasa Rao. R & E Research and Exposition in Mathematics, 6. Heldermann Verlag, Berlin.

[14] Lepski, O. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, no. 3, 459–470.

[15] Lepski, O. and Levit, B. (1997). Efficient adaptive estimation of infinitely differentiable function. *Math. Methods of Statistics*, submitted.

[16] Lepski, O., Mammen, E. and Spokoiny, V. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Statistics*, **25**, no.3, 929–947.

[17] Lepski, O. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, **25**, no.6,

[18] Liptser, R. and Shiryaev, A. (1989).*Theory of Martingales.* Kluwer Acad. Publ. 1989.

[19] Liptser, R. and Spokoiny, V. (1997). Moderate Deviations for integral functionals of diffusion process. Unpublished manuscript.

[20] Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Stat.*, **24**, no.6. 2477–2498.

[21] Tsybakov, A. (1986). Robust reconstruction of functions by the local approximation. *Prob. Inf. Transm.*, **22**, 133-146.

[22] Veretennikov, A. Yu. (1991) On the averaging principle for systems of stochastic differential equations. *Math. USSR Sborn.*, **69**, No. 1, 271-284.

[23] Veretennikov, A. Yu. (1992) On large deviations for ergodic empirical measures, *Topics in Nonparametric Estimation. Advances in Soviet Mathematics, AMS* **12**, 125-133.