DRIFT ESTIMATION FOR NONPARAMETRIC DIFFUSION MODEL: NONASYMPTOTIC APPROACH

SPOKOINY, V.

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

ABSTRACT. We consider a nonparametric diffusion process whose drift and diffusion coefficients are nonparametric functions of the state variable. The goal is to estimate the unknown drift coefficient. We apply a locally linear smoother with a data-driven bandwidth choice. The procedure is fully adaptive and nearly optimal up to a *log log* factor. The results about the quality of estimation are nonasymptotic and do not require any ergodic or mixing properties of the observed process.

1. Introduction

In this paper, we propose a procedure for adaptive estimation of the drift coefficient of a diffusion system described by the Itô equations

$$\mathrm{d}X_t \quad = \quad f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}w_t, \qquad X_0 = x_0, \qquad 0 \le t \le T. \tag{1.1}$$

Here w_t is a standard Wiener process and T is the *observation time.* The functions f, g, entering in (1.1), which are usually referred to as *drift* and *diffusion* coefficients, are unknown. The goal is to recover the unknown drift function f from the observations X_t, $0 \le t \le T$. We do not discuss here the problem of estimating the diffusion coefficient g since in the case of continuous observations, the required information about this function g can be exactly recovered from the data, Section 3.5 below. We also restrict ourselves to the problem of pointwise estimation, that is, given a point x, we estimate the value $f(x)$. The reader is referred to Lepski, Mammen and Spokoiny (1997) for a discussion of the relation between pointwise and global estimation. Note that the problem of the pointwise estimation of the drift function f is closely connected to the problem of forecasting the process X. Indeed, if we observe the process (X_t) until the time-point T, and if we

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

Key words and phrases. drift and diffusion coefficients, nonparametric estimation, bandwidth selection.

are interested in a behavior of the process in the nearest future after T, then we have to estimate $f(x)$ for $x = X_t$.

Statistical inference for stochastic processes and time series has attracted a lot of attention last years, especially in view of applications to finance mathematics. The estimation theory for diffusion type processes is well developed under the parametric modeling when underlying functions (drift and diffusion) are specified up to a value of a finite dimensional parameter (cf. Kutoyants, 1984b). In contrast, nonparametric estimation is not studied in details. The known results concern only with statistical inference for ergodic diffusion models with a small noise or for a large observation time T. Kutoyants (1984a) evaluated the minimax rate of estimation of the drift coefficient using a kernel type estimator. Genon-Catalot, Laredo and Picard (1992) applied wavelets. Locally polynomial estimators are described in Fan and Gijbels (1996). Milstein and Nussbaum (1994) established the Le Cam equivalence between the diffusion model and the "white noise model". Some pertinent results for autoregressive models in discrete time can be found in Doukhan and Ghindes (1980), Collomb and Doukhan (1983), Doukhan and Tsybakov (1993), Delyon and Juditsky (1997), Neumann (1998). A series of papers discusses simultaneous estimation of the drift and diffusion functions, among them Hall and Carroll (1989), Härdle and Tsybakov (1997), Ruppert et al (1997), Fan and Yao (1988).

It is worth mentioning that the stationarity assumption could be very restrictive for practical applications. Typically this assumption is fulfilled only in some local sense, that is, observed processes are only locally stationary. In other words, for every time point t, there is a time interval containing t and such the observed process is stationary or near stationary within this intervals, see e.g. Dahlhaus (1997) for more discussion. Statistical inference under local stationary assumption requires to study some nonasymptotic properties of statistical procedures. The reader is referred to the forthcoming paper by Härdle and Spokoiny (1999) for an example of parameter estimation for ARCH- and stochastic volatility models under local stationarity.

The present paper offers another approach to relax the stationarity assumption, so that neither ergodic property of the slow component nor large observation time T is assumed. This makes the problem much more complicated. We propose a locally linear estimator of $f(x)$ with a data-driven bandwidth choice and show that this method provides a nearly optimal accuracy of estimation up to a $\log\log T$ factor. The idea of proposed bandwidth selector goes back to Lepski (1990). Lepski, Mammen and Spokoiny (1997) presented a slightly modified version of the original Lepski's procedure and showed its optimality in the asymptotic minimax sense (over a wide range of Besov classes) and for the global L_p-risk in the "white noise model". Lepski and Spokoiny (1997) constructed an asymptotically sharp optimal pointwise adaptive procedure, again for the "white noise model". In this paper the procedure is adapted to locally linear smoothing in diffusion type model (1.1). The results compare the quality of the adaptive procedure to that of for an "ideal" estimate defined by the optimal choice of the smoothing parameter (bandwidth), see Section 4 for more discussion. In particular, it is shown that the accuracy of the adaptive procedure is worse than the "ideal" one by a factor $\log\log T$ which can be viewed as payment for the adaptive property.

The paper is organized as follows. The next section contains the description of a locally linear estimator. Its properties are discussed in Section 3. The data-driven bandwidth choice is presented in Section 4. All proofs are gathered in Sections 5.

2. A locally linear estimator

For fixed x, to estimate the value $f(x)$ we apply the locally linear smoother (cf. Katkovnik (1985), Tsybakov (1986), Fan and Gijbels (1996)).

We begin with some heuristic explanations of the method. Imagine for a moment that the observed process X_t, $0 \le t \le T$ satisfies the Itô equation with respect to Wiener process w_t:

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}w_t \tag{2.1}$$

with a linear function f of the form $f(u) = \theta_0 + \theta_1(u-x)/h$, depending on two parameters θ_0, θ_1, where x and $h > 0$ are fixed. The values θ_0 and θ_1 can be estimated by the least squares method:

$$(\widetilde{\theta}_0, \widetilde{\theta}_1) = \operatorname*{argmax}_{\theta_0,\theta_1} \left\{ \int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h}\right) \mathrm{d}X_t - \frac{1}{2}\int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h}\right)^2 \mathrm{d}t \right\}.$$

This quadratic optimization problem can be explicitly solved: with

$$\mu_k = \int_0^T \left(\frac{X_t - x}{h}\right)^k \mathrm{d}t, \qquad k = 0, 1, 2,$$

one has

$$\begin{aligned} \widetilde{\theta}_0 &= \frac{\mu_2 \int\limits_0^T \mathrm{d}X_t - \mu_1 \int\limits_0^T \frac{X_t - x}{h}\,\mathrm{d}X_t}{\mu_0\mu_2 - \mu_1^2}, \\ \widetilde{\theta}_1 &= \frac{-\mu_1 \int\limits_0^T \mathrm{d}X_t + \mu_0 \int\limits_0^T \frac{X_t - x}{h}\,\mathrm{d}X_t}{\mu_0\mu_2 - \mu_1^2}. \end{aligned}$$

Since clearly $f(x) = \theta_0$, the value $\widetilde{\theta}_0$ can be taken for estimating $f(x)$.

The locally linear smoother is defined in a similar way. The only difference is that the function f is not assumed to be linear but it is approximated by a linear function $\theta_0 + \theta_1(u - x)/h$ in a small neighborhood $[x - h, x + h]$ of the point x. Then the coefficients θ_0, θ_1 of this function can be estimated from the observations of X_t falling into the interval $[x-h, x+h]$. For formal description, let us introduce a *kernel* function $K(u)$ which is assumed to be smooth, non-negative, bounded by 1, and vanishing outside of $[-1,1]$. Then the locally linear estimate with the kernel K and a *bandwidth* h is defined as:

$$\widetilde{f}_h(x) = \frac{\mu_{2,h} \int\limits_0^T K\left(\frac{X_t - x}{h}\right) \mathrm{d}X_t - \mu_{1,h} \int\limits_0^T \frac{X_t - x}{h} K\left(\frac{X_t - x}{h}\right) \mathrm{d}X_t}{\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2}, \tag{2.2}$$

where

$$\mu_{k,h} = \int_0^T \left(\frac{X_t - x}{h}\right)^k K\left(\frac{X_t - x}{h}\right) \mathrm{d}t, \qquad k = 0,1,2. \tag{2.3}$$

The quality of estimate (2.2) essentially depends on the bandwidth h. Some useful properties of $\widetilde{f}_h(x)$ for the fixed h are described in Section 3. An adaptive (data-driven) choice of the bandwidth h is discussed in Section 4.

3. Some properties of the locally linear estimate

In this section we study some properties of the locally linear estimate $\widetilde{f}_h(x)$ from (2.2). We first formulate the required conditions on the coefficients f, g from (1.1). Then we present the result and discuss some its corollaries.

3.1. Conditions

In the sequel we suppose that the functions f, g from (1.1) obey the following conditions:

(A_s) Functions $f(u)$ and $g(u)$ are Lipschitz continuous in u and $f(u)$ is two times continuously differentiable in u. For some positive constants $g_{\min} \le g_{\max}$

$$g_{\min} \le |g(u)| \le g_{\max} \qquad \forall u.$$

It is worth mentioning that we do not impose any conditions which ensure ergodic or mixing properties of the process X. Our approach is essentially non-asymptotic and there is no difference between ergodic and non-ergodic cases.

3.2. Accuracy of the locally linear estimate

To state the result, we introduce some additional notations. With $\mu_{k,h}$ defined in (2.3), set

$$D_h \quad = \quad \mu_{0,h}\mu_{2,h} - \mu_{1,h}^2, \tag{3.1}$$

and

$$
\begin{aligned}
\sigma_h^2(x) &= \frac{1}{D_h^2}\int_0^T \left(\mu_{2,h}-\mu_{1,h}\frac{X_t-x}{h}\right)^2 K^2\left(\frac{X_t-x}{h}\right) g^2(X_t)\,\mathrm{d}t \\
&= v_{2,h}^2 V_{0,h} - 2v_{1,h}v_{2,h}V_{1,h} + v_{1,h}^2 V_{2,h}
\end{aligned} \tag{3.2}
$$

where

$$
\begin{aligned}
v_{k,h} &= \frac{\mu_{k,h}}{D_h} = \frac{\mu_{k,h}}{\mu_{0,h}\mu_{2,h}-\mu_{1,h}^2}, \qquad k=1,2, \\
V_{k,h} &= \int_0^T \left(\frac{X_t-x}{h}\right)^k K^2\left(\frac{X_t-x}{h}\right) g^2(X_t)\,\mathrm{d}t.
\end{aligned}
$$

Although the expressions for $V_{k,h}$, $k=0,1,2$, use the unknown diffusion coefficient $g^2(X_t)$, these values can be computed on the base of our observations $(X_t,\,0\le t\le T)$ only, see Section 3.5.

The value $\sigma_h^2(x)$ is called the *conditional variance* of the estimate $\widetilde{f}_h(x)$. This terminology is used by analogy with the regression case, where X_t is a deterministic design process and $\sigma_h^2(x)$ is really the variance of the least squares estimate $\widetilde{f}_h(x)$. Note that for the regression setup, some design regularity is required to ensure that $\sigma_h^2(x)$ is not too large.

In our case, X_t is the observed process which at the same time can be viewed as the design process. We therefore impose some conditions on the trajectories of the process X_t which are similar to that of used to describe the design regularity in the regression setting. Our results are also similar to that of can be obtained in the regression context, cf. Lepski, Mammen and Spokoiny (1997) or Lepski and Spokoiny (1997). In particular, we show that under the conditions imposed, the conditional variance $\sigma_h^2(x)$ helps to control the stochastic component of the estimate $\widetilde{f}_h(x)$.

For some $\rho\ge 0$, $r>0$, $b>0$ and $B\ge 1$ we introduce the set

$$
\mathcal{A}_h = \left\{ \begin{array}{ll}
\frac{b}{Th}\le v_{2,h}\le \frac{bB}{Th}, & \frac{b}{Th}\le \sigma_h^2(x)\le\frac{bB}{Th}, \\
\mu_{0,h}\le r\mu_{2,h}, & V_{0,h}\le rV_{2,h} \\
\mu_{1,h}^2\le\rho\,\mu_{0,h}\mu_{2,h}, & V_{1,h}^2\le\rho\,V_{0,h}V_{2,h}
\end{array}\right\}.
$$

Since X_t is the random process, the set $\mathcal{A}_h$ is random as well. In the sequel we study the properties of $\widetilde{f}_h(x)$ restricted to the set $\mathcal{A}_h$, see Section 3.3 for further discussion.

The quality of the approximation of $f(u)$ by a linear in u function in the neighborhood $u\in[x-h,x+h]$ is characterized by the following quantity

$$
\Delta_h(x) = \sup_{|u-x|\le h}|f(u)-f(x)-(u-x)f'(x)| \tag{3.3}
$$

where f' denotes the derivative of f. In the next theorem we describe some useful properties of the estimate (2.2).

Theorem 3.1. *Let* (A_s) *be fulfilled, and* $Th \geq 1$ *. Then for every* $\lambda \geq \sqrt{2}$

$$\boldsymbol{P}\left(\left|\widetilde{f}_h(x) - f(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x),\, \mathcal{A}_h\right) \\ \leq 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda\, e^{-\frac{\lambda^2}{2}}, \tag{3.4}$$

where $c = (1-\rho)^{-1/2}$.

Informally the result of the theorem means that the losses $|\widetilde{f}_h(x) - f(x)|$ of the estimate $\widetilde{f}_h(x)$, being restricted to $\mathcal{A}_h$, are bounded by the sum of two terms: $c\Delta_h(x)$ and $\lambda\sigma_h(x)$. The first one mimics the accuracy of approximating the function $f(u)$ by a linear in u function in the small vicinity $[x-h, x+h]$ of x . The second term is in proportion to the "stochastic standard deviation" $\sigma_h(x)$.

3.3. Some remarks related to the random set $\mathcal{A}_h$

The result of Theorem 3.1 describes the accuracy of the estimate $\widetilde{f}_h(x)$ on the random set $\mathcal{A}_h$ only. Here we briefly discuss some related questions.

3.3.1. Reason for restricting to $\mathcal{A}_h$

It was mentioned previously that restricting to $\mathcal{A}_h$ allows to eliminate irregular cases when, for instance, the trajectory $X_{[0,T]}$ does not pass through the interval $[x-h, x+h]$ and $\mu_{0,h} = \mu_{1,h} = \mu_{2,h} = D_h = 0$. Note that for typical applications to forecasting, we have to estimate $f(x)$ with $x = X_t$, and the trajectory $X_{[0,T]}$ obviously passes through x .

3.3.2. Verifying the condition $X_{[0,T]} \in \mathcal{A}_h$

Clearly the event $\mathcal{A}_h$ is completely determined by the known values $\mu_{k,h}$ and $V_{k,h}$, $k = 0, 1, 2$. It is therefore always possible to check whether the observed trajectory $X_{[0,T]}$ belongs to $\mathcal{A}_h$ or not. If the trajectory $X_{[0,T]}$ does not belong to $\mathcal{A}_h$, we are not able to guarantee a reasonable quality for the estimate $\widetilde{f}_h(x)$.

3.3.3. The conditions entering into the definition of $\mathcal{A}_h$

The conditions $0 \leq K(u) \leq 1$ and $K(u) = 0$ for $|u| \geq 1$ imply $\mu_{2,h} \leq \mu_{0,h}$ and $V_{2,h} \leq V_{0,h}$. Further, by the Cauchy-Schwarz inequality, it holds $\mu_{1,h}^2 \leq \mu_{0,h}\mu_{2,h}$ and $V_{1,h}^2 \leq V_{0,h}V_{2,h}$. The conditions $\mu_{0,h} \leq r\mu_{2,h}$, $V_{0,h} \leq rV_{2,h}$, $\mu_{1,h}^2 \leq \rho\mu_{0,h}\mu_{2,h}$ and $V_{1,h}^2 \leq \rho V_{0,h}V_{2,h}$ with $\rho < 1$ and $r \geq 1$ ensure that the local linear estimate is well defined. Note that these conditions are not completely independent. In particular, if

$g(x)$ is a constant function and if $K(u) = \mathbf{1}(|u| \le 1)$, then $\mu_{k,h} = V_{k,h}$ for $k = 0, 1, 2$ and $\sigma_h^2(x) = v_{2,h} = \mu_{2,h}/(\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2)$.

3.3.4. The choice of the constants ρ, b, B, r

The choice of constants ρ, b, B, r, entering in the definition of the set $\mathcal{A}_h$, is optional and they even may depend on T. Note that the upper bound (3.4) from Theorem 3.1 does not depends on b and it depends on B (which determines the range of different values for the conditional variance $\sigma_h^2(x)$) only via the log-factor $\log(4B^3)$.

3.3.5. Unconditional result under ergodicity

If the coefficients f and g obey some additional conditions which ensure ergodicity of the process X_t, see e.g. Veretennikov (1991), then, at least with growing T the normalized integrals $(Th)^{-1}\mu_{k,h}$ and $(Th)^{-1}V_{k,h}$ ($k = 0, 1, 2$) converge to some fixed values which depend only on the stationary distribution of the process X_t. Moreover, one can usually select fixed constants b, B and ρ, r in such a way that $1 - \boldsymbol{P}(\mathcal{A}_h)$ converges to zero exponentially fast as $T \to \infty$. Since obviously

$$\begin{aligned} &\boldsymbol{P}\left(\left|\widetilde{f}_h(x) - f(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x)\right) \\ &\qquad \le \boldsymbol{P}\left(\left|\widetilde{f}_h(x) - f(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x),\, \mathcal{A}_h\right) + \boldsymbol{P}(\mathcal{A}_h) \end{aligned}$$

we obtain in this situation an unconditional asymptotic bound for the risk of the estimate $\widetilde{f}_h(x)$.

3.4. Quality of estimation under smoothness assumptions

Due to the assumptions (A_s) from Section 3, the function f is twice continuously differentiable. Assume also that for every u from a small vicinity of x, the second derivative f'' is bounded by some fixed constant L:

$$\left|f''(u)\right| \le L. \tag{3.5}$$

Then the value $\Delta_h(x)$ defined in (3.3), is bounded above by $Lh^2/2$. On the other hand, on the set $\mathcal{A}_h$ the stochastic variance $\sigma_h^2(x)$ is of order $(Th)^{-1}$. Therefore, following to the standard approach in nonparametric estimation, the bandwidth h can be chosen by balancing the accuracy of approximation and the stochastic error:

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

This leads to the choice $h \asymp (T\,L^2)^{-1/5}$ and hence to the 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 this approach hardly applies in practice, since the constant L in (3.5) is typically unknown. An adaptive (data-driven) choice of the bandwidth is discussed in the next section.

3.5. Computation of $\sigma_h^2(x)$

Recall that with fixed h, the value $\sigma_h^2(x)$ is defined by the formula

$$\begin{aligned}\sigma_h^2(x) &= \frac{1}{D_h^2}\int_0^T K^2\left(\frac{X_t-x}{h}\right)\left(\mu_{2,h}-\mu_{1,h}\frac{X_t-x}{h}\right)^2 g^2(X_t)\,\mathrm{d}t\\ &= v_{2,h}^2 V_{0,h} - 2v_{1,h}v_{2,h}V_{1,h} + v_{1,h}^2 V_{2,h}\end{aligned}$$

with

$$\begin{aligned}\mu_{k,h} &= \int_0^T \left(\frac{X_t-x}{h}\right)^k K\left(\frac{X_t-x}{h}\right)\mathrm{d}t,\\ D_h &= \mu_{0,h}\mu_{2,h}-\mu_{1,h}^2,\\ v_{k,h} &= \frac{\mu_{k,h}}{D_h} = \frac{\mu_{k,h}}{\mu_{0,h}\mu_{2,h}-\mu_{1,h}^2},\\ V_{k,h} &= \int_0^T \left(\frac{X_t-x}{h}\right)^k K^2\left(\frac{X_t-x}{h}\right) g^2(X_t)\,\mathrm{d}t, \qquad k=0,1,2.\end{aligned}$$

The formula for $\sigma_h^2(x)$ includes the unknown diffusion coefficient $g^2(X_t)$. We now show that despite of this fact, the value $\sigma_h^2(x)$ can be computed via the observations $X_{[0,T]}$ only.

Let us introduce two random processes

$$Z_t' = \int_0^t K\left(\frac{X_s-x}{h}\right)\mathrm{d}X_s \quad \text{and} \quad Z_t'' = \int_0^t K\left(\frac{X_s-x}{h}\right)\frac{X_s-x}{h}\,\mathrm{d}X_s$$

which are completely determined on the time interval $[0,T]$ by $X_{[0,T]}$. Applying the Itô formula we get

$$\begin{aligned}(Z_T')^2 &= 2\int_0^T Z_t'\,\mathrm{d}Z_t' + V_{0,h}\\ (Z_T'')^2 &= 2\int_0^T Z_t''\,\mathrm{d}Z_t'' + V_{2,h}\\ Z_T'Z_T'' &= \int_0^T Z_t'\,\mathrm{d}Z_t'' + \int_0^T Z_t''\,\mathrm{d}Z_t' + V_{1,h}.\end{aligned}$$

Hence $V_{0,h} = (Z_T')^2 - 2\int_0^T Z_t'\,\mathrm{d}Z_t'$, so that $V_{0,h}$ is completely determined by $X_{[0,T]}$. Similar arguments apply for $V_{1,h}$ and $V_{2,h}$ and hence for $\sigma_h^2(x)$ as required.

4. Data-driven bandwidth selection

In this section we consider the problem of bandwidth selection for the locally linear estimator described in Section 2. It is assumed here that the method of estimation, that is, the locally linear smoother with the kernel K, is fixed and only the bandwidth h has to be chosen. The adaptive procedure originates from Lepski (1990), see also Lepski, Mammen and Spokoiny (1997) and Lepski and Spokoiny (1997).

4.1. An "ideal" bandwidth

First we introduce the notion of an "ideal" bandwidth. Let a set $\mathcal{H}$, of all admissible bandwidths 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. As in Section 3, we restrict ourselves only to those h from $\mathcal{H}$ for which the observed trajectory $X_{[0,T]}$ belongs to $\mathcal{A}_h$. Our goal is to select h from $\mathcal{H}$ providing the minimal in some sense error of estimation for the corresponding estimate $\widetilde{f}_h(x)$.

We begin with some heuristic explanations. Recall first, that the values $\sigma_h^2(x)$ can be exactly computed on the base of observations $X_{[0,T]}$, see Subsection 3.5. Note also that $\sigma_h^2(x)$ typically decreases in h. Indeed, an increase of h makes the estimation window $[x-h, x+h]$ larger and hence more observations can be used for estimating the underlying function f at the point x. This results in a smaller variance of the estimate. To simplify the exposition, we suppose that $\sigma_h^2(x)$ strongly decreases in $h \in \mathcal{H}$. (If this assumption is not fulfilled for the original set $\mathcal{H}$, i.e. if there is $h' < h \in \mathcal{H}$ with the property $\sigma_h^2(x) \ge \sigma_{h'}^2(x)$, then we simply exclude h from $\mathcal{H}$.)

The behavior of the bias term $\Delta_h(x)$ is just opposite. Namely, for a regular function f, the value $\Delta_h(x)$ is small when h is small, and it typically increases in h. Therefore, the minimization of the sum of the form $c\Delta_h(x) + \lambda\sigma_h(x)$ with some constants c, λ leads to the balance relation $\Delta_h(x) \asymp \sigma_h(x)$ and we define a "good" bandwidth h_{id} as the largest h from $\mathcal{H}$ such that $c\Delta_h(x)$ is still not larger than $D\sigma_h(x)$ with some prescribed constant D:

$$h_{\text{id}} = \max\{h \in \mathcal{H} : c\Delta_h(x) \le D\sigma_h(x)\}. \tag{4.1}$$

Since $\Delta_h(x)$ is unknown, the bandwidth h_{id} is unknown as well. In the sequel, following to Donoho and Johnstone (1994), h_{id} is referred to as an "ideal" bandwidth or "oracle". Due to Theorem 3.1, the losses of the "ideal" estimate $\widetilde{f}_{h_{\text{id}}}$ are bounded (with probability closed to one) by $(D+\lambda)\sigma_{h_{\text{id}}}(x)$ provided that λ is sufficiently large.

4.2. An adaptive bandwidth choice

Now we present our adaptive procedure and show that the corresponding accuracy of the estimation is essentially the same as if the "ideal" bandwidth applies. The procedure involves two positive parameters λ_1 and D. The last one is already mentioned in the definition of the "ideal" bandwidth. We discuss the choice of λ_1 and D at the end of this section.

The data-driven bandwidth $\widehat{h}$ is defined by the following rule:

$$\widehat{h} = \max\Bigl\{h \in \mathcal{H} : \bigl|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\bigr| \le \lambda_1\Bigl(\sigma_h(x) + \sigma_\eta(x)\Bigr) + 2D\sigma_h(x), \qquad \forall \eta \in \mathcal{H}, \eta < h\Bigr\}. \tag{4.2}$$

In words, the rule prescribes to take the largest value $h \in \mathcal{H}$ for which the corresponding estimate $\widetilde{f}_h(x)$ does not differ essentially from every estimate $\widetilde{f}_\eta(x)$ with a smaller bandwidth value $\eta \in \mathcal{H}$. The arguments for this choice are quite simple: if both η and h are not larger than h_{id}, then the "bias" terms $\Delta_\eta(x)$ and $\Delta_h(x)$ in the difference $\bigl|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\bigr|$ are bounded by $2D\sigma_{h_{\mathrm{id}}}(x) \le 2D\sigma_h(x)$ and therefore, the probability of the event

$$\Bigl\{\bigl|\widetilde{f}_h(x) - \widetilde{f}_\eta(x)\bigr| > \lambda_1\Bigl(\sigma_h(x) + \sigma_\eta(x)\Bigr) + 2D\sigma_h(x)\Bigr\}$$

is small provided that λ_1 is large enough (see Theorem 3.1). Hence, if we meet the opposite inequality for some $\eta < h$, this means that the bias $\Delta_h(x)$ is already too large and the bandwidth h is not a good one.

Finally, to define our adaptive estimate, we plug the data-driven bandwidth $\widehat{h}$ in the estimate $\widetilde{f}_h(x)$:

$$\widehat{f}(x) \equiv \widetilde{f}_{\widehat{h}}(x). \tag{4.3}$$

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

$$\mathcal{A}^* = \bigcap_{h \in \mathcal{H}} \mathcal{A}_h.$$

Theorem 4.1. *Let h_{id} be defined in (4.1) with $\lambda_1 \ge \sqrt{2}$. Then the estimate $\widehat{f}(x)$ fulfills the following property: for any λ with $\sqrt{2} \le \lambda \le \lambda_1$*

$$\begin{aligned} &\boldsymbol{P}\left(\left|\widehat{f}(x) - f(x)\right| > (\lambda + \lambda^*)\sigma_{h_{\mathrm{id}}}(x), \mathcal{A}^*\right) \\ \le\ & 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda_1^2\right)\lambda_1\left\{(\#\mathcal{H})^2 e^{-\frac{\lambda_1^2}{2}} + e^{-\frac{\lambda^2}{2}}\right\}, \end{aligned} \tag{4.4}$$

where

$$\lambda^* = 2\lambda_1 + 3D. \tag{4.5}$$

4.3. The choice of parameters λ_1, D

The choice of parameters λ_1, D, entering in (4.2), plays the important role. The bound in (4.4) shows that the probability for $\left|\widehat{f}(x) - f(x)\right|$ of being large is small, provided that the value $(\#\mathcal{H})^2\lambda_1^2 e^{-\lambda_1^2/2}$ is sufficiently small. This leads to the choice

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

so that

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

If $\mathcal{H}$ is taken in the form of the 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

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

There is much more degree of freedom in the choice of D. This parameter controls the balance between the accuracy of approximating the function f by a linear one and the stochastic error (see the definition (4.1) of the "ideal" bandwidth h_{id}). The results from Lepski and Spokoiny (1997) lead to the choice $D = \text{Const}\,\lambda_1$ (see also the next section). At the same time, Lepski and Levit (1997) argued that for a smooth function f, the relevant choice is $D = 0$. Simulation results show a reasonable performance of the presented procedure with $\lambda_1 \approx 3$ and $D = 0$.

4.4. The rate of adaptive estimation

We now compare the accuracy of the adaptive procedure (4.2) with the "optimal" one designed for the case of known smoothness properties of the underlying function f (see Section 3.4).

Assume $|f''(u)| \le L$, see (3.5). Then $\Delta_h(x) \le Lh^2/2$ and the constraints $c\Delta_h(x) \le D\sigma_h(x)$ and $b(hT)^{-1} \le \sigma_h^2(x) \le bB(hT)^{-1}$ yield for h_{id} from (4.1)

$$h_{\text{id}} \ge C_1\left(\frac{D^2}{TL^2}\right)^{1/5}$$

with $C_1 = (2bc^{-2})^{1/5}$, so that

$$\sigma_{h_{\text{id}}}(x) \le \left(\frac{bB}{Th_{\text{id}}}\right)^{1/2} \le C_2L^{1/5}(T^2D)^{-1/5}$$

with $C_2 = (bB/C_1)^{1/2}$. Hence, the above-mentioned choice $\lambda_1 \approx 2\sqrt{\log\log T}$ and $D = C\lambda_1$, leads due to Theorem 4.1 to the following accuracy of the adaptive estimation

$$(\lambda + 2\lambda_1 + 3D)\sigma_{h_{\text{id}}}(x) \le C_3L^{1/5}\left(\frac{2\log\log T}{T}\right)^{2/5}$$

with $C_3 = 3C_2(1+C)C^{-1/5}$. At the same time, the "ideal" choice of the bandwidth leads to the rate $L^{1/5}T^{-2/5}$, see Section 3.4. Thus, the accuracy of adaptive estimation is worse than the "ideal" one within a $loglogT$-factor only.

The origin of the $\log\log T$-factor in the rate of adaptive estimation can be easily explained. The total number $\#\mathcal{H}$ of considered estimates is logarithmic in the observation time T and the adaptive choice of the bandwidth leads to a worse accuracy by factor $\log(\#\mathcal{H})$ at some power.

The notion of "payment for adaptation" is now well understood in nonparametric estimation: if we have too many estimates to select between, we have to "pay" for the adaptive choice some additional factor in the risk of estimation. In particular, it is shown in Lepski (1990) and Brown and Low (1996) (see also Lepski and Spokoiny (1997)) that for the problem of pointwise adaptive estimation, the optimal adaptive rate has to be worse than the optimal one by a log-factor.

In our results a $\log\log$-factor appears. This fact is not in the contradiction with earlier issues, since the above-mentioned results correspond to the case of the power loss function $\ell(x) = |x|^p,\ p > 0$, while we consider the bounded loss function. It can be also shown that the rate achieved by our estimate is optimal for pointwise adaptive estimation with a bounded loss function (see Spokoiny (1997) for similar results in the adaptive testing problem).

5. Proofs

In this section we prove Theorems 3.1 and 4.1.

5.1. Decomposition of $\widetilde{f}_h(x)$

We use two obvious identities characterizing the local linear smoother: for $v_{1,h} = \frac{\mu_{1,h}}{D_h}$ and $v_{2,h} = \frac{\mu_{2,h}}{D_h}$

$$\begin{aligned} \int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right) \mathrm{d}s &= 1 \\ \int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h}\frac{X_s - x}{h} - v_{1,h}\frac{(X_s - x)^2}{h^2}\right) \mathrm{d}s &= 0 \end{aligned}$$

and hence

$$\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right) f(x)\,\mathrm{d}s = f(x) \tag{5.1}$$

$$\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h}\frac{X_s - x}{h} - v_{1,h}\frac{(X_s - x)^2}{h^2}\right) f'(x)\,\mathrm{d}s = 0. \tag{5.2}$$

Due to (2.2) and (1.1), the estimate $\widetilde{f}_h(x)$ can be represented as follows:

$$\begin{aligned}
\widetilde{f}_h(x) &= v_{2,h}\int_0^T K\left(\frac{X_s-x}{h}\right)\mathrm{d}X_s - v_{1,h}\int_0^T K\left(\frac{X_s-x}{h}\right)\frac{X_s-x}{h}\,\mathrm{d}X_s \\
&= \int_0^T K\left(\frac{X_s-x}{h}\right)\left(v_{2,h}-v_{1,h}\frac{X_s-x}{h}\right) f(X_s)\,\mathrm{d}s \\
&\quad + v_{2,h}\int_0^T K\left(\frac{X_s-x}{h}\right) g(X_s)\,\mathrm{d}w_s \\
&\quad - v_{1,h}\int_0^T K\left(\frac{X_s-x}{h}\right)\frac{X_s-x}{h}\, g(X_s)\,\mathrm{d}w_s.
\end{aligned}$$

Now (5.1) and (5.2) imply the following decomposition

$$\widetilde{f}_h(x) = f(x) + \xi_h + r_h \tag{5.3}$$

where, with $\delta(X_s,x) = f(X_s) - f(x) - \frac{X_s-x}{h}\, f'(x)$,

$$\begin{aligned}
r_h &= \int_0^T K\left(\frac{X_s-x}{h}\right)\left(v_{2,h}-v_{1,h}\frac{X_s-x}{h}\right)\delta(X_s,x)\,\mathrm{d}s, \\
\xi_h &= v_{2,h}\int_0^T K\left(\frac{X_s-x}{h}\right) g(X_s)\,\mathrm{d}w_s \\
&\quad - v_{1,h}\int_0^T K\left(\frac{X_s-x}{h}\right)\frac{X_s-x}{h}\, g(X_s)\,\mathrm{d}w_s.
\end{aligned}$$

Below we evaluate separately each term in this decomposition.

5.2. An upper bound for $|r_h|$

Since $K\left(\frac{u-x}{h}\right)$ vanishes for any $u \notin [x-h, x+h]$ and $|\delta(X_s,x)| \le \Delta_h(x)$ for $|X_s - x| \le h$, we get

$$\begin{aligned}
|r_h| &\le \int_0^T K\left(\frac{X_s-x}{h}\right)\left(v_{2,h}-v_{1,h}\frac{X_s-x}{h}\right)|\delta(X_s,x)|\,\mathrm{d}s \\
&\le \Delta_h(x)\int_0^T K\left(\frac{X_s-x}{h}\right)\left|v_{2,h}-v_{1,h}\frac{X_s-x}{h}\right|\,\mathrm{d}s.
\end{aligned} \tag{5.4}$$

The properties $|K(u)| \le 1$ and $K(u) = 0,\ |u| \ge 1$ imply the inequality $\mu_{2,h} \le \mu_{0,h}$. In addition we know that it holds on $\mathcal{A}_h$

$$\mu_{1,h}^2 \le \rho\,\mu_{0,h}\mu_{2,h}. \tag{5.5}$$

We now show that

$$|r_h| \le (1-\rho)^{-1/2}\Delta_h(x) \qquad \text{on} \quad \mathcal{A}_h. \tag{5.6}$$

The Cauchy-Schwarz inequality applied to (5.4) gives

$$|r_h| \le \Delta_h(x)\left\{\int_0^T K\left(\frac{X_s-x}{h}\right)\mathrm{d}s\int_0^T K\left(\frac{X_s-x}{h}\right)\left(v_{2,h}-v_{1,h}\frac{X_s-x}{h}\right)^2\mathrm{d}s\right\}^{1/2}.$$

Next,

$$\int_0^T K\left(\frac{X_s - x}{h}\right) \mathrm{d}s = \mu_{0,h},$$

and using $v_{k,h} = \mu_{k,h}/D_h$, with $D_h = \mu_{2,h}\mu_{0,h} - \mu_{1,h}^2$, $k = 0, 1, 2$, we get

$$\begin{aligned}
&\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right)^2 \mathrm{d}s \\
&\quad = \frac{1}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\left(\mu_{2,h} - \mu_{1,h}\frac{X_s - x}{h}\right)^2 \mathrm{d}s \\
&\quad = \frac{\mu_{2,h}^2}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right) \mathrm{d}s + \frac{\mu_{1,h}^2}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{(X_s - x)^2}{h^2} \mathrm{d}s \\
&\qquad - \frac{2\mu_{1,h}\mu_{2,h}}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h} \mathrm{d}s \\
&\quad = \frac{\mu_{2,h}^2\mu_{0,h} - \mu_{2,h}\mu_{1,h}^2}{D_h^2} \\
&\quad = \mu_{2,h}/D_h.
\end{aligned}$$

Hence, in view of (5.5),

$$|r_h| \quad \le \quad \Delta_h(x)\left(\frac{\mu_{0,h}\,\mu_{2,h}}{D_h}\right)^{1/2} = \Delta_h(x)\left(\frac{\mu_{0,h}\,\mu_{2,h}}{\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2}\right)^{1/2} \le \Delta_h(x)\left(\frac{1}{1-\rho}\right)^{1/2}$$

as required.

5.3. An upper bound for $\boldsymbol{\xi_h}$

We study here some properties of the “stochastic term”

$$\begin{aligned}
\xi_h \quad = \quad & v_{2,h}\int_0^T K\left(\frac{X_s - x}{h}\right) g(X_s)\,\mathrm{d}w_s \\
& - v_{1,h}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h} g(X_s)\,\mathrm{d}w_s.
\end{aligned}$$

Namely, we intend to show that the probability of the event $\{\xi_h > \lambda\sigma_h(x)\}$ with $\sigma_h(x)$ from (3.2) is small provided that λ is large enough. Set for $t \le T$

$$\begin{aligned}
M_{0,t} \quad &= \quad \int_0^t K\left(\frac{X_s - x}{h}\right) g(X_s)\ \mathrm{d}w_s, \\
M_{1,t} \quad &= \quad \int_0^t K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h} g(X_s)\ \mathrm{d}w_s.
\end{aligned}$$

The Itô integrals $M_{0,t}$ and $M_{1,t}$ are continuous local martingales with the predictable quadratic variations (see e.g. Liptser and Shiryayev (1989))

$$\begin{aligned}
\langle M_0 \rangle_t &= \int_0^t K^2\left(\frac{X_s - x}{h}\right) g^2(X_s)\,\mathrm{d}s, \\
\langle M_0, M_1 \rangle_t &= \int_0^t K^2\left(\frac{X_s - x}{h}\right) \frac{X_s - x}{h}\, g^2(X_s)\,\mathrm{d}s, \\
\langle M_1 \rangle_t &= \int_0^t K^2\left(\frac{X_s - x}{h}\right) \left(\frac{X_s - x}{h}\right)^2 g^2(X_s)\,\mathrm{d}s,
\end{aligned}$$

so that $\langle M_0 \rangle_T = V_{0,h}$, $\langle M_0, M_1 \rangle_T = V_{1,h}$ and $\langle M_1 \rangle_T = V_{2,h}$. This yields

$$\begin{aligned}
\xi_h(x) &= v_{2,h} M_{0,T} - v_{1,h} M_{1,T}, \\
\sigma_h^2(x) &= v_{2,h}^2 \langle M_0 \rangle_T - 2 v_{1,h} v_{2,h} \langle M_0, M_1 \rangle_T + v_{1,h}^2 \langle M_1 \rangle_T.
\end{aligned}$$

Denote

$$u_h = \frac{v_{1,h}}{v_{2,h}} = \frac{\mu_{1,h}}{\mu_{2,h}}.$$

Obviously

$$\begin{aligned}
&\boldsymbol{P}\left(|\xi_h| > \lambda \sigma_h(x), \mathcal{A}_h\right) \\
&\quad = \boldsymbol{P}\left(|M_{0,T} - u_h M_{1,T}| > \lambda \sqrt{\langle M_0 \rangle_T - 2u_h \langle M_0, M_1 \rangle_T + u_h^2 \langle M_1 \rangle_T},\ \mathcal{A}_h\right).
\end{aligned}$$

To evaluate from above the right side of this equality, we apply the general result from Proposition 6.2, see Appendix. First we check the required conditions. The value $|u_h|$, being restricted to $\mathcal{A}_h$, can be bounded as:

$$|u_h| \le \left| \frac{\sqrt{\rho\, \mu_{0,h}\, \mu_{2,h}}}{\mu_{2,h}} \right| \le \sqrt{\rho r}.$$

Note now that

$$\begin{aligned}
\frac{\langle M_1 \rangle_T}{\langle M_0 \rangle_T - 2u_h \langle M_0, M_1 \rangle_T + u_h^2 \langle M_1 \rangle_T} &= \frac{V_{2,h}}{V_{0,h} - 2u_h V_{1,h} + u_h^2 V_{2,h}} \\
&= \frac{V_{2,h}^2}{V_{0,h} V_{2,h} - V_{1,h}^2 + (V_{1,h} - u_h V_{2,h})^2},
\end{aligned}$$

and it holds on $\mathcal{A}_h$ in view of $V_{2,h} \le V_{0,h}$

$$\frac{\langle M_1 \rangle_T}{\langle M_0 \rangle_T - 2u_h \langle M_0, M_1 \rangle_T + u_h^2 \langle M_1 \rangle_T} \le \frac{V_{2,h}^2}{(1-\rho) V_{0,h} V_{2,h}} \le \frac{1}{1-\rho}.$$

In addition, the definition of $\mathcal{A}_h$ provides the following bounds for $\sigma_h^2(x)$ on this set

$$\frac{\sigma_h^2(x)}{Th\, v_{2,h}^2} = \frac{Th\, \sigma_h^2(x)}{(Th\, v_{2,h})^2} \le \frac{bB}{b^2} = \frac{B}{b},$$

$$\frac{\sigma_h^2(x)}{Th\, v_{2,h}^2} = \frac{Th\, \sigma_h^2(x)}{(Th\, v_{2,h})^2} \ge \frac{b}{(bB)^2} = \frac{1}{bB^2}.$$

Applying now Proposition 6.2 we get

$$\boldsymbol{P}\left(|\xi_h| > \lambda\sigma_h(x), \mathcal{A}_h\right) \le 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}. \tag{5.7}$$

5.4. Proof of Theorem 3.1

Summing up the decomposition (5.3) and the bounds (5.6), (5.7), we get

$$\begin{aligned}\boldsymbol{P}&\left(\left|\widetilde{f}_h(x) - f(x)\right| > c\Delta_h(x) + \lambda\sigma_h(x),\, \mathcal{A}_h\right)\\ &\le 4e\log(4B^3)\left(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\,\lambda^2\right)\lambda\exp\left(-\frac{\lambda^2}{2}\right).\end{aligned}$$

This leads to the required bound from Theorem 3.1.

5.5. Proof of Theorem 4.1

Let h_{id} be shown in the theorem. Recall that $\mathcal{A}^* = \bigcap_{h\in\mathcal{H}} \mathcal{A}_h$. We use an obvious inequality

$$\begin{aligned}\boldsymbol{P}&\left(\left|\widehat{f}(x) - f(x)\right| > (\lambda+\lambda^*)\sigma_{h_{\mathrm{id}}}(x),\, \mathcal{A}^*\right)\\ &\le \boldsymbol{P}\left(\left|\widehat{f}(x) - f(x)\right| > (\lambda+\lambda^*)\sigma_{h_{\mathrm{id}}}(x),\, \widehat{h} \ge h_{\mathrm{id}},\, \mathcal{A}^*\right) + \boldsymbol{P}\left(\widehat{h} < h_{\mathrm{id}},\, \mathcal{A}^*\right).\end{aligned}$$

Since $\sigma_h(x)$ decreases in h, we have on the set $\{\widehat{h} \ge h_{\mathrm{id}}\} \cap \mathcal{A}^*$ in view of the definition of $\widehat{h}$

$$|\widetilde{f}_{\widehat{h}}(x) - \widetilde{f}_{h_{\mathrm{id}}}(x)| \le \lambda_1\bigl(\sigma_{\widehat{h}}(x) + \sigma_{h_{\mathrm{id}}}(x)\bigr) + 2D\sigma_{\widehat{h}}(x) \le 2(\lambda_1 + D)\sigma_{h_{\mathrm{id}}}(x).$$

Further, using the inequality $c\Delta_{h_{\mathrm{id}}}(x) \le D\sigma_{h_{\mathrm{id}}}(x)$ and Theorem 3.1, we get

$$\begin{aligned}\boldsymbol{P}&\left(|\widetilde{f}_{h_{\mathrm{id}}}(x) - f(x)| > (D+\lambda)\sigma_{h_{\mathrm{id}}}(x), \mathcal{A}^*\right)\\ &\le \boldsymbol{P}\left(|\widetilde{f}_{h_{\mathrm{id}}}(x) - f(x)| > \lambda\sigma_{h_{\mathrm{id}}}(x) + c\Delta_{h_{\mathrm{id}}}(x)\,,\, \mathcal{A}^*\right)\\ &\le \left(C_1\lambda + C_2\lambda^3\right)e^{-\frac{\lambda^2}{2}},\end{aligned}$$

where

$$\begin{aligned}C_1 &= 4e\log(4B^3),\\ C_2 &= 4e\log(4B^3)\,4r\sqrt{\frac{1+r}{1-\rho}}\,.\end{aligned}$$

Hence

$$\boldsymbol{P}\left(|\widehat{f}(x)-f(x)|>(\lambda+\lambda^{*})\sigma_{h_{\mathrm{id}}}(x),\,\mathcal{A}^{*},\widehat{h}\geq h_{\mathrm{id}}\right)\leq\left(C_{1}\lambda+C_{2}\lambda^{3}\right)e^{-\frac{\lambda^{2}}{2}} \tag{5.8}$$

and it only remains to evaluate $\boldsymbol{P}(\widehat{h}<h_{\mathrm{id}},\,\mathcal{A}^{*})$. Due to the definition of $\widehat{h}$, we have

$$\begin{aligned}&\{\widehat{h}<h_{\mathrm{id}},\,\mathcal{A}^{*}\}\\&\quad\subseteq\bigcup_{h\in\mathcal{H}\,:\,h<h_{\mathrm{id}}}\ \bigcup_{\eta\in\mathcal{H}\,:\,\eta<h}\left\{|\widehat{f}_{h}(x)-\widehat{f}_{\eta}(x)|>\lambda_{1}\Big(\sigma_{h}(x)+\sigma_{\eta}(x)\Big)+2D\sigma_{h}(x),\,\mathcal{A}^{*}\right\}.\end{aligned}$$

We now use that for every $\eta,h\in\mathcal{H}$ with $\eta<h<h_{\mathrm{id}}$

$$\begin{aligned}c\Delta_{h}(x)&\leq c\Delta_{h_{\mathrm{id}}}(x)\leq D\sigma_{h_{\mathrm{id}}}(x)\leq D\sigma_{h}(x),\\c\Delta_{\eta}(x)&\leq c\Delta_{h_{\mathrm{id}}}(x)\leq D\sigma_{h_{\mathrm{id}}}(x)\leq D\sigma_{h}(x).\end{aligned}$$

Therefore by Theorem 3.1

$$\begin{aligned}&\boldsymbol{P}\left(|\widetilde{f}_{h}(x)-\widetilde{f}_{\eta}(x)|>\lambda_{1}\big(\sigma_{h}(x)+\sigma_{\eta}(x)\big)+2D\sigma_{h}(x)\,,\,\mathcal{A}^{*}\right)\\&\quad\leq\boldsymbol{P}\left(|\widetilde{f}_{h}(x)-f(x)|>\lambda_{1}\sigma_{h}(x)+c\Delta_{h}(x)\,,\,\mathcal{A}_{h}\right)\\&\qquad+\boldsymbol{P}\left(|\widetilde{f}_{\eta}(x)-f(x)|>\lambda_{1}\sigma_{\eta}(x)+c\Delta_{\eta}(x)\,,\,\mathcal{A}_{\eta}\right)\\&\quad\leq2\left(C_{1}\lambda_{1}+C_{2}\lambda_{1}^{3}\right)e^{-\frac{\lambda_{1}^{2}}{2}}.\end{aligned}$$

Clearly the total number of pairs $\eta,h\in\mathcal{H}$, satisfying $\eta<h<h_{\mathrm{id}}$, is at most $(\#\mathcal{H})^{2}/2$. Therefore

$$\boldsymbol{P}\left(\widehat{h}<h_{\mathrm{id}}\right)\leq(\#\mathcal{H})^{2}\left(C_{1}\lambda_{1}+C_{2}\lambda_{1}^{3}\right)e^{-\frac{\lambda_{1}^{2}}{2}}.$$

This bound coupled with (5.8) implies the desired assertion.

6. Appendix. Deviation probabilities for martingales

In the Appendix we present two general results for continuous martingales. The first result describes some properties of real-valued martingales, while the second one deals with martingales valued in $\mathbb{R}^{2}$.

6.1. The scalar case

Let M_{t} be a continuous martingale with $M_{0}=0$ and with the predictable quadratic variation $\langle M\rangle_{t}$.

Proposition 6.1. *For every $T > 0$, $\vartheta > 0$, $S \ge 1$ and $\lambda \ge 1$*

$$\boldsymbol{P}\left(|M_T| > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \le 4\lambda\sqrt{e}\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}.$$

Proof. We use

$$\begin{aligned}\boldsymbol{P}&\left(|M_T| > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right)\\ &\le \boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right)\\ &\quad + \boldsymbol{P}\left(M_T < -\lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right).\end{aligned}$$

We estimate separately each term in the right side of this inequality.

Given $a > 1$, introduce the geometric series $\vartheta_k = \vartheta a^k$ and define the sequence of random events $\mathcal{C}_k = \{\vartheta_k \le \sqrt{\langle M\rangle_T} < \vartheta_{k+1}\}$, $k = 0, 1, \ldots$. Then clearly

$$\begin{aligned}(6.1)\qquad \boldsymbol{P}&\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right)\\ &\le \sum_{k\ge 0}^{K} \boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S,\ \mathcal{C}_k\right).\end{aligned}$$

where K is the integer part of $\log_a S$. We now bound each term in this sum. Let, with $\gamma \in \mathbb{R}$,

$$Z_t(\gamma) = \exp\left(\gamma M_t - \frac{\gamma^2}{2}\langle M\rangle_t\right).$$

The random process $Z_t(\gamma)$ is the continuous local martingale and, being positive, it is the supermartingale (see Problem 1.4.4 in Liptser and Shiryayev (1986)). Therefore for every $T > 0$,

$$(6.2)\qquad \boldsymbol{E}Z_T(\gamma) \le 1.$$

For fixed k, we pick $\gamma_k = \frac{\lambda}{\vartheta_k}$ and use (6.2) for the inequality

$$1 \ge \boldsymbol{E}Z_T(\gamma_k)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \mathcal{C}_k\right)$$

which implies

$$\begin{aligned}1 &\ge \boldsymbol{E}\exp\left(\frac{\lambda}{\vartheta_k}M_T - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right)\\ &\ge \boldsymbol{E}\exp\left(\frac{\lambda^2}{\vartheta_k}\sqrt{\langle M\rangle_T} - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right)\\ &\ge \boldsymbol{E}\exp\left\{\inf_{\vartheta_k \le v \le \vartheta_{k+1}}\left(\frac{\lambda^2 v}{\vartheta_k} - \frac{\lambda^2 v^2}{2\vartheta_k^2}\right)\right\}\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right).\end{aligned}$$

It is easy to check that "$\inf_{\vartheta_k \le v \le \vartheta_{k+1}}$" is attained at the point $v = \vartheta_{k+1} = a\vartheta_k$ so that

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \le \exp\left\{-\lambda^2\left(a - \frac{a^2}{2}\right)\right\}.$$

Combining this bound with (6.1) and the use of $K \le \log_a S$ yields

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \le (1+\log_a S)\exp\left\{-\lambda^2\left(a-\frac{a^2}{2}\right)\right\}.$$

Since the left hand side of this inequality does not depend on a, its right side can be optimized w.r.t. a. This leads to the choice $a = 1 + 1/\lambda$. Then

$$\lambda^2\left(a-\frac{a^2}{2}\right) = \lambda^2\left\{1+\frac{1}{\lambda}-\frac{1}{2}\left(1+\frac{1}{\lambda}\right)^2\right\} = \frac{1}{2}(\lambda^2-1)$$

and, since $\log(1+1/\lambda) \ge 1/(2\lambda)$ for $\lambda \ge 1$, it also holds $\log_a S \le 2\lambda \log S$. Hence

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \le 2\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}.$$

In the similar way we obtain

$$\boldsymbol{P}\left(M_T < -\lambda\sqrt{\langle M\rangle_T},\, \vartheta \le \sqrt{\langle M\rangle_T} \le \vartheta S\right) \le 2\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}$$

and the assertion follows. □

6.2. **The vector case**

Here, we consider continuous vector martingale M_t valued in $\mathbb{R}^2$ with components $M_{0,t}$ and $M_{1,t}$. Define

$$\begin{aligned} V_{0,t} &= \langle M_0\rangle_t \\ V_{1,t} &= \langle M_0, M_1\rangle_t \\ V_{2,t} &= \langle M_1\rangle_t. \end{aligned}$$

Let u be a random variable and

$$\sigma_t^2 = V_{0,t} - 2uV_{1,t} + u^2 V_{2,t}.$$

For a fixed time moment T and constants $\vartheta > 0$, $S \ge 1$, $\beta \ge 0$ and $\rho \in (0,1)$, introduce the event

$$\mathcal{A}_T = \left\{\begin{array}{c} \vartheta \le \sigma_T^2 \le \vartheta S \\ V_{1,T}^2 \le \rho V_{0,T} V_{2,T} \\ |u| \le \beta \end{array}\right\}. \tag{6.3}$$

Proposition 6.2. *Let M_t be a martingale with values in $\mathbb{R}^2$ such that $V_{0,T} \ge V_{2,T}$. Then, with $\mathcal{A}_T$ from (6.3), it holds for every $\lambda \ge \sqrt{2}$,*

$$\boldsymbol{P}\left(|M_{0,T} - uM_{1,T}| > \lambda\sigma_T,\, \mathcal{A}_T\right) \le 4e\log(4S)\left(1+4\beta\sqrt{\frac{1+\beta}{1-\rho}}\,\lambda^2\right)\lambda e^{-\frac{\lambda^2}{2}}.$$

Proof. For fixed β, ρ, and λ define δ by the equality

$$\frac{2\delta(1+\beta)}{1-\rho} = \lambda^{-2} \tag{6.4}$$

and denote by $D_\delta = \{\alpha_k = k\delta : k \in \mathbb{N}, \, |\alpha| \le \beta\}$ the discrete grid with the step δ in the interval $[-\beta, \beta]$.

Let ν_+ (respectively ν_-) be the random variable valued in D_δ which is closest to u from above (respectively from below). Then clearly

$$\begin{aligned} |\nu_\pm - u| &\le \delta. && (6.5) \\ |M_{0,T} - uM_{1,T}| &\le \max\{|M_{0,T} - \nu_- M_{1,T}|, |M_{0,T} - \nu_+ M_{1,T}|\}. && (6.6) \end{aligned}$$

Let now ν be one of ν_- and ν_+. Then by the construction $|\nu - u| \le \delta$. The next step is to show that on the set $\mathcal{A}_T$ it holds

$$1 - \lambda^{-2} \le \frac{V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T}}{\sigma_T^2} \le 1 + \lambda^{-2} \tag{6.7}$$

Indeed

$$\begin{aligned} \sigma_T^2 &= V_{0,T} - 2uV_{1,T} + u^2 V_{2,T} \\ &= V_{0,T} - \frac{V_{1,T}^2}{V_{2,T}} + V_{2,T}\left(u - \frac{V_{1,T}}{V_{2,T}}\right)^2 \\ &\ge \frac{V_{0,T}V_{2,T} - V_{1,T}^2}{V_{2,T}} \\ &\ge (1-\rho)V_{0,T} \end{aligned}$$

and the use of $V_{2,T} \le V_{0,T}$ leads to the bound

$$\begin{aligned} \frac{|V_{1,T}|}{\sigma_T^2} &\le \frac{\sqrt{\rho V_{0,T}V_{2,T}}}{(1-\rho)V_{0,T}} \le \frac{\sqrt{\rho}}{1-\rho} \le (1-\rho)^{-1}, \\ \frac{V_{2,T}}{\sigma_T^2} &\le \frac{V_{2,T}}{(1-\rho)V_{0,T}} \le (1-\rho)^{-1}. \end{aligned}$$

Since on the set $\mathcal{A}$ it holds $|u| \le \beta$ and by construction $\nu \le \beta$ we obtain, using the definition (6.4) of δ,

$$\begin{aligned} &\left|V_{0,T} - 2uV_{1,T} + u^2 V_{2,T} - (V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T})\right| \\ &\quad \le 2|V_{1,T}||u - \nu| + V_{2,T}\left|u^2 - \nu^2\right| \\ &\quad \le 2\delta(1-\rho)^{-1}\sigma_T^2 + 2\beta\delta(1-\rho)^{-1}\sigma_T^2 \\ &\quad = \sigma_T^2 \lambda^{-2} \end{aligned}$$

and (6.7) follows.

Since on the set $\mathcal{A}_T$ the value σ_T^2 is between ϑ and ϑS, we also get for $\nu = \nu_\pm$

$$(1-\lambda^{-2})\vartheta \le V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T} \le (1+\lambda^{-2})\vartheta S. \tag{6.8}$$

Now (6.6), (6.7) and (6.8) imply

$$
\begin{aligned}
&\{M_{0,T} - uM_{1,T}| > \lambda\sigma_T,\, \mathcal{A}_T\} \\
&\quad \subseteq \left\{ M_{0,T} - \nu_- M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{V_{0,T} - 2\nu_- V_{1,T} + \nu_-^2 V_{2,T}}\ ,\, \mathcal{A}_T \right\} \\
&\qquad \cup \left\{ M_{0,T} - \nu_+ M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{V_{0,T} - 2\nu_+ V_{1,T} + \nu_+^2 V_{2,T}}\ ,\, \mathcal{A}_T \right\} \\
&\quad \subseteq \bigcup_{\alpha \in D_\delta} \left\{ |M_{0,T} - \alpha M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}}\,,\, \mathcal{A}_{\alpha,T} \right\},
\end{aligned}
$$

where

$$
A_{\alpha,T} = \left\{ (1-\lambda^{-2})\vartheta \le V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T} \le (1+\lambda^{-2})\vartheta S \right\}.
$$

Now, for every $\alpha \in D_\delta$, the process $M_{0,t} - \alpha M_{1,t}$ is the continuous local martingale with $\langle M_0 - \alpha M_1 \rangle_T = V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}$. Proposition 6.1 and the inequalities $\lambda^2 \ge 2$ and

$$
\frac{\lambda^2}{1+\lambda^{-2}} \ge \lambda^2(1-\lambda^{-2}) = \lambda^2 - 1,
$$

yield

$$
\begin{aligned}
&\boldsymbol{P}\left(|M_{0,T} - \alpha M_{1,T}| > \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{V_{0,T} - 2\alpha V_{1,T} + \alpha^2 V_{2,T}}\,,\, A_{\alpha,T} \right) \\
&\qquad \le 4 \frac{\lambda}{\sqrt{1+\lambda^{-2}}} \left(1 + \log \frac{(1+\lambda^{-2})\vartheta S}{(1-\lambda^{-2})\vartheta} \right) \exp\left(-\frac{\lambda^2}{2(1+\lambda^{-2})} + \frac{1}{2} \right) \\
&\qquad \le 4\lambda \left(1 + \log \frac{3S}{2} \right) \exp\left(-\frac{\lambda^2}{2} + 1 \right).
\end{aligned}
$$

Since the number of different elements in D_δ is at most $1 + 2\beta\delta^{-1}$ and since δ from (6.4) fulfills $\delta^{-1} = \frac{2\lambda^2(1+\beta)}{1-\rho}$, it follows

$$
\begin{aligned}
\boldsymbol{P}\left(|M_{0,T} - uM_{1,T}| > \lambda\sigma_T\,,\, \mathcal{A}_T \right) &\le 4e \left(1 + \log \frac{3S}{2} \right) \left(1 + 2\beta\delta^{-1} \right) \lambda e^{-\frac{\lambda^2}{2}} \\
&\le 4e \log(4S) \left(1 + 4\beta \sqrt{\frac{1+\beta}{1-\rho}}\, \lambda^2 \right) \lambda e^{-\frac{\lambda^2}{2}}
\end{aligned}
$$

as required.

□

References

[1] Brown, L.D. and Low, M.G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann.-Statist.* **24** (1996), no. 6, 2524–2535.

[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] Dahlhaus, R. (1997). Fitting time series to nonstationary processes. *Ann.-Statist.* **25** (1997), no. 1, 1–37.

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

[5] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, no. 3, 425–455.

[6] 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.

[7] Doukhan, P. and Tsybakov, A.B. (1993). Nonparametric recurrent estimation in nonlinear ARX models. *Problems Inform. Trans.* **29**, no. 4, 318–327.

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

[9] Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** no. 3, 645–660.

[10] Genon-Catalot, V., Laredo, C. and Picard, D. (1992). Nonparametric estimation of the diffusion coefficient by wavelet methods, *Scand. J. Statist.* **19**, 317-335.

[11] Härdle, W. and Spokoiny, V. (1999). Adaptive estimation for a time inhomogeneous stochastic-volatility model. Unpublished manuscript.

[12] Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometr.* **81** 233–242.

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

[14] Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimation of the mean. *J. R. Stat. Soc.* **B 51** 3–14.

[15] Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models, *Prob. Theory and Rel. Fields,* **111**, 167–214.

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

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

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

[19] 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.

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

[21] Lepski, O. and Levit, B. (1997). Efficient adaptive estimation of infinitely differentiable function. Unpublished manuscript.

[22] 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.

[23] Lepski, O. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, **25**, no.6, 2512–2546.

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

[25] Milstein, G. and Nussbaum, M. (1994). Nonparametric estimation of a nonparametric diffusion model. *Prob. Theory and Rel. Fields.* To appear.

[26] Neumann, M.H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Annals of Statistics*, **26**, no. 5, 2014–2048.

[27] Ruppert, D. and Wand, M.P. , Holst, U. and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39** 262–273.

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

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

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

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

E-mail address: spokoiny@wias-berlin.de