The Annals of Statistics
2000, Vol. 28, No. 3, 815–836

ADAPTIVE DRIFT ESTIMATION FOR NONPARAMETRIC DIFFUSION MODEL[1]

BY VLADIMIR G. SPOKOINY

Weierstrass Institute for Applied Analysis and Stochastics

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

$$dX_t = f(X_t)\,dt + g(X_t)\,dw_t, \qquad X_0 = x_0, \qquad 0 \le t \le T. \tag{1.1}$$

Here w_t is a standard Wiener process and T is the *observation time*. The functions f, g entering in (1.1) are usually referred to as *drift* and *diffusion* coefficients. 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; see 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 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 recently, especially in view of applications to financial mathematics. The estimation theory for diffusion-type processes is well developed under parametric modelling when the 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

Received December 1997; revised March 2000.

[1]Supported by the Deutsche Forschungsgemeinschaft, SFB 373 "Simulation and Quantification of Economic Processes" at Humboldt University.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62M99.

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

detail. The known results concern only 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 (1998) 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 (2000), 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, Wand, Holst and Hössjer (1997), Fan and Yao (1998).

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 such that the observed process is stationary or near stationary within this interval; see, for example, Dahlhaus (1997) for more discussion. Statistical inference under local stationary assumption requires studying some nonasymptotic properties of statistical procedures. The reader is referred to forthcoming paper by Mercurio and Spokoiny (2000) 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 an ergodic property of the slow component nor large observation time T is assumed. We propose a locally linear estimator of $f(x)$ with a data-driven bandwidth choice which goes back to Lepski (1990). Lepski, Mammen and Spokoiny (1997) presented a slightly modified version of the original Lepski 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 bandwidth selector for kernel smoothing, again for the "white noise model." In this paper the procedure is adapted to locally linear smoothing in a diffusion-type model (1.1). The results compare the quality of the adaptive procedure to that of 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 $\log\log T$ to some power 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 collected in Sections 5 and 6.

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 (1.1) 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,

$$(\tilde{\theta}_0, \tilde{\theta}_1) = \underset{\theta_0,\,\theta_1}{\operatorname{argmax}} \left\{ \int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h} \right) dX_t - \frac{1}{2} \int_0^T \left(\theta_0 + \theta_1 \frac{X_t - x}{h} \right)^2 dt \right\}.$$

This quadratic optimization problem can be explicitly solved: with

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

one has

$$\tilde{\theta}_0 = \frac{\mu_2 \int_0^T dX_t - \mu_1 \int_0^T (X_t - x)/h \, dX_t}{\mu_0 \mu_2 - \mu_1^2},$$

$$\tilde{\theta}_1 = \frac{-\mu_1 \int_0^T dX_t + \mu_0 \int_0^T (X_t - x)/h \, dX_t}{\mu_0 \mu_2 - \mu_1^2}.$$

Since clearly $f(x) = \theta_0$, the value $\tilde{\theta}_0$ 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 a formal description, let us introduce a *kernel* function $K(u)$ which is assumed to be smooth, nonnegative, 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

$$\tilde{f}_h(x) = \frac{\mu_{2,h} \int_0^T K((X_t - x)/h) dX_t - \mu_{1,h} \int_0^T ((X_t - x)/h) K((X_t - x)/h) dX_t}{\mu_{0,h} \mu_{2,h} - \mu_{1,h}^2}, \tag{2.1}$$

where

$$\mu_{k,h} = \int_0^T \left(\frac{X_t - x}{h} \right)^k K\left(\frac{X_t - x}{h} \right) dt, \qquad k = 0, 1, 2. \tag{2.2}$$

The quality of estimate (2.1) essentially depends on the bandwidth h. Some useful properties of $\tilde{f}_h(x)$ 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 $\tilde{f}_h(x)$ from (2.1). We first formulate the required conditions on the coefficients f, g from (1.1). Then we present the result and discuss some of its corollaries.

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

(A_s) The function $f(u)$ is two times continuously differentiable in u. The function $g(u)$ is Lipschitz continuous in u and 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 nonasymptotic and there is no difference between the ergodic and nonergodic cases.

3.2. *Accuracy of the locally linear estimate.* To state the result, we introduce some additional notation. With $\mu_{k,h}$ defined in (2.2), set

$$\begin{aligned} \sigma_h^2(x) &= \frac{1}{D_h^2}\int_0^T \left(\mu_{2,h} - \mu_{1,h}\frac{X_t - x}{h}\right)^2 K^2\left(\frac{X_t - x}{h}\right) g^2(X_t)\,dt \\ &= v_{2,h}^2 V_{0,h} - 2v_{1,h}v_{2,h}V_{1,h} + v_{1,h}^2 V_{2,h}, \end{aligned} \tag{3.1}$$

where

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

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 $\tilde{f}_h(x)$. 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 $\tilde{f}_h(x)$. Note that for the regression set-up, 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 used to describe the design regularity in the regression setting. Our results are also similar to

those that 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 $\tilde{f}_h(x)$.

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

$$\mathscr{A}_h = \begin{cases} \dfrac{b}{Th} \leq v_{2,h} \leq \dfrac{bB}{Th}, \ \dfrac{b}{Th} \leq \sigma_h^2(x) \leq \dfrac{bB}{Th}, \\ \mu_{0,h} \leq r\mu_{2,h}, \qquad V_{0,h} \leq rV_{2,h}, \\ \mu_{1,h}^2 \leq \rho\mu_{0,h}\mu_{2,h}, \ V_{1,h}^2 \leq \rho V_{0,h}V_{2,h}. \end{cases}$$

Since X_t is the random process, the set $\mathscr{A}_h$ is random as well. In the sequel we study the properties of $\tilde{f}_h(x)$ restricted to the set $\mathscr{A}_h$; 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:

$$\Delta_h(x) = \sup_{|u-x| \leq h} \bigl| f(u) - f(x) - (u-x)f'(x) \bigr|, \tag{3.2}$$

where f' denotes the derivative of f. The next theorem describes some useful properties of the estimate (2.1).

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

$$P\biggl(\Bigl| \tilde{f}_h(x) - f(x) \Bigr| > c\Delta_h(x) + \lambda\sigma_h(x), \mathscr{A}_h \biggr) \leq \alpha(\lambda) \tag{3.3}$$

with

$$\alpha(\lambda) = 4e \log(4B^3) \biggl(1 + 4r\sqrt{\frac{1+r}{1-\rho}}\lambda^2 \biggr) \lambda \exp\biggl(-\frac{\lambda^2}{2} \biggr) \tag{3.4}$$

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

Informally, the result of the theorem means that for sufficiently large λ, the losses $|\tilde{f}_h(x) - f(x)|$ of the estimate $\tilde{f}_h(x)$, being restricted to $\mathscr{A}_h$, 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 $\mathscr{A}_h$.* The result of Theorem 3.1 describes the accuracy of the estimate $\tilde{f}_h(x)$ on the random set $\mathscr{A}_h$ only. Here we briefly discuss some related questions.

3.3.1. *Reason for restricting to* $\mathscr{A}_h$. It was mentioned previously that restricting to $\mathscr{A}_h$ allows eliminating 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, one has to estimate $f(x)$ with $x = X_t$, and the path $X_{[0,T]}$ obviously passes through x.

3.3.2. *Verifying the condition* $X_{[0,T]} \in \mathscr{A}_h$. Clearly the event $\mathscr{A}_h$ 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 path $X_{[0,T]}$ belongs to $\mathscr{A}_h$ or not. If $X_{[0,T]}$ does not belong to $\mathscr{A}_h$, we are not able to guarantee a reasonable quality for the estimate $\tilde{f}_h(x)$.

3.3.3. *The conditions entering into the definition of* $\mathscr{A}_h$. The conditions $0 \le K(u) \le 1$ and $K(u) = 0$ for $|u| \ge 1$ imply $\mu_{2,h} \le \mu_{0,h}$ and $V_{2,h} \le V_{0,h}$. Further, by the Cauchy–Schwarz inequality, it holds $\mu_{1,h}^2 \le \mu_{0,h}\mu_{2,h}$ and $V_{1,h}^2 \le V_{0,h}V_{2,h}$. The conditions $\mu_{0,h} \le r\mu_{2,h}$, $V_{0,h} \le V_{2,h}$, $\mu_{1,h}^2 \le \rho\mu_{0,h}\mu_{2,h}$, and $V_{1,h}^2 \le \rho V_{0,h}V_{2,h}$ with $\rho < 1$ and $r \ge 1$ 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 $K(u) = 1(|u| \le 1)$, then $\mu_{k,h} = V_{k,h}$ for $k = 0, 1, 2$ and $\sigma_h^2(x) = v_{2,h}/(\mu_{0,h}\mu_{2,h} - \mu_{1,h}^2)$.

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 $\mathscr{A}_h$, is optional and they even may depend on T.

For a regular design in the regression set-up, it holds $\mu_{1,h} = V_{1,h} = 0$. If, in addition, $g(u)$ is constant in the interval $[x-h, x+h]$, then $\mu_{0,h} = r(K)\mu_{2,h}$ and $V_{0,h} = r(K)V_{2,h}$ with $r(K) = \int K(u)\,du(\int u^2K(u)\,du)^{-1}$. Therefore, one reasonable choice would be $\rho = 1/2$ and $r = 2r(K)$.

Concerning the choice of the parameters b, B, note that the upper bound (3.3) from Theorem 3.1 does not depend 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 $\log(4B^3)$. Simple heuristic consideration prompt a possible choice $b = h_{\min}$ and $B = T$.

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 $1-\mathbf{P}(\mathscr{A}_h)$ converges to zero exponentially fast as $T \to \infty$. Since obviously

$$\mathbf{P}\Big(\Big|\tilde{f}_h(x) - f(x)\Big| > c\Delta_h(x) + \lambda\sigma_h(x)\Big)$$
$$\le \mathbf{P}\Big(\Big|\tilde{f}_h(x) - f(x)\Big| > c\Delta_h(x) + \lambda\sigma_h(x), \mathscr{A}_h\Big) + \mathbf{P}(\mathscr{A}_h),$$

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

3.4. *Quality of estimation under smoothness assumptions.* Due to the assumptions (A_s) from Section 3.1, 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,

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

Then the value $\Delta_h(x)$ defined in (3.2), is bounded above by $Lh^2/2$. On the other hand, on the set $\mathscr{A}_h$ the stochastic variance $\sigma_h^2(x)$ is of order $(Th)^{-1}$. Therefore, following the standard approach in nonparametric estimation, the bandwidth h can be chosen by balancing the accuracy of approximation and the stochastic error: $Lh^2 \asymp (Th)^{-1/2}$. (The symbol $\asymp$ here means equivalence in order; that is, the ratio remains bounded as T grows.) This leads to the choice $h \asymp (TL^2)^{-1/5}$ 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) for the related results for the "white noise" model]. 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 the value $\sigma_h^2(x)$ is defined as

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

with

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

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

Let us introduce two random processes,

$$Z'_t = \int_0^t K\Big(\frac{X_s - x}{h}\Big)dX_s \quad \text{and} \quad Z''_t = \int_0^t K\Big(\frac{X_s - x}{h}\Big)\frac{X_s - x}{h}\,dX_s$$

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

$$(Z'_T)^2 = 2\int_0^T Z'_t dZ'_t + V_{0,h},$$

$$(Z''_T)^2 = 2\int_0^T Z''_t dZ''_t + V_{2,h},$$

$$Z'_T Z''_T = \int_0^T Z'_t dZ''_t + \int_0^T Z''_t dZ'_t + V_{1,h}.$$

Hence $V_{0,h} = (Z'_T)^2 - 2\int_0^T Z'_t dZ'_t$, 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. Below we discuss one adaptive (data driven) approach which goes back to the idea of pointwise adaptive estimation; see Lepski (1990), Lepski and Spokoiny (1997) and Spokoiny (1998).

The idea of the method can be explained as follows. In the light of Theorem 3.1, we could be interested in selecting a bandwidth h which leads to a possibly small sum of the form $c\Delta_h(x) + \lambda\sigma_h(x)$ among all considered bandwidth values h. This sum comprises two terms. The first one ("bias") characterizes the accuracy of local approximation of the underlying drift function f by the linear functions and it typically increases with h. The second term is proportional to the conditional standard deviation $\sigma_h(x)$ which typically decreases with 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 $h \in \mathscr{H}$. [If this assumption is not fulfilled for the original set $\mathscr{H}$, i.e., if there is $h' < h \in \mathscr{H}$ with the property $\sigma_h^2(x) \geq \sigma_{h'}^2(x)$, then we simply exclude h from $\mathscr{H}$.]

Therefore, a "good" (or "ideal") choice h_{id} corresponds to a possibly large bandwidth h (which makes the stochastic component of the estimate small) still providing that the "bias" component $c\Delta_h(x)$ is not significant larger than $\sigma_h(x)$. (We call h_{id} an "ideal" bandwidth since its definition relies on the unknown function $\Delta_h(x)$.) The latter property is clearly fulfilled for all smaller bandwidths $h \leq h_{\mathrm{id}}$. Therefore, if h_{id} is "good" and $h < h_{\mathrm{id}}$, then the two corresponding estimates $\tilde{f}_{h_{\mathrm{id}}}(x)$ and $\tilde{f}_h(x)$ should not differ significantly.

The proposed procedure can be viewed as a family of tests for whether the estimate $\tilde{f}_h(x)$ for a bandwidth candidate h differs significantly from estimates $\tilde{f}_\eta(x)$ with smaller bandwidths $\eta < h$. The latter is done on the base of Theorem 3.1 which allows bounding with a large probability the difference

$|\tilde{f}_h(x) - \tilde{f}_\eta(x)|$ by $\lambda\sigma_h(x) + \lambda\sigma_\eta(x) + c\Delta_h(x) + c\Delta_\eta(x)$ provided λ is sufficiently large. The terms $c\Delta_h(x)$ and $c\Delta_\eta(x)$ in this sum are unknown but, if h is "good," that is, if $\Delta_h(x) \ll \sigma_h(x)$, then their contribution is negligible. In opposition a significant deviation of $|\tilde{f}_h(x) - \tilde{f}_\eta(x)|$ over the level $\lambda\sigma_h(x) + \lambda\sigma_\eta(x)$ can be explained only by a large bias component indicating that h is not a "good" bandwidth. The procedure searches for the largest bandwidth h such that the hypothesis $\tilde{f}_h(x) = \tilde{f}_\eta(x)$ is not rejected for all $\eta < h$.

Now we present a formal description. Suppose a family $\mathscr{H}$ of bandwidth candidates h is fixed. For technical reasons, we assume that this set is finite and denote by H the number of its elements. Usually $\mathscr{H}$ is taken as a geometric grid,

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

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 $\mathscr{H}$ for which the observed path $X_{[0,T]}$ belongs to $\mathscr{A}_h$.

With every bandwidth value h we associate the estimate $\tilde{f}_h(x)$ of $f(x)$ and the corresponding conditional standard deviations $\sigma_h(x)$ which can be precisely calculated as described in Section 3.5.

Now, with two constants λ_1 and λ_2, define the adaptive choice of bandwidth by the following iterative procedure.

Initialization. Select the smallest bandwidth in $\mathscr{H}$.

Iteration. Select the next bandwidth h in $\mathscr{H}$ and calculate the corresponding estimate $\tilde{f}_h(x)$ and the conditional standard deviation $\sigma_h(x)$.

Testing. Reject h, if there exists one $\eta \in \mathscr{H}$ with $\eta < h$ such that

$$\left|\tilde{f}_h(x) - \tilde{f}_\eta(x)\right| > \lambda_1\sigma_\eta(x) + \lambda_2\sigma_h(x). \tag{4.1}$$

Loop. If h is not rejected, then continue with the *iteration step* by choosing a larger bandwidth h in $\mathscr{H}$. Otherwise, set $\hat{h}$ = "the latest nonrejected h."

The proposed rule can be packed in the following form:

$$\begin{aligned}\hat{h} = \max\{h \in \mathscr{H}:\ & |\tilde{f}_{h'}(x) - \tilde{f}_\eta(x)| \\ & \le \lambda_1\sigma_\eta(x) + \lambda_2\sigma_{h'}(x)\ \forall\, h',\ \eta \in \mathscr{H},\ \eta < h' \le h\}\end{aligned} \tag{4.2}$$

The choice of the parameters λ_1, λ_2 and the set $\mathscr{H}$ is discussed in Section 4.1.

Finally, to define our adaptive estimate, we plug the data-driven bandwidth $\hat{h}$ in the estimate $\tilde{f}_h(x)$, that is $\hat{f}(x) \equiv \tilde{f}_{\hat{h}}(x)$.

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

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

THEOREM 4.1. *The estimate* $\hat{f}(x) \equiv \tilde{f}_{\hat{h}}(x)$ *with* $\hat{h}$ *from* (4.2) *and* $\lambda_2 \geq \lambda_1$ *fulfills the following property*:

$$P\bigl(|\hat{f}(x) - f(x)| > (2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x), \mathscr{A}^*\bigr) \leq \sum_{\eta \in \mathscr{H}:\ \eta \leq h_{\mathrm{id}}} \alpha(\lambda_\eta), \tag{4.3}$$

where $\alpha(\lambda)$ *is defined in* (3.4) *and*

$$\lambda_\eta = \lambda_1 - c\Delta_\eta(x)/\sigma_\eta(x). \tag{4.4}$$

4.1. *The choice of parameters* $\lambda_1, \lambda_2, h_{\min}, h_{\max}$ *and* a. Different proposals for the choice of grid $\mathscr{H}$ are discussed in Lepski, Mammen and Spokoiny (1997) and in Lepski and Spokoiny (1997). One possible choice for the grid $\mathscr{H}$ reads as follows: $h_{\min} = 1/T$, $h_{\max} = 1$, $a = \sqrt{2}$, although these values can be changed without essential influence on the quality of the procedure.

The choice of parameters λ_1, λ_2, entering in (4.2), plays a more important role. We start with the following general remark: the upper bound for the risk from Theorem 4.1 is rather rough and should be used with care for the parameter selection. However, it delivers some useful qualitative information about this choice which can be used for a theoretical study. The bound in (4.3) shows that the probability for $|\hat{f}(x) - f(x)|$ of being large is small, provided that the value $\sum_{\eta \in \mathscr{H}:\ \eta \leq h_{\mathrm{id}}} \alpha(\lambda_\eta)$ is sufficiently small. Here we discuss briefly the specific case when the values $\Delta_\eta(x)$ vanish. The general case can be relatively easily reduced to that one. Indeed, a "good" bandwidth h_{id} can be defined by trade-off arguments between the "bias" $c\Delta_{h_{\mathrm{id}}}(x)$ and the conditional standard deviation $\sigma_{h_{\mathrm{id}}}(x)$; that is, h_{id} is the maximal h from $\mathscr{H}$ with $c\Delta_h(x) \leq D\sigma_h(x)$ for some fixed value D. Taking D small enough provides that $c\Delta_\eta(x) \ll \sigma_\eta(x)$ for all $\eta \leq h_{\mathrm{id}}$.

If $\Delta_\eta(x)$ vanishes for all such η, then $\lambda_\eta = \lambda_1$ and $\sum_{\eta \in \mathscr{H}:\ \eta \leq h_{\mathrm{id}}} \alpha(\lambda_\eta) \leq H\alpha(\lambda_1)$. Therefore, λ_1 should be selected in a way to provide that $H\alpha(\lambda_1)$ is sufficiently small. This leads to the choice

$$\lambda_1 \approx \sqrt{2\,\log(H) + \lambda^2}$$

with some fixed constant λ so that

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

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

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

There are many degrees of freedom in the choice of λ_2. The constraint $\lambda_2 \geq \lambda_1$ from Theorem 4.1 is of technical matter and it is used only in theoretical investigations. It can be skipped in practical applications. Simulation results show a reasonable (and very similar) performance of the presented procedure with $\lambda_1 \approx 2$ and $\lambda_2 = 1$, or $\lambda_1 = \lambda_2 = 1.5$ in most cases. We refer to the forthcoming paper by Mercurio and Spokoiny (2000) for a more detailed discussion of

practical issues and for a proposal for a data-driven choice of the parameters λ_1 and λ_2 in the context of applications to finance time series.

4.2. *Accuracy 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 conditions $\sigma_h^2(x) \asymp (hT)^{-1}$ and the balance relation $c\Delta_h(x) \le D\sigma_h(x)$ yield for h_{id}:

$$h_{\mathrm{id}} \asymp (TL^2)^{-1/5}$$

so that $\sigma_{h_{\mathrm{id}}}(x) \asymp L^{1/5}T^{-2/5}$. Hence, the above-mentioned choice $\lambda_1 \approx \sqrt{2\log\log T}$ and $\lambda_2 = \lambda_1$ leads, due to Theorem 4.1, to the following accuracy of the adaptive estimation:

$$(2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x) \asymp L^{1/5}\left(\frac{\log\log T}{T}\right)^{2/5}.$$

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 then the "ideal" one within a $\log\log T$-factor only.

The origin of the $\log\log T$-factor in the rate of adaptive estimation can be easily explained. The total number 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 $\log(H)$ at some power.

The notion of "payment for adaptation" is now well understood in nonparametric estimation: if we have too many estimates to select among, we have to "pay" for the adaptive choice with 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 contradiction to 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 also be shown that the rate achieved by our estimate is optimal for pointwise adaptive estimation with a bounded loss function (see Spokoiny (1996) for similar results in the adaptive testing problem).

5. Proofs. In this section we prove Theorems 3.1 and 4.1.

5.1. *Proof of Theorem* 3.1. The proof of the theorem will be split into a few separate steps.

5.1.1. *Decomposition of* $\tilde{f}_h(x)$. We use two obvious identities characterizing the locally linear smoother: for $v_{1,h} = \mu_{1h}/D_h$ and $v_{2,h} = \mu_{2,h}/D_h$,

$$\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right) ds = 1,$$

$$\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h}\frac{X_s - x}{h} - v_{1,h}\frac{(X_s - x)^2}{h^2}\right) ds = 0$$

and hence

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

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

Due to (2.1) and (1.1), the estimate $\tilde{f}_h(x)$ can be represented as follows:

$$\begin{aligned}
\tilde{f}_h(x) &= v_{2,h}\int_0^T K\left(\frac{X_s - x}{h}\right) dX_s - v_{1,h}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h}\, dX_s \\
&= \int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right) f(X_s)\, ds \\
&\quad + v_{2,h}\int_0^t K\left(\frac{X_s - x}{h}\right) g(X_s)\, dw_s \\
&\quad - v_{1,h}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h}\, g(X_s)\, dw_s.
\end{aligned}$$

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

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

where, with $\delta(X_s, x) = f(X_s) - f(x) - ((X_s - x)/h)f'(x)$,

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

Below we evaluate separately each term in this decomposition.

5.1.2. *An upper bound for* $|r_h|$. Since $K^{((u-x)/h)}$ vanishes for any $u \notin [x-h, x+h]$ and $|\delta(X_x, x)| \le \Delta_h(x)$ for $|X_s - x| \le h$, we get

$$|r_h| \le \int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right)|\delta(X_s,\ x)|ds \tag{5.4}$$

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

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 $\mathscr{A}_h$,

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

We now show that

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

The Cauchy–Schwarz inequality applied to (5.4) gives

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

Next,

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

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

$$\int_0^T K\left(\frac{X_s - x}{h}\right)\left(v_{2,h} - v_{1,h}\frac{X_s - x}{h}\right)^2 ds$$

$$= \frac{1}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\left(\mu_{2,h} - \mu_{1,h}\frac{X_s - x}{h}\right)^2 ds$$

$$= \frac{\mu_{2,h}^2}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right) ds + \frac{\mu_{1,h}^2}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{(X_s - x)^2}{h^2} ds$$

$$-\frac{2\mu_{1,h}\mu_{2,h}}{D_h^2}\int_0^T K\left(\frac{X_s - x}{h}\right)\frac{X_s - x}{h} ds$$

$$= \frac{\mu_{2,h}^2\mu_{0,h} - \mu_{2,h}\mu_{1,h}^2}{D_h^2} = \frac{\mu_{2,h}}{D_h}.$$

Hence, in view of (5.5),

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

as required.

5.1.3. *An upper bound for* ξ_h. We study here some properties of the "stochastic term,"

$$\xi_h = \nu_{2,h}\int_0^T K\Big(\frac{X_s - x}{h}\Big)g(X_s)\,dw_s - \nu_{1,h}\int_0^T K\Big(\frac{X_s - x}{h}\Big)\frac{X_s - x}{h}g(X_s)\,dw_s.$$

Namely, we intend to show that the probability of the event $\{\xi_h > \lambda\sigma_h(x)\}$ with $\sigma_h(x)$ from (3.1) is small provided that λ is large enough. Set for $t \le T$,

$$M_{0,t} = \int_0^t K\Big(\frac{X_s - x}{h}\Big)g(X_s)\,dw_s,$$

$$M_{1,t} = \int_0^t K\Big(\frac{X_s - x}{h}\Big)\frac{X_s - x}{h}g(X_s)\,dw_s.$$

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)]

$$\langle M_0\rangle_t = \int_0^t K^2\Big(\frac{X_s - x}{h}\Big)g^2(X_s)\,ds,$$

$$\langle M_0, M_1\rangle_t = \int_0^t K^2\Big(\frac{X_s - x}{h}\Big)\frac{X_s - x}{h}g^2(X_s)\,ds,$$

$$\langle M_1\rangle_t = \int_0^t K^2\Big(\frac{X_s - x}{h}\Big)\Big(\frac{X_s - x}{h}\Big)^2 g^2(X_s)\,ds,$$

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

$$\xi_h(x) = \nu_{2,h}M_{0,T} - \nu_{1,h}M_{1,T},$$

$$\sigma_h^2(x) = \nu_{2,h}^2\langle M_0\rangle_T - 2\nu_{1,h}\nu_{2,h}\langle M_0, M_1\rangle_T + \nu_{1,h}^2\langle M_1\rangle_T.$$

Denote

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

Obviously,

$$\mathbf{P}(|\xi_h| > \lambda\sigma_h(x), \mathscr{A}_h) = \mathbf{P}\Big(|M_{0,T} - u_hM_{1,T}| > \lambda\sqrt{\langle M_0\rangle_T - 2u_h\langle M_0, M_1\rangle_T + u_h^2\langle M_1\rangle_T}, \mathscr{A}_h\Big).$$

To evaluate from above the right side of this equality, we apply the general result from Proposition A.2; see the Appendix. First we check the required conditions. The value $|u_h|$, being restricted to $\mathscr{A}_h$, 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 $\mathscr{A}_h$ in view of $V_{2,h} \le V_{0,h}$,

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

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

$$\begin{aligned} \frac{\sigma_h^2(x)}{Th\nu_{2,h}^2} &= \frac{Th\sigma_h^2(x)}{(Th\nu_{2,h})^2} \le \frac{bB}{b^2} = \frac{B}{b}, \\ \frac{\sigma_h^2(x)}{Th\nu_{2,h}^2} &= \frac{Th\sigma_h^2(x)}{(Th\nu_{2,h})^2} \ge \frac{b}{(bB^2)} = \frac{1}{bB^2}. \end{aligned}$$

Applying now Proposition A.2 we get

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

5.1.4. *End of the proof.* Summing up the decomposition (5.3) and the bounds (5.6), (5.7), we get

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

This leads to the required bound from Theorem 3.1.

5.2. *Proof of Theorem 4.1.* Let h_{id} be a "good" bandwidth. We intend to show that

$$\big\{ |\hat{f}(x) - f(x)| > (2\lambda_1 + \lambda_2) \sigma_{h_{\mathrm{id}}}(x) \big\} \subseteq \bigcup_{\eta \in \mathscr{H}(h_{\mathrm{id}})} \big\{ |\tilde{f}_\eta(x) - f(x)| > \lambda_1 \sigma_\eta(x) \big\},$$

where $\mathscr{H}(h) = \{\eta \in \mathscr{H}: \eta \le h\}$. This statement is equivalent to saying that the inequality $|\hat{f}(x) - f(x)| > (2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x)$ is impossible if

$$|\tilde{f}_\eta(x) - f(x)| \le \lambda_1 \sigma_\eta(x) \qquad \forall\, \eta \in \mathscr{H}(h_{\mathrm{id}}). \tag{5.8}$$

Obviously,

$$\begin{aligned}&\{|\hat{f}(x) - f(x)| > (2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x)\}\\ &\quad\subseteq \{|\hat{f}(x) - f(x)| > (2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x), \hat{h} > h_{\mathrm{id}}\} + \{h_{\mathrm{id}} \text{ is rejected}\}.\end{aligned}$$

We consider separately each event in the right side of this inequality.

It holds on the event $\{\hat{h} > h_{\mathrm{id}}\}$ in view of the definition of $\hat{h}$,

$$|\tilde{f}_{\hat{h}}(x) - \tilde{f}_{h_{\mathrm{id}}}(x)| \le \lambda_1 \sigma_{h_{\mathrm{id}}}(x) + \lambda_2 \sigma_{\hat{h}}(x) \le (\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x).$$

Next, by (5.8),

$$|\tilde{f}_{h_{\mathrm{id}}}(x) - f(x)| \le \lambda_1 \sigma_{h_{\mathrm{id}}}(x).$$

Hence, $\{\hat{h} > h_{\mathrm{id}}\}$ and (5.8) imply

$$\begin{aligned}|\hat{f}(x) - f(x)| &\le |\hat{f}_{\hat{h}}(x) - \tilde{f}_{h_{\mathrm{id}}}(x)| + |\tilde{f}_{h_{\mathrm{id}}}(x) - f(x)|\\ &\le (2\lambda_1 + \lambda_2)\sigma_{h_{\mathrm{id}}}(x).\end{aligned}$$

Now we study the event $\{h_{\mathrm{id}} \text{ is rejected}\}$. By definition,

$$\{h_{\mathrm{id}} \text{ is rejected}\} = \bigcup_{h \in \mathscr{H}(h_{\mathrm{id}})} \bigcup_{\eta \in \mathscr{H}(h)} \{|\tilde{f}_h(x) - \tilde{f}_\eta(x)| > \lambda_2 \sigma_h(x) + \lambda_1 \sigma_\eta(x)\}.$$

Condition (5.8) yields for every pair $\eta < h \in \mathscr{H}(h_{\mathrm{id}})$

$$|\tilde{f}_h(x) - \tilde{f}_\eta(x)| \le |\tilde{f}_h(x) - f(x)| + |\tilde{f}_\eta(x) - f(x)| \le \lambda_1(\sigma_h(x) + \sigma_\eta(x))$$

so that the event $\{h_{\mathrm{id}} \text{ is rejected}\}$ is impossible under (5.8) in view of $\lambda_2 \ge \lambda_1$.

It remains to bound the probability of the event in (5.8). With $\lambda_\eta = \lambda_1 - c\Delta_\eta(x)/\sigma_\eta(x)$, it holds by Theorem 3.1,

$$\begin{aligned}\mathbf{P}\bigl(|\tilde{f}_\eta(x) - f(x)| > \lambda_1 \sigma_\eta(x)\bigr) &= \mathbf{P}\bigl(|\tilde{f}_\eta(x) - f(x)| > \lambda_\eta \sigma_\eta(x) + c\Delta_\eta(x)\bigr)\\ &\ge \alpha(\lambda_\eta),\end{aligned}$$

where $\alpha(\lambda)$ is from (3.4) and hence,

$$\mathbf{P}\bigl(|\tilde{f}_\eta(x) - f(x)| \le \lambda_1 \sigma_\eta(x), \forall\, \eta \in \mathscr{H}(h_{\mathrm{id}})\bigr) \ge 1 - \sum_{\eta \in \mathscr{H}(h_{\mathrm{id}})} \alpha(\lambda_\eta).$$

This completes the proof of the theorem. □

APPENDIX

Deviation probabilities for martingales. 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 $\mathbb{R}^2$.

A.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 A.1. *For every $T > 0$, $\vartheta > 0$, $S \geq 1$ and $\lambda \geq 1$,*

$$\mathrm{P}\Big(|M_T| > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big) \leq 4\lambda\sqrt{e}(1+\log S)\exp\Big(-\frac{\lambda^2}{2}\Big).$$

PROOF. We use

$$\begin{aligned}\mathrm{P}&\Big(|M_T| > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big)\\ &\leq \mathrm{P}\Big(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big)\\ &\quad + \mathrm{P}\Big(M_T < -\lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big).\end{aligned}$$

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 $\mathscr{C}_k = \{\vartheta_k \leq \sqrt{\langle M\rangle_T} < \vartheta_{k+1}\}$, $k = 0, 1, \ldots$. Then clearly,

$$\begin{aligned}&\mathrm{P}\Big(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big)\\ &\quad\leq \sum_{k\geq 0}^{K} \mathrm{P}\big(M_T > \lambda\sqrt{\langle M\rangle_T},\, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S,\, \mathscr{C}_k\big),\end{aligned} \tag{A.1}$$

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

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

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 (1989)]. Therefore for every $T > 0$,

$$\mathrm{E}Z_T(\gamma) \leq 1. \tag{A.2}$$

For fixed k, we pick $\gamma_k = \lambda/\vartheta_k$ and use (A.22) for the inequality

$$1 \geq \mathrm{E}Z_T(\gamma_k)\mathrm{I}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathscr{C}_k\Big),$$

which implies

$$\begin{aligned}
1 &\geq \mathrm{E}\exp\Big(\frac{\lambda}{\vartheta_k}M_T - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\Big)\mathrm{I}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathscr{C}_k\Big), \\
&\geq \mathrm{E}\exp\Big(\frac{\lambda^2}{\vartheta_k}\sqrt{\langle M\rangle_T} - \frac{\lambda^2}{2\vartheta_k}\langle M\rangle_T\Big)\mathrm{I}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathscr{C}_k\Big), \\
&\geq \mathrm{E}\exp\Big\{\inf_{\vartheta_k \leq v \leq \vartheta_{k+1}}\Big(\frac{\lambda^2 v}{\vartheta_k} - \frac{\lambda^2 v^2}{2\vartheta_k^2}\Big)\Big\}\mathrm{I}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathscr{C}_k\Big),
\end{aligned}$$

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

$$\mathrm{P}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathscr{C}_k\Big) \leq \exp\Big\{-\lambda^2\Big(a - \frac{a^2}{2}\Big)\Big\}.$$

Combining this bound with (A.1) and the use of $K \leq \log_a S$ yields

$$\mathrm{P}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big) \leq (1 + \log_a S)\exp\Big\{-\lambda^2\Big(a - \frac{a^2}{2}\Big)\Big\}.$$

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

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

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

$$\mathrm{P}\Big(M_T > \lambda\sqrt{\langle M\rangle_T}, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big) \leq 2\sqrt{e}\lambda(1 + \log S)\exp\Big(-\frac{\lambda^2}{2}\Big).$$

In the similar way, we obtain

$$\mathrm{P}\Big(M_T < -\lambda\sqrt{\langle M\rangle_T}, \vartheta \leq \sqrt{\langle M\rangle_T} \leq \vartheta S\Big) \leq 2\sqrt{e}\lambda(1 + \log S)\exp\Big(-\frac{\lambda^2}{2}\Big)$$

and the assertion follows.

A.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

$$V_{0,t} = \langle M_0\rangle_t, \qquad V_{1,t} = \langle M_0, M_1\rangle_t, \qquad V_{2,t} = \langle M_1\rangle_t.$$

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 $\vartheta > 0$, $S \geq 1$, $\beta \geq 0$ and $\rho \in (0,1)$, introduce the event

$$\mathscr{A}_t = \begin{cases} \vartheta \leq \sigma_T^2 \leq \vartheta S, \\ V_{1,T}^2 \leq \rho V_{0,T} V_{2,T}, \\ |u| \leq \beta. \end{cases} \tag{A.3}$$

PROPOSITION A.2. *Let M_t be a martingale with values in $\mathbb{R}^2$ such that $V_{0,T} \geq V_{2,T}$. Then, with $\mathscr{A}_T$ from* (A.3)*, it holds for every $\lambda \geq \sqrt{2}$,*

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

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

$$\frac{2\delta(1+\beta)}{1-\rho} = \lambda^{-2} \tag{A.4}$$

and denote by $D_\delta = \{\alpha_k = k^\delta : k \in \mathbb{N}, |\alpha| \leq \beta\}$ 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,

$$|\nu_\pm - u| \leq \delta, \tag{A.5}$$

$$|M_{0,T} - uM_{1,T}| \leq \max\{|M_{0,T} - \nu_- M_{1,T}|, |M_{0,T} - \nu_+ M_{1,T}|\}. \tag{A.6}$$

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

$$1 - \lambda^{-2} \leq \frac{V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T}}{\sigma_T^2} \leq 1 + \lambda^{-2}. \tag{A.7}$$

Indeed,

$$\sigma_T^2 = V_{0,T} - 2uV_{1,T} + u^2 V_{2,T} = V_{0,T} - \frac{V_{1,T}^2}{V_{2,T}} + V_{2,T}\left(u - \frac{V_{1,T}}{V_{2,T}}\right)^2$$

$$\geq \frac{V_{0,T}V_{2,T} - V_{1,T}^2}{V_{2,T}} \geq (1-\rho)V_{0,T}$$

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

$$\frac{|V_{1,T}|}{\sigma_T^2} \leq \frac{\sqrt{\rho V_{0,T} V_{2,T}}}{(1-\rho)V_{0,T}} \leq \frac{\sqrt{\rho}}{1-\rho} \leq (1-\rho)^{-1},$$

$$\frac{V_{2,T}}{\sigma_T^2} \leq \frac{V_{2,T}}{(1-\rho)V_{0,T}} \leq (1-\rho)^{-1}.$$

Since on the set $\mathscr{A}$ it holds $|u| \le \beta$ and by construction $\nu \le \beta$ we obtain, using the definition (A.4) of δ,

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

and (A.7) follows.

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

$$(1-\lambda^{-2})\vartheta \le V_{0,T} - 2\nu V_{1,T} + \nu^2 V_{2,T} \le (1+\lambda^{-2})\vartheta S. \tag{A.8}$$

Now (A.6), (A.7) and (A.8) imply

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

where

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

Now, for every $\alpha \in D_\delta$, the process $M_{0,t} - \alpha M_{1,t}$ is a 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 A.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

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

Since the number of different elements in D_δ is at most $1 + 2\beta\delta^{-1}$ and since δ from (A.4) fulfills $\delta^{-1} = 2\lambda^2(1+\beta)/(1-\rho)$, it follows that

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

as required.

REFERENCES

BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535.

COLLOMB, G. and DOUKHAN, P. (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 Sér. I* **296** 859–862.

DAHLHAUS, R. (1997). Fitting time series to nonstationary processes. *Ann. Statist.* **25** 1–37.

DELYON, B. and JUDITSKY, A. (2000). On minimax identification of nonparametric autoregressive models. *Probab. Theory Relat. Fields* **116** 21–39.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DOUKHAN, P. and GHINDES, M. (1980). Estimations dans le processus "$X_{n+1} = f(X_n) + \varepsilon_n$". *C.R. Acad. Sci. Paris Sér. A* **291** 61–64.

DOUKHAN, P. and TSYBAKOV, A. B. (1993). Nonparametric recurrent estimation in nonlinear ARX models. *Problems Inform. Transmission* **29** 318–327.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J. and YAO, Q. (1988). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660.

GENON-CATALOT, V., LAREDO, C. and PICARD, D. (1992). Nonparametric estimation of the diffusion coefficient by wavelet methods. *Scand. J. Statist.* **19** 317–335.

HALL, P. and CARROLL, R. J. (1989). Variance function estimation in regression: the effect of estimation of the mean. *J. Roy. Statist. Soc. Ser. B* **51** 3–14.

HÄRDLE, W. and TSYBAKOV, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81** 233–242.

HÄRDLE, W. and VIEU, P. (1992). Kernel regression smoothing of time series. *J. Time Ser. Anal.* **13** 209–232.

GRAMA, I. and NUSSBAUM, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields* **111** 167–214.

IBRAGIMOV, J. A. and KHASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KATKOVNIK, V. JA. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach* (in Russian). Nauka, Moscow.

KUTOYANTS, YU. A. (1984a). On nonparametric estimation of trend coefficients in a diffusion process. In *Statistics and Control of Stochastic Processes* 230–250. Moscow.

KUTOYANTS, YU. A. (1984b). Parameter estimation for stochastic processes. In *R & E Research and Exposition in Mathematics* **6** (B. L. S. Prakasa, Rao ed.) Heldermann, Berlin.

LEPSKI, O. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 459–470.

LEPSKI, O. and LEVIT, B. (1997). Efficient adaptive estimation of infinitely differentiable function. Unpublished manuscript.

LEPSKI, O., MAMMEN, E. and SPOKOINY, V. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Ann. Statist.* **25** 929–947.

LEPSKI, O. and SPOKOINY, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546.

LIPTSER, R. and SHIRYAEV, A. (1989). *Theory of Martingales*. Kluwer, Dordrecht.

MERCURIO, D. and SPOKOINY, V. (2000). Statistical inference for time-inhomogeneous volatility models. http: //www.wias-berlin.de WIAS preprint 583, Berlin.

MILSTEIN, G. and NUSSBAUM, M. (1994). Nonparametric estimation of a nonparametric diffusion model. *Probab. Theory Related Fields*. **112** 535–543.

NEUMANN, M. H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* **26** 2014–2048.

RUPPERT, D., WAND, M. P., HOLST, U. and HÖSSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39** 262–273.

SPOKOINY, V. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498.

SPOKOINY, V. (1998) Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.* **26** 1356–1378.

TSYBAKOV, A. (1986). Robust reconstruction of functions by the local approximation. *Problems Inform. Transmission* **22** 133–146.

VERETENNIKOV, A. YU. (1991). On the averaging principle for systems of stochastic differential equations. *Math. USSR Sborn.* **69** 271–284.

WEIERSTRASS INSTITUTE
MOHRENSTR. 39
10117 BERLIN
GERMANY
E-MAIL: spokoiny@wias-berlin.de