Weierstraß–Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations

Michael H. Neumann

submitted: 12th December 1996

Weierstrass Institute
for Applied Analysis
and Stochastics
Mohrenstraße 39
D – 10117 Berlin
Germany

Preprint No. 295
Berlin 1996

1991 Mathematics Subject Classification. Primary 62G07; secondary 62G09, 62M07.

Key words and phrases. Density estimation, strong approximation, bootstrap, weak dependence, mixing, whitening by windowing, simultaneous confidence bands, nonparametric tests.

Edited by
Weierstraß–Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2044975
e–mail (X.400): c=de;a=d400-gw;p=WIAS-BERLIN;s=preprint
e–mail (Internet): preprint@wias-berlin.de

ABSTRACT. We derive a useful approximation of a density estimator based on weakly dependent random vectors by a density estimator built from independent random vectors. We construct, on a sufficiently rich probability space, such a pairing of the random variables of both experiments that the set of observations $\{X_1, \ldots, X_n\}$ from the time series model is nearly the same as the set of observations $\{Y_1, \ldots, Y_n\}$ from the i.i.d. model. The set $(\{X_1, \ldots, X_n\} \Delta \{Y_1, \ldots, Y_n\}) \cap ([a_1, b_1] \times \ldots \times [a_d, b_d])$ has with a high probability at most $O(\{[n^{1/2} \prod (b_i - a_i)] + 1\} \log(n))$ elements. Although this does not imply very much for parametric problems, it has important implications in nonparametric statistics. It yields a strong approximation of a kernel estimator of the stationary density by a kernel density estimator in the i.i.d. model. Moreover, we show that such a strong approximation is also valid for the standard bootstrap and the smoothed bootstrap. Using these results we derive simultaneous confidence bands as well as supremum-type nonparametric tests based on reasoning for the i.i.d. model.

1. INTRODUCTION

Density estimation on the basis of i.i.d. observations is one of the most often studied problems in nonparametric statistics. Important asymptotic properties concerning the pointwise as well as the joint probabilistic behaviour of commonly used estimators are now well-known and allow for powerful methods of statistical inference like, for example, tests for certain hypotheses or simultaneous confidence bands which guarantee asymptotically the desired error probability of the first kind and coverage probability, respectively.

In contrast, much less is known in the case of dependent observations. This case is very important from the practical point of view, since data from time series usually show some dependence. In order to develop analogous tools as in the independent case, it seems to be on first sight unavoidable to account for the dependence by specific corrections. This might, however, turn out to be quite a difficult and messy task, such that one could be tempted to seek for conditions which ensure asymptotically the same behaviour of certain statistics as known from the i.i.d. setting.

Whereas long-range dependence usually leads to phenomena essentially different from those under independence, there seems to be some hope for asymptotic similarities to the independent case under short-range dependence. Some commonly imposed conditions for weak dependence are strong (α-) mixing and absolute regularity (β-mixing). Provided the corresponding mixing coefficients decay fast enough, then commonly used nonparametric estimators converge with the same *rates* as in the independent case; cf. Györfi, Härdle, Sarda and Vieu (1989). The fact that desirable properties of the estimators remain valid in the dependent case provides a strong motivation for applying just the same estimation techniques as under the assumption of independence. However, some important tools for statistical inference require a more accurate knowledge of the asymptotic properties of the underlying estimators. Assuming mixing and some additional, not very restrictive condition on the boundedness of the joint densities of consecutive random variables, Robinson (1983), Masry (1994) and Hart (1995) showed that certain nonparametric estimators have actually

the same asymptotic variance as in the independent case. This phenomenon, which was described as "whitening by windowing" by Hart, is in sharp contrast to what happens in (finite-dimensional) parametric problems. For example, the asymptotic variance of the mean of time-series data does of course depend on the covariances as well. Results like those of Robinson (1983), Masry (1994) and Hart (1995) on the *pointwise* behaviour of nonparametric estimators allow, for example, to neglect the dependence structure when one establishes pointwise confidence intervals for the density function.

On the other hand, other problems of statistical inference require an even stronger notion of asymptotic equivalence. For example, the construction of simultaneous confidence bands or the determination of critical values for certain tests against a nonparametric alternative require knowledge about the *joint* distribution of the nonparametric estimator used to define the corresponding statistic. A first step in this direction has been done by Neumann and Kreiss (1996). They characterized the asymptotic equivalence of nonparametric autoregression and nonparametric regression by a strong approximation of a local polynomial estimator of the autoregression function by a local polynomial estimator in an appropriate regression setup. However, the nonparametric autoregressive model automatically imposes certain *structural* conditions on the data-generating process, which were essential for the approximation method used. Since this restricts the applicability of such a method in practice, it would be very desirable to develop similar results without any such structural assumptions.

In the present paper we show quite a surprising similarity between the observations that stem from a time-series model and a set of independent observations. Let $X_1, \ldots, X_n$ be d-dimensional, weakly dependent random vectors with a stationary density f. As a counterpart we consider i.i.d. random vectors $Y_1, \ldots, Y_n$ with the same density f. Let $\Delta_n = \{X_1, \ldots, X_n\} \Delta \{Y_1, \ldots, Y_n\}$ be the symmetric difference of both sets of observations. We show that there exists, on a sufficiently rich probability space, a pairing of the random variables of both models, which preserves the respective joint distributions, such that the following fact is true. Let $[a, b] = [a_1, b_1] \times \ldots \times [a_d, b_d]$ be an arbitrary hyperrectangle. Then the relation

$$\#(\Delta_n \cap [a,b]) = O\left(\left\{\left[n^{1/2} \prod_{i=1}^{d} (b_i - a_i)\right] + 1\right\} \log(n)\right)$$

is satisfied with a probability exceeding $1 - O(n^{-\lambda})$, where $\lambda < \infty$ is an arbitrarily large constant. The link is achieved by embedding both the random variables from the time series model and the i.i.d. model into a common Poisson process on $(0, \infty) \times \mathbb{R}^d$. It turns out that most of the randomness of the kernel estimators in both models is driven by the same part of the Poisson process. This leads to an approximation of the kernel estimators in the time series model by the kernel estimator in the i.i.d. model with an error that is of lower order of magnitude than the noise level of one of these estimators.

Let $\widehat{f}_h(x) = (nh^d)^{-1} \sum_{i=1}^{n} K((x - X_i)/h))$ and $\widetilde{f}_h(x) = (nh^d)^{-1} \sum_{i=1}^{n} K((x - Y_i)/h))$ be

kernel estimators of $f(x)$, where K is a compactly supported kernel function. Then we see that $\#\{\Delta_n \cap supp(K((x-.)/h))\} = O(n^{1/2}h^d \log(n))$, and, therefore,

$$\widehat{f}_h(x) - \widetilde{f}_h(x) \;=\; O\left(n^{-1/2}\log(n)\right)$$

are satisfied with a large probability. Such a result can be shown to be valid in a uniform manner for $x \in \mathbb{R}^d$. In view of the fact that $\sup_x\{var(\widehat{f}_h(x))\} \asymp (nh^d)^{-1}$, we have a useful strong approximation of the kernel estimator $\{\widehat{f}_h(x)\}_{x\in\mathbb{R}^d}$ by $\{\widetilde{f}_h(x)\}_{x\in\mathbb{R}^d}$. As some interesting applications we establish simultaneous confidence bands for f as well as certain tests based on the supremum norm between the above kernel estimator $\widehat{f}_h$ and estimators corresponding to hypotheses of lower-dimensional parametric or semiparametric structures. To determine the required tuning parameters, that is the width of the bands and the critical value for the test, respectively, we propose two bootstrap methods, both developed under the assumption of independence.

2. The approximation scheme

The main goal in this section is to establish a link between density estimation under weak dependence and density estimation based on independent observations. This will be achieved in a mainly constructive way, by embedding the random variables of both models in a common Poisson process indexed by time as well as spatial position in $\mathbb{R}^d$. The apparently quite involved problem of finding a global (in x) connection between kernel estimators $\widehat{f}_h(x)$ and $\widetilde{f}_h(x)$ in these models will be reduced to a collection of one-dimensional problems, which can be analysed separately from each other. Hence, in contrast to many other papers on strong approximations, the pleasant fact with our approximation method is that the technical part of the calculations becomes quite elementary.

2.1. The model and basic assumptions. Assume we have d-dimensional realizations $X_1, \ldots, X_n$ of a weakly dependent, stationary and time-homogeneous process with a stationary density f. To obtain some kind of asymptotic equivalence to the case of i.i.d. random variables, we impose the following conditions:

Assumption 1
$\{X_i\}$ is absolutely regular (that is, β-mixing) with mixing coefficients satisfying

$$\beta_k \;\le\; C\exp(-C_1 k).$$

Assumption 2
Let $f_{X_i|\mathcal{F}_{i-1}}$ be the density of the conditional distribution $\mathcal{L}(X_i \mid X_{i-1}, \ldots, X_1)$ and let $f_{X_i|X_{i-1},\ldots,X_{i-\gamma}}$ be the density of $\mathcal{L}(X_i \mid X_{i-1}, \ldots, X_{i-\gamma})$. We assume that

$$\sup_i \sup_{x\in\mathbb{R}^d} \left\{\left|f_{X_i|\mathcal{F}_{i-1}}(x) \;-\; f_{X_i|X_{i-1},\ldots,X_{i-\gamma}}(x)\right|\right\} \;=\; O\left(\exp(-C_2\gamma)\right)$$

and

$$\sup_i \sup_{x\in\mathbb{R}^d} \left\{f_{X_i|\mathcal{F}_{i-1}}(x)\right\} \;\le\; C_3$$

hold for some $C_2 > 0$ and $C_3 < \infty$.

Remark 1.

(i) Our assumption of exponentially decaying mixing coefficients is rather strong and can possibly be relaxed on the expense of a larger error in our approximation. Nevertheless, it is known that still many interesting processes are actually exponentially β-mixing. Mokkadem (1990, Theorem 2.1) provides sufficient conditions for a Markov chain to be geometrically β-mixing. Ango Nze (1992) used this result to derive sufficient conditions for a vector autoregressive process with conditional heteroscedasticity given as

$$X_{t+1} = m(X_t) + g(X_t)\varepsilon_{t+1},$$

ε_t i.i.d., to be geometrically ergodic, which implies geometrical β-mixing if the chain is stationary. An overview is given by Doukhan (1994).

(ii) Assumption 1 may be seen as some kind of minimal condition which brings the time series model close to an i.i.d. situation. This is however not enough to get the desired asymptotic equivalence. We need some additional condition which ensures that closely neighbored (in time) observations do not behave too different from an i.i.d. situation. Whereas Robinson (1983), Masry (1994) and Hart (1995) imposed a condition on the boundedness of the joint densities, we set this slightly stronger Assumption 2, which also reflects a rapidly decaying memory of the process $\{X_i\}$.

2.2. Embedding the random variables into a common Poisson process. Now we relate the random vectors $X_1, \ldots, X_n$ from the above setup to i.i.d. random vectors $Y_1, \ldots, Y_n$ having a density f. For that, we define on a sufficiently rich probability space copies $X_1^*, \ldots, X_n^*$ and $Y_1^*, \ldots, Y_n^*$ with the same joint distribution as $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively. As the connecting device, which determines both $X_1^*, \ldots, X_n^*$ and $Y_1^*, \ldots, Y_n^*$, we use a Poisson process N on $(0, \infty) \times \mathbb{R}^d$ with an intensity function equal to the Lebesgue measure. For details concerning the definition and construction of N, see Reiss (1993, Section 2.1). In contrast to Reiss, we use the equivalent formulation of a set-valued process instead of a point measure-valued process. Furthermore, since it is unlikely that this causes any confusion, we do not distinguish between X_i and X_i^* as well as Y_i and Y_i^* and denote the versions of these random variables on the common probability space simply by X_i and Y_i, respectively. The notation X_i^* will be reserved for bootstrap resamples of the X_i's to be introduced in Section 3.

First we describe in detail how the Poisson process N is used to generate the observations $X_1, \ldots, X_n$, retaining the joint distribution of these random vectors. The embedding of $Y_1, \ldots, Y_n$ is completely analogous, since independence is a special case of weak dependence.

(i) *Embedding of* X_1

Let $\{(U_j, V_j), j = 1, 2, \ldots\}$ denote a realization of N, where $U_j \in (0, \infty)$ and $V_j \in \mathbb{R}^d$. The basic idea of how X_1 is represented by $N^{(1)} = N$ may be explained as follows:

consider the graph $(tf_{X_1}(v), v)$ of the function $g_t(v) = tf_{X_1}(v)$, which spreads out, starting from $\{0\} \times \mathbb{R}^d$, with a velocity proportional to $f_{X_1}(v)$. We define

$$X_1 = V_{j_1},$$

where (U_{j_1}, V_{j_1}) is the first realization of $N^{(1)}$ hitted by $(tf_{X_1}(v), v)$ as t grows from zero to infinity. In other words, we have

$$j_1 = \arg\inf\{U_j / f_{X_1}(V_j)\}.$$

Note that $\{(U_j/f_{X_1}(V_j), V_j), j = 1, 2, \dots\}$ is a Poisson process on $(0, \infty) \times \mathbb{R}^d$ with intensity function $p(u, v) = f_{X_1}(v)$. Hence, it is clear that X_1 has just the desired density $f_{X_1} = f$.

To explain the following steps in a formally correct way, we introduce stopping times $\tau_v^{(i)}$, $i = 0, \dots, n$. Define

$$\tau_v^{(0)} \equiv 0$$

and

$$\tau_v^{(1)} = \tau_v^{(0)} + [U_{j_1}/f_{X_1}(V_{j_1})] f_{X_1}(v).$$

If we order $\{(U_j/f_{X_1}(V_j), V_j), j = 1, 2, \dots\}$ with respect to the first component, we may alternatively construe this object as a marked Poisson point process where the second argument has the density f_{X_1}. If we denote the corresponding realizations of this process by (S_j, W_j), $S_1 < S_2 < \dots$, then X_1 is just equal to W_1. By the strong Markov property of a marked Poisson point process, the remaining part of N,

$$N^{(2)} = \left\{(U_j - \tau_{V_j}^{(1)}, V_j)\right\} \cap \left((0, \infty) \times \mathbb{R}^d\right),$$

is again a Poisson process on $(0, \infty) \times \mathbb{R}^d$.

(ii) *Embedding of X_i*

Assume that $X_1, \dots, X_{i-1}$ have already been embedded into N, according to their conditional distributions $\mathcal{L}(X_k \mid X_{k-1}, \dots, X_1\}$. We embed X_i in the remaining part of N, that is

$$N^{(i)} = \left\{(U_j - \tau_{V_j}^{(i-1)}, V_j)\right\} \cap \left((0, \infty) \times \mathbb{R}^d\right).$$

In other words, we use from the whole set of realizations $\{(U_j, V_j)\}$ of N only those from the subset $\{(U_j, V_j) \mid U_j > \tau_{V_j}^{(i-1)}\}$. By the strong Markov property of the corresponding marked Poisson point process, $N^{(i)}$ is again a Poisson process on $(0, \infty) \times \mathbb{R}^d$. Now we define

$$X_i = V_{j_i},$$

where

$$j_i = \arg\inf\left\{(U_j - \tau_{V_j}^{(i-1)})/f_{X_i|\mathcal{F}_{i-1}}(V_j), \quad U_j > \tau_{V_j}^{(i-1)}\right\}.$$

Further, we set

$$\tau_v^{(i)} = \tau_v^{(i-1)} + [(U_{j_i} - \tau_{V_{j_i}}^{(i-1)})/f_{X_i|\mathcal{F}_{i-1}}(V_{j_i})] f_{X_i|\mathcal{F}_{i-1}}(v).$$

Finally, we obtain that

$$\{X_1, \ldots, X_n\} = \{V_j \mid \quad U_j \leq \tau_{V_j}^{(n)}\}. \tag{2.1}$$

(iii) *Embedding of* $Y_1, \ldots, Y_n$

The embedding of $Y_1, \ldots, Y_n$ is completely analogous to that of $X_1, \ldots, X_n$. Since $\mathcal{L}(Y_i \mid \ Y_{i-1}, \ldots, Y_1) = \mathcal{L}(Y_1)$, we have to deform the time axis only once.
Let $\{(\widetilde{T}_j, \widetilde{W}_j), j = 1, 2, \ldots\}$ be the marked Poisson point process corresponding to $\{(U_j/f(V_j), V_j), j = 1, 2, \ldots\}$. That is, we have in particular $\widetilde{T}_1 < \widetilde{T}_2 < \ldots$. Then we define

$$Y_i = \widetilde{W}_i, \quad i = 1, \ldots, n.$$

We may introduce stopping times $\tilde{\tau}_v^{(i)}$ analogous to the $\tau_v^{(i)}$'s. We obtain $\tilde{\tau}_v^{(n)} = \widetilde{T}_n f(v)$, which implies that

$$\{Y_1, \ldots, Y_n\} = \{V_j \mid \quad U_j \leq \tilde{\tau}_{V_j}^{(n)}\}. \tag{2.2}$$

Remark 2. It may well happen that the X_i's emerge in a different chronological order than the Y_i's. Since the transition densities are usually different from the stationary density, the construction for the time-series model "borrows" some probability mass assigned to future time points in the i.i.d. model. This is just the reason why we introduce a "time axis" for our embedding method.

2.3. Approximation results. To get estimates for the number of elements of Δ_n that fall in certain intervals, we derive first an estimate for the distance between $\tau_v^{(n)}$ and $\tilde{\tau}_v^{(n)}$, respectively, and their common expectation $nf(v)$.
Since many assertions in this article are of the type that a certain random variable is below some threshold with a high probability, we introduce the following notation.

Definition 2.1. Let $\{Z_n\}$ be a sequence of random variables and let $\{\alpha_n\}$ and $\{\gamma_n\}$ be sequences of positive reals. We write

$$Z_n = \tilde{O}(\alpha_n, \gamma_n),$$

if

$$P(|Z_n| > C\alpha_n) \leq C\gamma_n$$

holds for $n \geq 1$ and some $C < \infty$.

This definition is obviously stronger than the usual O_P and it is well suited for our particular purposes of constructing confidence bands and nonparametric tests; see its application in Section 3.
Further, we make throughout the paper the convention that $\delta > 0$ will denote an arbitrarily small and $\lambda < \infty$ an arbitrarily large constant.

Lemma 2.1. *Suppose that Assumptions 1 and 2 hold. Then*

$$|\tau_v^{(n)} - nf(v)| + |\tilde{\tau}_v^{(n)} - nf(v)| = \tilde{O}\left(n^{1/2}\log(n), n^{-\lambda}\right).$$

Whereas the pointwise (in v) similar behavior of $\tau_v^{(n)}$ and $\tilde{\tau}_v^{(n)}$ does not imply anything essential, a uniform version of the result given in Lemma 2.1 will finally yield the desired result about the difference set Δ_n. To derive such a uniform version, we impose the following smoothness condition on the conditional densities:

Assumption 3
There exists some constant $C < \infty$ such that

$$\sup_i \left| f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|\mathcal{F}_{i-1}}(v') \right| \leq C\|v - v'\|.$$

Lemma 2.2. *Suppose that Assumptions 1 through 3 are fulfilled. Then we have for any hyperrectangle* $[a,b] = [a_1,b_1] \times \ldots \times [a_d,b_d]$ *that*

$$\sup_{v \in [a,b]} \left\{ |\tau_v^{(n)} - nf(v)| + |\tilde{\tau}_v^{(n)} - nf(v)| \right\} = \tilde{O}\left(n^{1/2}\log(n), n^{-\lambda}\right).$$

Now we are in a position to relate both experiments to a common experiment given by the restriction of N to

$$S_n = \left\{ (u,v) \middle| \quad 0 < u \leq nf(v), \ \ v \in \mathbb{R}^d \right\}.$$

Let

$$\{Z_1, \ldots, Z_\nu\} = \{V_j \mid \ U_j \leq nf(V_j)\}. \tag{2.3}$$

Now we obtain estimates for the cardinality of the sets $(\{X_1,\ldots,X_n\}\Delta\{Z_1,\ldots,Z_\nu\}) \cap [a,b]$ as well as $(\{Y_1,\ldots,Y_n\}\Delta\{Z_1,\ldots,Z_\nu\}) \cap [a,b]$ from Lemma 2.2 and an appropriate exponential inequality for Poisson processes.

Proposition 2.1. *Suppose that Assumptions 1 through 3 hold. Then*

$$\left.\begin{array}{l} \#\{(\{X_1,\ldots,X_n\}\Delta\{Z_1,\ldots,Z_\nu\}) \cap [a,b]\} \\ \#\{(\{Y_1,\ldots,Y_n\}\Delta\{Z_1,\ldots,Z_\nu\}) \cap [a,b]\} \end{array}\right\} = \tilde{O}\left(\{[n^{1/2}\prod(b_i - a_i)] + 1\}\log(n), n^{-\lambda}\right).$$

Now we obtain, as an immediate consequence of Proposition 2.1, the desired strong approximation of a kernel estimator $\hat{f}_h$ in the time series model by a kernel estimator $\tilde{f}_h$ in the i.i.d. model. Let

$$\hat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

and

$$\tilde{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x - Y_i}{h}\right).$$

For simplicity we impose the following condition:

Assumption 4
The kernel K is supported on $[-1,1]^d$ and $\sup_x\{|K(x)|\} \leq K_0$.

It is obvious that

$$\begin{aligned}
&|\widehat{f}_h(x) - \widetilde{f}_h(x)| \\
&\leq \frac{K_0}{nh^d}\#\{(\{X_1,,\ldots,X_n\}\Delta\{Y_1,\ldots,Y_n\}) \cap ([x_1-h,x_1+h]\times\ldots\times[x_d-h,x_d+h])\} \\
&= \tilde{O}\left([n^{-1/2}+(nh^d)^{-1}]\log(n), n^{-\lambda}\right) \qquad (2.4)
\end{aligned}$$

holds for arbitrary $x \in \mathbb{R}^d$. With a simple extra argument, we can show that this result holds simultaneously over $\mathbb{R}^d$.

Theorem 2.1. *Suppose that Assumptions 1 through 4 are fulfilled. Then*

$$\sup_{x\in\mathbb{R}^d}\left\{|\widehat{f}_h(x) - \widetilde{f}_h(x)|\right\} = \tilde{O}\left([n^{-1/2}+(nh^d)^{-1}]\log(n), n^{-\lambda}\right).$$

Now it becomes clear what we have achieved by our embedding of $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$ in a common Poisson process: the apparently quite difficult task of getting a *uniform* (in x) approximation of $\widehat{f}_h(x)$ by $\widetilde{f}_h(x)$ is reduced to the technically much simpler task of proving a *pointwise* result as in Lemma 2.1.

3. Application to simultaneous confidence bands and nonparametric tests

Theorem 2.1 in the previous section provides an approximation of a kernel estimator in the time series model by a kernel estimator in an i.i.d. model. Besides the more fundamental message that weak dependence is asymptotically negligible, the practical significance lies on the possibility to transfer methods of inference originally developed under the assumption of independence to the case of weakly dependent random variables. As two important applications, we propose in this section confidence bands and supremum-type tests based on a bootstrap approximation of the distribution of the L_∞-distance between $\widehat{f}_h$ and $E\widehat{f}_h$. We did not attempt to develop versions of these methods based on asymptotic theory instead of the bootstrap. Although, at least in the one-dimensional case, the process $\{(\widehat{f}_h(x) - E\widehat{f}_h(x))/\sqrt{var(\widehat{f}_h(x))}\}_{x\in[a,b]}$ can be well approximated by a Gaussian process, the approximation of the supremum of this Gaussian process by its limit, as proposed by Bickel and Rosenblatt (1973), converges with the very slow rate $(\log(n))^{-1}$; cf. Hall (1991). In contrast, it will be shown that the bootstrap approximation converges with a certain algebraic rate.

3.1. Two bootstrap proposals. We consider two methods of bootstrapping the empirical process, the standard bootstrap and the smoothed bootstrap. Both versions were proposed by Efron (1979) in the context of i.i.d. observations.
Denote by P_n the empirical distribution based on $\{X_1, \ldots, X_n\}$. In the standard bootstrap, we draw with replacement n independent bootstrap resamples $X_1^*, \ldots, X_n^*$. That is, the unknown distribution P is replaced by its empirical analog P_n. In the smoothed bootstrap, we draw n independent bootstrap resamples $X_1^{*,g}, \ldots, X_n^{*,g}$ from a smoothed version $P_{n,g}$ of P_n. $P_{n,g}$ is the distribution function which corresponds to the kernel estimate

$$\hat{f}_g(x) = \frac{1}{ng^d} \sum_{i=1}^{n} L\left(\frac{x - X_i^*}{g}\right)$$

of $f(x)$. We use the letters L and g to indicate that one may use a kernel and a bandwidth different from K and h, respectively. It will turn out that there is very much freedom for the choice of g.
A discussion about the relative merits of the standard bootstrap and the smoothed bootstrap as well as some examples may be found in Efron (1979, 1982), Silverman and Young (1987) and Hall (1992). Roughly speaking, smoothing does not improve the convergence rate of the bootstrap estimate, if that estimate can be expressed as (or well approximated by) a smooth function of a vector sample mean. In other cases, like in estimating the mean squared error of a quantile estimate, the smoothed bootstrap can significantly outperform the unsmoothed one; cf. Hall (1992, Appendix IV).
The derivation of asymptotic properties of the bootstrap methods goes again via strong approximations. We begin with the smoothed bootstrap and construct a pairing of $(Y_1, \ldots, Y_n)$ and $(X_1^{*,g}, \ldots, X_n^{*,g})$, which are both vectors of i.i.d. random variables, as follows. First we draw n independent Bernoulli random variables $B_i \sim \mathit{Bernoulli}(p)$, where $p = \int (f(x) \wedge \hat{f}_g(x))\, dx$. If $B_i = 1$, then we generate Y_i according to the density $(f(x) \wedge \hat{f}_g(x))/p$, and set $X_i^{*,g} = Y_i$. If $B_i = 0$, then we draw independently Y_i according to the density $[f(x) - (f(x) \wedge \hat{f}_g(x))]/(1-p)$ and $X_i^{*,g}$ with the density $[\hat{f}_g(x) - (f(x) \wedge \hat{f}_g(x))]/(1-p)$. It is easy to see that $Y_1, \ldots, Y_n$ are i.i.d. with density f and $X_1^{*,g}, \ldots, X_n^{*,g}$ are i.i.d. with density $\hat{f}_g$. The following assertion shows that this construction actually leads to a useful approximation of $\{\hat{f}_h(x) - E\hat{f}_h(x)\}_{x \in \mathbb{R}^d}$ by $\{\hat{f}_h^{*,g}(x) - E\hat{f}_h^{*,g}(x)\}_{x \in \mathbb{R}^d}$, where

$$\hat{f}_h^{*,g}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i^{*,g}}{h}\right).$$

Since the proofs of the assertions of this section use approximations of the kernel estimators on fine grids, we impose the following additional conditions:

Assumption 5
The kernel K is Lipschitz continuous and of second order.
Assumption 6
The kernel L is Lipschitz continuous and of second order.

Theorem 3.1. *Suppose that Assumptions 1 through 6 are fulfilled. Let*

$$\mu_n = g^2 + (ng^d)^{-1/2}\sqrt{\log(n)}.$$

Then there exists a pairing of the random variables $X_1, \dots, X_n$ and $X_1^{*,g}, \dots, X_n^{*,g}$ such that

$$\sup_{x\in\mathbb{R}^d}\left\{\left|(\widehat{f}_h(x) - E\widehat{f}_h(x)) - (\widehat{f}_h^{*,g}(x) - E\widehat{f}_h^{*,g}(x))\right|\right\}$$
$$= \tilde{O}\left(n^{-1/2}\log(n) + (nh^d)^{-1}\log(n) + (nh^d)^{-1/2}\mu_n^{1/2}\sqrt{\log(n)}, n^{-\lambda}\right).$$

In contrast to the case of the smoothed bootstrap, the distributions P and P_n are actually orthogonal. Hence, there is no hope to find such a pairing of both experiments that enough random variables from them coincide. However, obviously one can define a pairing of $(X_1^{*,g}, \dots, X_n^{*,g})$ and $(X_1^*, \dots, X_n^*)$ such that $\|X_i^{*,g} - X_i^*\| \le \sqrt{d}g$ for all i. Hence, for $g \ll h$, $\widehat{f}_h^{*,0}(x)$ is well approximated by $\widehat{f}_h^{*,g}(x)$, which finally provides the desired strong approximation of $\{\widehat{f}_h(x) - E\widehat{f}_h(x)\}_{x\in\mathbb{R}^d}$ by $\{\widehat{f}_h^{*,0}(x) - E\widehat{f}_h^{*,0}(x)\}_{x\in\mathbb{R}^d}$.

Theorem 3.2. *Suppose that Assumptions 1 through 5 are fulfilled. Then there exists a pairing of the random variables $X_1, \dots, X_n$ and $X_1^{*,0}, \dots, X_n^{*,0}$ such that*

$$\sup_{x\in\mathbb{R}^d}\left\{\left|(\widehat{f}_h(x) - E\widehat{f}_h(x)) - (\widehat{f}_h^{*,0}(x) - E\widehat{f}_h^{*,0}(x))\right|\right\}$$
$$= \tilde{O}\left(n^{-1/2}\log(n) + (nh^d)^{-1}\log(n) + (nh^d)^{-1/2}\sqrt{\log(n)}\inf_g\{(ng^d)^{-1/2}\sqrt{\log(n)} + g/h\}, n^{-}\right.$$

In order to assess the significance of the above strong approximation results for the desired approximation of the distribution of the maximal deviation of $\widehat{f}_h$ from its expectation, we still need an upper bound for the probabilities that this supremum falls into small intervals.

Proposition 3.1. *Suppose that Assumptions 1 through 5 are fulfilled.*

$$P\left(\sup_{x\in\mathbb{R}^d}\left\{|\widehat{f}_h(x) - E\widehat{f}_h(x)|\right\} \in [c,d]\right)$$
$$= O\left((d-c)(nh^d)^{1/2}(\log(n))^{1/2} + h\log(n) + (nh^d)^{-1/4}(\log(n))^{5/4} + \right.$$
$$\left. + h^{d/2}(\log(n))^{3/2} + (nh^d)^{-1/2}(\log(n))^{3/2}\right).$$

This estimate will finally imply, in conjunction with Theorems 2.1, 3.1 and 3.2, the validity of the bootstrap for the supremum functional. We apply this to the construction of simultaneous confidence bands and nonparametric tests in the next subsections.

3.2. Simultaneous confidence bands. Confidence bands are an important universal tool which provide some impression about the exactness of a nonparametric estimator. Similarly to nonparametric tests, they can indicate whether there is empirical evidence for certain conjectured features of the curve.

There already exists a considerable amount of literature on the construction of confidence bands in the context of independent observations. Work on simultaneous confidence bands in nonparametric density estimation dates back to the seminal paper by Bickel and Rosenblatt (1973) who used a first-order asymptotic approximation of the distribution of the supremum of a certain Gaussian process that approximates the deviation of the kernel estimator from its mean. The use of the bootstrap to determine an appropriate width for confidence bands for a univariate density was proposed by Faraway and Jhun (1990) on a heuristic level and investigated in more detail by Hall (1993). One of the main messages in Hall (1991, 1993) is that the application of the bootstrap leads to much smaller errors in coverage probability than the approach of Bickel and Rosenblatt (1973).

In contrast to the papers mentioned above, we consider confidence bands of uniform size rather than bands with a varying size, proportional to $\left(\widehat{var}(\hat{f}_h(x))\right)^{1/2}$. The latter bands seem to be somewhat more natural and they work well as long as they are restricted to some compact set on which the density f is bounded away from zero. One has to exclude regions of sparse design, because the performance of the bootstrap approximation deteriorates there. Such a truncation is not necessary with uniform bands, because then the problematic regions are automatically faded out.

Let t^*_α be the $(1-\alpha)$-quantile of the distribution of $\sup\{|\hat{f}^{*,g}_h(x) - E\hat{f}^{*,g}_h(x)|\}$, that is

$$P\left(\sup_{x\in\mathbb{R}^d}\left\{|\hat{f}^{*,g}_h(x) - E\hat{f}^{*,g}_h(x)|\right\} > t^*_\alpha \,\middle|\, X_1,\ldots,X_n\right) = \alpha. \tag{3.1}$$

For simplicity, we restrict the following considerations to the smoothed bootstrap. Using Theorem 3.2 instead of Theorem 3.1, one may derive results similar to the following theorems for t^*_α based on the standard bootstrap.

Let K_h be the smoothing operator defined by

$$K_h(f) = \int \frac{1}{h^d} K\left(\frac{x-z}{h}\right) f(z)\,dz. \tag{3.2}$$

Although statisticians usually focus on confidence intervals or bands for the density itself, we consider first simultaneous confidence bands for $K_h(f)$. The reason is that this problem is much easier to deal with, and with bands for $K_h(f)$ we have also more freedom to choose h. Theorems 2.1 and 3.2 and Proposition 3.1 imply the following theorem:

Theorem 3.3. *Suppose that Assumptions 1 through 6 are fulfilled. Then*

$$P\left(K_h(f)(x) \in [\widehat{f}_h(x) - t^*_\alpha, \widehat{f}_h(x) + t^*_\alpha] \quad \textit{for all} \quad x \in \mathbb{R}^d\right)$$
$$= 1 - \alpha + O\left(h^{d/2}(\log(n))^{3/2} + (nh^d)^{-1/2}(\log(n))^{3/2} + \mu_n^{1/2}\log(n) + \right.$$
$$\left. + h\log(n) + (nh^d)^{-1/4}(\log(n))^{5/4}\right).$$

If

$$h = o\left((\log(n))^{-(3\vee d)}\right), \tag{3.3}$$

$$(nh^d)^{-1} = o\left((\log(n))^{-5}\right) \tag{3.4}$$

and

$$\mu_n = o\left((\log(n))^{-2}\right), \tag{3.5}$$

then the confidence band will have asymptotically the prescribed coverage probability for $K_h(f)$. Certain *qualitative* features of f like unimodality or monotonicity in some region remain valid for the smoothed version $K_h(f)$ under mild regularity assumptions on the kernel K. Hence, the confidence band for $K_h(f)$ can also be used as a criterion to assess whether there is enough evidence for such a feature. This is, of course, closely related to the formal test proposed in Subsection 3.3.
Since density estimation is an ill-posed inverse problem, there are certain limitations for a *pointwise* inference about $f(x)$. For example, one cannot consistently distinguish between two densities that differ only on an interval shrinking at a sufficiently fast rate. This is in some way reflected in the bias problem one necessarily encounters in the construction of confidence bands for f. Nevertheless, there seems to be considerable interest in such bands, because they provide an easily accessible quantitative characterization of the precision of a nonparametric estimator.
To determine the width of the confidence band, we will use again the $(1-\alpha)$-quantile t^*_α of the bootstrapped maximal deviation of the density estimator from its mean. We will obtain an asymptotically correct coverage probability, if the bias of $\widehat{f}_h$ is of smaller order of magnitude than its standard deviation. Our theory is valid for an undersmoothed estimator $\widehat{f}_h$, which excludes the usual mean-squared-error optimal choice of h.

Theorem 3.4. *Suppose that Assumptions 1 through 6 are fulfilled. Then*

$$P\left(f(x) \in [\widehat{f}_h(x) - t^*_\alpha, \widehat{f}_h(x) + t^*_\alpha] \quad \textit{for all} \quad x \in \mathbb{R}^d\right)$$
$$= 1 - \alpha + O\left(h^{d/2}(\log(n))^{3/2} + (nh^d)^{-1/2}(\log(n))^{3/2} + \mu_n^{1/2}\log(n) + \right.$$
$$\left. + h\log(n) + (nh^d)^{-1/4}(\log(n))^{5/4} + h^2(nh)^{1/2}(\log(n))^{1/2}\right).$$

We see from this theorem that the confidence band has asymptotically the desired coverage probability, if, besides (3.3), (3.4) and (3.5),

$$h^2 = o\left((nh^d)^{-1/2}(\log(n))^{-1/2}\right) \tag{3.6}$$

is satisfied. (3.6) means that we have to undersmooth in order to make the bias of $\widehat{f}_h$, which was not mimicked by the bootstrap, negligible. A well-known alternative consists in an explicit bias correction, which allows then also bandwidths $h = h_n$ decaying at the mean-squared-error optimal rate $n^{-1/(4+d)}$.
We do not dwell on the effect of a data-driven bandwidth choice which is important for a real application of this method. Usually data-driven bandwidths $\widehat{h}$ are intended to approximate a certain nonrandom bandwidth h_n. If $(\widehat{h} - h_n)/h_n$ converges at an appropriate rate, then the estimators $\widehat{f}_{\widehat{h}}$ and $\widehat{f}_{h_n}$ are sufficiently close to each other, such that the results obtained in this paper remain valid; see Neumann (1995) for a detailed investigation of these effects for pointwise confidence intervals in nonparametric regression.

3.3. A nonparametric test. Tests against a nonparametric alternative are an important tool to assess the appropriateness of a parametric or a semiparametric model. In contrast to tests like the Kolmogorov-Smirnov or the Cramér-von Mises test, our density-based test seems to be more powerful for local deviations from the assumed model. Moreover, by considering the supremum statistic, we exploit the whitening-by-windowing principle, which allows to neglect the dependence structure. We allow a composite hypothesis, that is

$$H_0: \quad f \in \mathcal{F},$$

where the only requirement is that the functional class $\mathcal{F}$ allows a faster rate of convergence than the full nonparametric model. We will assume

Assumption 7

There exists an estimator $\widehat{\widehat{f}}$ of f such that, for $f \in \mathcal{F}$,

$$\sup_{x \in \mathbb{R}^d} \left\{ \left| \int h^{-d} K\left(\frac{x-z}{h}\right) [\widehat{\widehat{f}}(z) - f(z)]\, dz \right| \right\} = o_P\left((nh^d)^{-1/2}(\log(n))^{-1/2}\right).$$

In the case $d = 1$, this includes some parametric models,

$$\mathcal{F} = \{f_\theta, \quad \theta \in \Theta\}.$$

In the higher-dimensional case, one may test for parametric but also for certain semiparametric models like, for example, a multiplicative nonparametric model that corresponds to the assumption that the components of the X_i's are independent,

$$\mathcal{F} = \left\{ f(x) = \prod_{i=1}^{d} f_i(x_i) \,\middle|\, \quad f_i \text{ "sufficiently smooth"} \right\},$$

or a semiparametric model proposed by Friedman, Stuetzle and Schroeder (1984),

$$\mathcal{F} = \left\{ f(x) = f_0(x) \prod_{i=1}^{M} f_i(\alpha_i' x) \,\middle|\, f_i \text{ "sufficiently smooth"} \right\}.$$

In accordance to our theory above, we consider the maximal deviation between $\widehat{f}_h$ and $K_h(\widehat{\widehat{f}})$, that is

$$T = \sup_{x \in \mathbb{R}^d} \left\{ \left| \widehat{f}_h(x) - \int h^{-d} K\left(\frac{x-z}{h}\right) \widehat{\widehat{f}}(z)\, dz \right| \right\}.$$

The next theorem shows that the prescribed error of the first kind is asymptotically guaranteed.

Theorem 3.5. *Suppose that Assumptions 1 through 7 as well as (3.3), (3.4) and (3.5) are fulfilled. Then*

$$P_{H_0}(T > t_\alpha^*) \longrightarrow \alpha \quad as \quad n \to \infty.$$

Remark 3. It seems that L_2-tests, like that proposed by Härdle and Mammen (1993) in the regression setup, are the most popular ones among nonparametric statisticians. Such tests can be optimal for testing against smooth alternatives, whereas supremum-type tests have less power in in such a situation. On the other hand, supremum-type tests can also outperform L_2-tests for testing against local alternatives having the form of sharp peaks; see Konakov, Läuter and Liero (1995) and Spokoiny (1996) for more details.
Our methodology is obviously restricted to supremum-type tests. The author conjectures that weak dependence as considered in the present paper cannot be neglected for L_2-test statistics.

4. Discussion

1) *A knock out for traditional mixing conditions?*

By now strong mixing and absolute regularity have been accepted as being benchmark conditions to characterize weak dependence. A lot of efforts have been devoted to show that estimation problems under weak dependence allow the same *rates* of convergence as under independence.
However, see see in this paper, as well as in Robinson (1983), Masry (1994) and Hart (1995), that suitable extra conditions on the joint densities lead to qualitatively much stronger results: then we obtain asymptotic equivalence on the level of *constants*. In many instances such an extra condition is not very restrictive and leads to an immediate applicability of important statistical methods developed under the assumption of independence. Hence, although the classical mixing conditions seem to be suitable in *parametric* (that is, finite-dimensional) estimation problems, we think

that this concept is perhaps not the most adequate one in *nonparametric* (that is, infinite-dimensional) problems of inference.

2) *Does a multiscale approach lead to a better approximation?*

In many cases one obtains better rates for strong approximations by a multiscale approach based on a dyadic partition of the interval of interest. A classical example is the construction by Komlós, Major and Tusnády (1975). A dyadic approximation scheme has also been employed by Neumann and Kreiss (1996) for constructing a strong approximation of nonparametric autoregression by nonparametric regression. The simultaneous consideration of different resolution scales makes sense for the above examples, because the *relative* approximation rate deteriorates as one moves to smaller intervals.

However, in our context, the possibility to approximate density estimators under weak dependence by density estimators under independence is essentially based on the "whitening by windowing"-principle. Therefore, the relative approximation rate becomes even better for finer scales. It seems to be unlikely that a multiscale approach leads to better approximation rates between kernel estimators from both models.

3) *Are these non-standard proofs really necessary?*

Compared to existing literature on similar topics, the proving methods in this paper are somehow non-standard. In particular, all proofs are based on certain constructive pairing techniques instead of the commonly used first-order approximation by the supremum of the limiting Gaussian process. This is done for the following two reasons: First, a purely analytical derivation of the asymptotic distribution of the maximal deviation between $\hat{f}_h$ and its expectation is presumably very technical and neither pleasant for the author nor for the reader. Second, it is well-known that first-order asymptotic theory leads to very poor rates of convergence in this context. Once we had used such an approximation at any point, we were not be able to prove that the bootstrap actually leads to better rates of convergence.

There exists an extensive literature on strong approximations for empirical cumulative distribution functions by certain Gaussian processes. For example, Dhompongsa (1984) showed for absolutely regular processes that the cumulative distribution function can be approximated by a Gaussian process with an error of order $n^{-1/2-\lambda}$, for a certain $\lambda > 0$. Such a result can also be used to show that a kernel density estimator is approximated by a certain Gaussian process. However, in dependence on the value of λ, there are limitations for the significance of such results. Kernel estimators with small bandwidths h will require more localized approximations.

4) *Alternative bootstrap methods*

Even if the effect of the dependence vanishes asymptotically, it is still present in higher order terms. Instead of neglecting it, one could also try to mimic the dependence structure by the bootstrap. One standard tool is the blockwise bootstrap introduced by Künsch (1989). Bühlmann (1994) showed that the blockwise bootstrap consistently estimates the distribution of a multivariate empirical process based on α-mixing observations, and applied this result to a nonlinear estimator of a finite-dimensional parameter. On the other hand, the blockwise bootstrap requires the estimation of

much more features of the data-generating process, which in turn leads to new fluctuations of the resulting estimates. It seems to be an important and challenging task to explore whether such an approach can really improve the rate of approximation.

5. Proofs

Proof of Lemma 2.1. Define

$$T_i = (U_{j_i} - \tau_{V_{j_i}}^{(i-1)})/f_{X_i|\mathcal{F}_{i-1}}(V_{j_i}).$$

We split up

$$\tau_v^{(n)} = \sum_{i=1}^{n} T_i f_{X_i|\mathcal{F}_{i-1}}(v) = nf(v) + R_1 + R_2, \tag{5.1}$$

where

$$\begin{aligned} R_1 &= \sum_{i=1}^{n} T_i \left[f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|X_{i-1},\dots,X_{i-\gamma_n}}(v) \right], \\ R_2 &= \sum_{i=1}^{n} \left[T_i f_{X_i|X_{i-1},\dots,X_{i-\gamma_n}}(v) - f(v) \right] \end{aligned}$$

and γ_n is chosen such that $\log(n)/(4C_2) \le \gamma_n < \log(n)/(4C_2) + 1$, C_2 given by Assumption 2.
It is easy to see that the vector $(T_1, \dots, T_n)$ is independent of $(X_1, \dots, X_n)$ and that $T_i \sim \text{Exp}(1)$ are i.i.d.
(To see this, consider for a moment the situation where we start with independent vectors $(\tilde{T}_1, \dots, \tilde{T}_n)$ and $(\widetilde{X}_1, \dots, \widetilde{X}_n)$, where $\tilde{T}_i \sim \text{Exp}(1)$ are i.i.d. and $\mathcal{L}(\widetilde{X}_i \mid \widetilde{X}_{i-1} = x_{i-1}, \dots, \widetilde{X}_1 = x_1) = \mathcal{L}(X_i \mid X_{i-1} = x_{i-1}, \dots, X_1 = x_1)$. Now we easily see that the conditional distributions $\mathcal{L}((T_i, X_i) \mid (T_{i-1}, X_{i-1}), \dots, (T_1, X_1))$ and $\mathcal{L}((\tilde{T}_i, \widetilde{X}_i) \mid (\tilde{T}_{i-1}, \widetilde{X}_{i-1}), \dots, (\tilde{T}_1, \widetilde{X}_1))$ coincide, which implies that $(T_1, \dots, T_n)$ and $(X_1, \dots, X_n)$ are actually independent.)
According to Assumption 2, we have that

$$\left| f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|X_{i-1},\dots,X_{i-\gamma_n}}(v) \right| = O(n^{-1/4}).$$

R_1 is a weighted sum of the T_i's, where $(T_1, \dots, T_n)$ is independent of the weights $([f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|X_{i-1},\dots,X_{i-\gamma_n}}(v)], i = 1, \dots, n)$. Hence, we obtain by Theorem 4 of Amosova (1972) that

$$P\left(|R_1| \ge \kappa \sqrt{\sum [f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|X_{i-1},\dots,X_{i-\gamma_n}}(v)]^2} \sqrt{\log(n)} \,\middle|\, X_1, \dots, X_n \right) = O(n^{-\kappa^2/2})$$

holds for arbitrary $\kappa < \infty$ and uniformly in $X_1, \dots, X_n$. This implies that

$$R_1 = \tilde{O}\left(n^{1/2} \sqrt{\log(n)}, n^{-\lambda} \right). \tag{5.2}$$

To estimate R_2, we consider blocks of observations $\{X_j, j \in J_i\}$, where $J_i = \{(i-1)\rho_n - \gamma_n + 1, \ldots, i\rho_n\}$ and $\rho_n \geq (\lambda+1)\log(n)/C_1 + \gamma_n - 1$, $\rho_n = O(\log(n))$. Without loss of generality, we consider the blocks with odd numbers. Note that we have

$$\beta\left(\sigma(\{X_j, j \in J_i\}), \sigma(\{X_j, j \in J_k\}, k = i+2, i+4, \ldots)\right) \leq C\exp(-C_1(\rho_n - \gamma_n + 1)).$$

By Proposition 2 of Doukhan, Massart and Rio (1995, page 407), there exists a sequence of independent blocks $\{\widetilde{X}_j, j \in J_i\}$, i odd, where the $\widetilde{X}_j$'s are independent of the T_j's, $\mathcal{L}((\widetilde{X}_j, j \in J_i)) = \mathcal{L}((X_j, j \in J_i))$, and

$$P\left((\widetilde{X}_j, j \in J_i) \neq (X_j, j \in J_i)\right) \leq C\exp(-C_1(\rho_n - \gamma_n + 1)) = O(n^{-\lambda-1}). \quad (5.3)$$

Now we have

$$var\left(\sum_{j=(i-1)\rho_n+1}^{i\rho_n} T_j f_{\widetilde{X}_j|\widetilde{X}_{j-1},\ldots,\widetilde{X}_{j-\gamma_n}}(v)\right) \leq \rho_n \sum_{j=(i-1)\rho_n+1}^{i\rho_n} var(T_j f_{\widetilde{X}_j|\widetilde{X}_{j-1},\ldots,\widetilde{X}_{j-\gamma_n}}(v)),$$

which implies, again by Theorem 4 of Amosova (1972), that

$$\begin{aligned}
&\sum_{i \text{ odd}} \sum_{j=(i-1)\rho_n+1}^{i\rho_n} \left(T_j f_{X_j|X_{j-1},\ldots,X_{j-\gamma_n}}(v) - f(v)\right) \\
&= \tilde{O}\left(\sqrt{\rho_n \sum_{i \text{ odd}} \sum_{j=(i-1)\rho_n+1}^{i\rho_n} var(T_j f_{X_j|X_{j-1},\ldots,X_{j-\gamma_n}}(v))}\sqrt{\log(n)}, n^{-\lambda}\right) \\
&= \tilde{O}\left(n^{1/2}\log(n), n^{-\lambda}\right).
\end{aligned}$$

An analogous result can be shown for the blocks with even numbers, which implies, in conjunction with (5.3), that

$$R_2 = \tilde{O}\left(n^{1/2}\log(n), n^{-\lambda}\right). \quad (5.4)$$

The proof of the assertion about $\tilde{\tau}_v^{(n)}$ is analogous, which finishes the proof. □

Proof of Lemma 2.2. We prove the assertion only for $\sup_{v\in[a,b]}\{|\tau_v^{(n)} - nf(v)|\}$. Let $\mathcal{N}_n = \{v_1, \ldots, v_{\gamma_n}\}$ be an $n^{-1/2}$-net for the hyperrectangle $[a, b]$ of cardinality $\#\mathcal{N}_n = \gamma_n = O(n^{d/2})$. It is clear from Lemma 2.1 that

$$\sup_{1\leq j\leq\gamma_n}\left\{\left|\tau_{v_j}^{(n)} - nf(v_j)\right|\right\} = \tilde{O}\left(n^{1/2}\log(n), n^{-\lambda}\right) \quad (5.5)$$

holds. Let $v \in [a, b]$ be arbitrary. Then there exists a $j(v) \in \{1, \ldots, \gamma_n\}$ such that $\|v - v_{j(v)}\| = O(n^{-1/2})$. Since

$$\sum_{i=1}^{n} T_i = O(n) + \tilde{O}\left(n^{1/2}\sqrt{\log(n)}, n^{-\lambda}\right),$$

we have that

$$|\tau_v^{(n)} - \tau_{v_{j(v)}}^{(n)}| \leq \sum_{i=1}^{n} T_i \left|f_{X_i|\mathcal{F}_{i-1}}(v) - f_{X_i|\mathcal{F}_{i-1}}(v_{j(v)})\right| = \tilde{O}(n^{1/2}, n^{-\lambda}), \quad (5.6)$$

which yields, in conjunction with $|f(v) - f(v_{j(v)})| = O(n^{-1/2})$, the assertion. □

Proof of Proposition 2.1. According to Lemma 2.2, we have that

$$\begin{aligned}&(\{X_1, \dots, X_n\} \Delta \{Z_1, \dots, Z_\nu\}) \cap [a,b] \\ &\quad \subseteq \left\{V_j \in [a,b] \,\middle|\, nf(v) - C_\lambda n^{1/2} \log(n) \le U_j \le nf(v) + C_\lambda n^{1/2} \log(n)\right\}\end{aligned}$$

holds with a probability exceeding $1 - O(n^{-\lambda})$, where C_λ is an appropriate constant. To get an estimate for the cardinality of the latter set, we apply an exponential inequality to the restriction N_D of the Poisson process N to

$$D = \left\{(u,v) \,\middle|\, nf(v) - C_\lambda n^{1/2} \log(n) \le u \le nf(v) + C_\lambda n^{1/2} \log(n),\ v \in [a,b]\right\}.$$

It is clear that N_D is a Poisson process with intensity $\mu(D) = O(n^{1/2} \log(n) \prod(b_i - a_i))$. If $\mu(D) \ge (8/3)\lambda \log(n)$, then we obtain by Inequality 14.5.1 on page 569, and Proposition 11.1.1(10) on page 441 in Shorack and Wellner (1986) that

$$P(N_D > 2\mu(D)) \le \exp\left(-\frac{\mu(D)}{2}\psi(1)\right) \le \exp\left(-\frac{\mu(D)}{2}\frac{3}{4}\right) = O(n^{-\lambda}). \quad (5.7)$$

If $\mu(D) < (8/3)\lambda \log(n)$, then we obtain, again by Inequality 14.5.1 and Proposition 11.1.1(10) of Shorack and Wellner (1986), that

$$\begin{aligned}&P\left(N_D - \mu(D) > \frac{8}{3}\lambda \log(n)\right) \\ &\le \exp\left(-\frac{((8/3)\lambda \log(n))^2}{2\mu(D)}\psi\left(\frac{(8/3)\lambda \log(n)}{\mu(D)}\right)\right) \\ &\le \exp\left(-\frac{((8/3)\lambda \log(n))^2}{2\mu(D)}\frac{3}{4}\frac{\mu(D)}{(8/3)\lambda \log(n)}\right) = O(n^{-\lambda}). \quad (5.8)\end{aligned}$$

(5.7) and (5.8) imply (i). (ii) follows from the same reasoning. □

Proof of Theorem 2.1. It remains to show that (2.4) holds simultaneously for all $x \in \mathbb{R}^d$. For that, we show that the assertion of Proposition 2.1 holds simultaneously over all hypercubes $I_k = [(k_1-1)h, k_1 h] \times \ldots \times [(k_d-1)h, k_d h]$, where $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$. Since $K((x - .)/h)$ is supported on a finite number of these I_k's, we will immediately get the assertion of the theorem from a relation like (2.4).
The main argument of this proof will be based on Bernstein's inequality, which we quote for reader's convenience from Shorack and Wellner (1986, p. 855):
Let $Z_1, \dots, Z_n$ be i.i.d. random variables with $EZ_1 = 0$ and $|Z_1| \le K_n$ almost surely. Then, for $Z = \sum Z_i$,

$$\begin{aligned}P(Z > c) &\le \exp\left(-\frac{c^2/2}{var(Z) + (K_n c)/3}\right) \\ &\le \exp\left(-\frac{c^2}{4\,var(Z)}\right) + \exp\left(-\frac{3c}{4K_n}\right)\end{aligned}$$

holds for arbitrary $c > 0$.

Setting

$$c_\lambda = \sqrt{var(Z)}\sqrt{4\lambda \log(n)} + (4/3)K_n \lambda \log(n)$$

we get

$$P(|Z| > c_\lambda) \leq 4\exp(-\lambda \log(n)).$$

In other words, we have that

$$Z = \tilde{O}\left(\sqrt{var(Z)}\sqrt{\log(n)} + K_n \log(n), n^{-\lambda}\right). \tag{5.9}$$

We consider two index sets:

$$\mathcal{K}_1 = \left\{k \mid \quad P(X_1 \in I_k) \geq n^{-\tau}\right\}$$

and

$$\mathcal{K}_2 = \left\{k \mid \quad P(X_1 \in I_k) < n^{-\tau}\right\},$$

where $\tau > \lambda + 1$. The set $\mathcal{K}_1$ contains at most n^τ elements, hence, (2.4) holds simultaneously over $x \in \bigcup_{k \in \mathcal{K}_1} I_k$.
Now we combine the indices from $\mathcal{K}_2$ to disjoint sets $\mathcal{K}_{21}, \ldots, \mathcal{K}_{2c_n}$, where $c_n \leq n^\tau$, $\bigcup_{i=1}^{c_n} \mathcal{K}_{2i} = \mathcal{K}_2$ and

$$P\left(X_1 \in \bigcup_{k \in \mathcal{K}_{2i}} I_k\right) \leq 2n^{-\tau}.$$

As in the proof of Lemma 2.1, we decompose the set $\{1, \ldots, n\}$ again into blocks of length $\rho_n \asymp \log(n)$. We consider again, without loss of generality, the blocks with odd numbers. After having replaced these blocks by independent ones, we can apply (5.9). It is obvious that

$$P\left(\sum_{j \in \mathcal{J}_l} I\left(X_j \in \bigcup_{k \in \mathcal{K}_{2i}} I_k\right) > 1 \quad \text{for any} \quad l\right) = O\left(\log(n) n^{-\tau+1}\right)$$

and

$$var\left(\sum_{j \in \mathcal{J}_l} I\left(X_j \in \bigcup_{k \in \mathcal{K}_{2i}} I_k\right)\right) = O(n^{-1}\log(n)),$$

which implies by (5.9) that

$$\sum_{l \text{ odd}} \sum_{j \in \mathcal{J}_l} I\left(X_j \in \bigcup_{k \in \mathcal{K}_{2i}} I_k\right) = \tilde{O}\left(\log(n), n^{-\lambda}\right).$$

For $\{Y_1, \ldots, Y_n\}$ we can show the same results, which yields the assertion. □

Proof of Theorem 3.1. It is easy to show that

$$\sup_{x \in \mathbb{R}^d} \left\{\left|\hat{f}_g(x) - f(x)\right|\right\} = \tilde{O}\left(\mu_n, n^{-\lambda}\right). \tag{5.10}$$

We show in this proof that there exists a pairing of the random variables $Y_1, \ldots, Y_n$ and $X_1^{*,g}, \ldots, X_n^{*,g}$ such that

$$\sup_{x \in \mathbb{R}^d} \left\{ \left| [\tilde{f}_h(x) - E\tilde{f}_h(x)] - [\hat{f}_h^{*,g}(x) - E\hat{f}_h^{*,g}(x)] \right| \right\}$$
$$= \tilde{O}\left((nh^d)^{-1/2} \mu_n^{1/2} \sqrt{\log(n)} + (nh^d)^{-1} \log(n), n^{-\lambda} \right). \tag{5.11}$$

The assertion of the theorem follows then in conjunction with Theorem 2.1.
Since $Y_i = X_i^{*,g}$ if $B_i = 1$, we have

$$[\tilde{f}_h(x) - E\tilde{f}_h(x)] - [\hat{f}_h^{*,g}(x) - E\hat{f}_h^{*,g}(x)] \tag{5.12}$$
$$= \frac{1}{nh^d} \sum_i \left\{ I(B_i = 0)[K(\frac{x - Y_i}{h}) - K(\frac{x - X_i^{*,g}}{h})] - \int K(\frac{x - z}{h})[f(z) - \hat{f}_g(z)]\, dz \right\}.$$

To estimate the right-hand side of (5.12), we proceed as in the proof of Theorem 2.1 and distinguish between two sets of hypercubes:

$$\mathcal{K}_1 = \left\{ k \left| \int_{I_k} |\hat{f}_g(z) - f(z)|\, dz \geq n^{-\tau} \right. \right\},$$
$$\mathcal{K}_2 = \left\{ k \left| \int_{I_k} |\hat{f}_g(z) - f(z)|\, dz < n^{-\tau} \right. \right\},$$

where $\tau > \lambda + 1$.
First we investigate the case of $x \in I_k$, $k \in \mathcal{K}_1$. Let $\mathcal{N}_n = \{x_1, \ldots, x_{c_n}\}$ be an n^{-1}-net of $\bigcup_{k \in \mathcal{K}_1} I_k$, where $c_n = O(n^{d+\tau})$. Because of

$$var\left(\frac{1}{nh^d} \sum_i \left\{ I(B_i = 0)[K(\frac{x - Y_i}{h}) - K(\frac{x - X_i^{*,g}}{h})] - \int K(\frac{x - z}{h})[f(z) - \hat{f}_g(z)]\, dz \right\} \right)$$
$$= O\left(n(nh^d)^{-2} \int_{supp(K((x-.)/h))} |f(z) - \hat{f}_g(z)|\, dz \right) = O\left((nh^d)^{-1} \mu_n \right),$$

we obtain by (5.9) that

$$\sup_{x \in \mathcal{N}_n} \left\{ \left| [\tilde{f}_h(x) - E\tilde{f}_h(x)] - [\hat{f}_h^{*,g}(x) - E\hat{f}_h^{*,g}(x)] \right| \right\}$$
$$= \tilde{O}\left((nh^d)^{-1/2} \mu_n^{1/2} \sqrt{\log(n)} + (nh^d)^{-1} \log(n), n^{-\lambda} \right). \tag{5.13}$$

Let $x \in I_k$, $k \in \mathcal{K}_1$, be arbitrary. Then there exists a $j(x) \in \{1, \ldots, c_n\}$ such that $\|x - x_{j(x)}\| = O(n^{-1})$. Since

$$|\tilde{f}_h(x) - \tilde{f}_h(x_{j(x)})| + |\hat{f}_h^{*,g}(x) - \hat{f}_h^{*,g}(x_{j(x)})| = O\left(h^{-d} n^{-1} \right)$$

is satisfied with probability 1, we have that

$$\sup_{k \in \mathcal{K}_1} \sup_{x \in I_k} \left\{ \left| [\tilde{f}_h(x) - E\tilde{f}_h(x)] - [\hat{f}_h^{*,g}(x) - E\hat{f}_h^{*,g}(x)] \right| \right\}$$
$$= \tilde{O}\left((nh^d)^{-1/2} \mu_n^{1/2} \sqrt{\log(n)} + (nh^d)^{-1} \log(n), n^{-\lambda} \right). \tag{5.14}$$

Concerning the set $\mathcal{K}_2$, we show, analogously to the corresponding part of the proof of Theorem 2.1, that

$$\sup_{k\in\mathcal{K}_2}\{\#(\{Y_1,\dots,Y_n\}\Delta\{X_1^{*,g},\dots,X_n^{*,g}\})\cap I_k\} = \tilde{O}\left(\log(n),n^{-\lambda}\right),$$

which implies

$$\sup_{k\in\mathcal{K}_2}\sup_{x\in I_k}\left\{\left|[\tilde{f}_h(x)-E\tilde{f}_h(x)]-[\hat{f}_h^{*,g}(x)-E\hat{f}_h^{*,g}(x)]\right|\right\} = \tilde{O}\left((nh^d)^{-1}\log(n),n^{-\lambda}\right). \tag{5.15}$$

(5.14) and (5.15) imply (5.11), which yields the assertion in conjunction with Theorem 2.1. □

Proof of Theorem 3.2. As already mentioned, we cannot use the idea of the proof of Theorem 3.1, because the probability measures P and P_n are orthogonal. However, we may exploit the pairing of $X_1,\dots,X_n$ and $X_1^{*,g},\dots,X_n^{*,g}$ used for proving Theorems 2.1 and 3.1 as an intermediate step to show the closeness of $[\hat{f}_h(x)-E\hat{f}_h(x)]$ and $[\hat{f}_h^{*,0}(x)-E\hat{f}_h^{*,0}(x)]$. In addition to this paring we pair the $X_i^{*,0}$'s with the $X_i^{*,g}$'s in such a way that

$$\left\|X_i^{*,0}-X_i^{*,g}\right\| \le \sqrt{d}g \tag{5.16}$$

holds with probability 1. Since

$$K\left(\frac{x-X_i^{*,0}}{h}\right)-K\left(\frac{x-X_i^{*,g}}{h}\right) = O(g/h),$$

we obtain by an approximation on a sufficiently fine grid that

$$\sup_{x\in\mathbb{R}^d}\left\{\left|[\hat{f}_h^{*,0}(x)-E\hat{f}_h^{*,0}(x)]-[\hat{f}_h^{*,g}(x)-E\hat{f}_h^{*,g}(x)]\right|\right\} = O\left((nh^d)^{-1/2}(g/h)\sqrt{\log(n)},n^{-\lambda}\right).$$

This yields, in conjunction with Theorem 3.1, that

$$\sup_{x\in\mathbb{R}^d}\left\{\left|[\hat{f}_h(x)-E\hat{f}_h(x)]-[\hat{f}_h^{*,0}(x)-E\hat{f}_h^{*,0}(x)]\right|\right\}$$
$$= O\left(n^{-1/2}\log(n)+(nh^d)^{-1}\log(n)+(nh^d)^{-1/2}\sqrt{\log(n)}[(ng^d)^{-1/2}\sqrt{\log(n)}+g/h],n^{-\lambda}\right).$$

□

Proof of Proposition 3.1. (i) *Upper estimates for Poisson probabilities*
Before we turn directly to the proof of the assertion, we first derive some technical results to be applied in the main part of this proof.
Let $P_s(\{k\}) = e^{-s}s^k/k!$ be a Poisson probability. Let $k = s\pm s^{1/2}\sqrt{r_s\log(s)}$ be an integer. Then we obtain by formula 11.9.19 in Shorack and Wellner (1986, page 486) that

$$P_s(\{k\}) = \frac{e^{-a(k)}}{\sqrt{2\pi s}}\frac{1}{\sqrt{1\pm s^{-1/2}\sqrt{r_s\log(s)}}}\exp\left(-\frac{-r_s\log(s)}{2}\psi(\pm s^{-1/2}\sqrt{r_s\log(s)})\right),$$

where $1/(12k+1) < a(k) < 1/(12k)$. Using the estimate for $\psi(.)$ given in Proposition 11.1.1(10) in Shorack and Wellner (1986, page 441), we get, for an appropriate c_λ,

$$P_s(\{k\}) = O(s^{-\lambda}), \quad \text{if } |k-s| \geq s^{1/2}\sqrt{c_\lambda \log(s)}. \tag{5.17}$$

For $k < s + s^{1/2}\sqrt{c_\lambda \log(s)}$ we obtain that

$$\begin{aligned}
&P_s(\{k, k+1, \dots\})/P_s(\{k\}) \\
&= 1 + \frac{s}{k+1} + \frac{s}{k+1}\frac{s}{k+2} + \dots \\
&\geq \left[\left(\frac{s}{c_\lambda \log(s)}\right)^{1/2}\right] \left(\frac{s}{s + \sqrt{sc_\lambda \log(s)} + [\sqrt{s/(c_\lambda \log(n))}]}\right)^{[\sqrt{s/(c_\lambda \log(s))}]} \\
&\geq \left[\left(\frac{s}{c_\lambda \log(s)}\right)^{1/2}\right] \left(1 - \frac{\sqrt{sc_\lambda \log(s)} + \sqrt{s/(c_\lambda \log(s))}}{s}\right)^{[\sqrt{s/(c_\lambda \log(s))}]} . \\
&\geq C\sqrt{s/\log(s)}.
\end{aligned} \tag{5.18}$$

Analogously we get, for $k > s - s^{1/2}\sqrt{c_\lambda \log(s)}$, that

$$\begin{aligned}
&P_s(\{k, k-1, \dots\})/P_s(\{k\}) \\
&= 1 + \frac{k}{s} + \frac{k}{s}\frac{k-1}{s} + \dots \\
&\geq \left[\left(\frac{s}{c_\lambda \log(s)}\right)^{1/2}\right] \left(\frac{s - \sqrt{sc_\lambda \log(s)} - \sqrt{s/(c_\lambda \log(s))}}{s}\right)^{[\sqrt{s/(c_\lambda \log(s))}]} \\
&\geq C\sqrt{s/\log(s)}.
\end{aligned} \tag{5.19}$$

(5.17) and (5.18) imply

$$P_s(\{k\}) \leq C_\lambda \left[\sqrt{\log(s)/s}\, P_s(\{k, k+1, \dots\}) + s^{-\lambda}\right], \tag{5.20}$$

and (5.17) and (5.19) yield

$$P_s(\{k\}) \leq C_\lambda \left[\sqrt{\log(s)/s}\, P_s(\{k, k-1, \dots\}) + s^{-\lambda}\right] \tag{5.21}$$

for all $k \in \mathbb{Z}$.

(ii) *Some preparatory considerations*

We consider instead of $\widehat{f}_h$ the artificial quantity

$$\bar{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{\nu} K\left(\frac{x - Z_i}{h}\right), \tag{5.22}$$

where $\{Z_1, \dots, Z_\nu\}$ were defined by (2.3).

The crucial point is that $\bar{f}_h$ is based on a Poisson process instead of an empirical process. Therefore, $\bar{f}_h(x_1)$ and $\bar{f}_h(x_2)$ are independent, if the supports of the corresponding kernels are disjoint.
We decompose the $\mathbb{R}^d$ into nonoverlapping hypercubes of sidelength $2h$, that is

$$I_k = [2(k_1-1)h, 2k_1h) \times \ldots \times [2(k_d-1)h, 2k_dh).$$

Further, we divide the set $\mathbb{Z}^d$ into 2^d subsets,

$$\mathcal{K}_l = \{k = (k_1, \ldots, k_d) \mid \; k_i = 2j_i + l_i, \; j_i \in \mathbb{Z}\},$$

where $l = (l_1, \ldots, l_d) \in \{0,1\}^d$. We fix l and consider

$$Z_l = \sup_{k \in \mathcal{K}_l} \sup_{x \in I_k} \left\{ |\bar{f}_h(x) - E\hat{f}_h(x)| \right\}.$$

It can be seen from the following considerations that

$$P\left(Z_l < C_\lambda (nh^d)^{-1/2} (\log(n))^{1/2}\right) = O(n^{-\lambda}) \tag{5.23}$$

holds for sufficiently small C_λ.
Let

$$\mu_k = \sup_{x \in I_k} \left\{ \int h^{-d} K\left(\frac{x-z}{h}\right) f(z)\, dz \right\}.$$

Similarly to the considerations in the proof of Theorem 2.1, we can show that

$$P\left(\sup_{k:\, \mu_k < \bar{\mu}} \sup_{x \in I_k} \left\{ |\bar{f}_h(x) - E\hat{f}_h(x)| \right\} \geq C_\lambda (nh^d)^{-1/2} (\log(n))^{1/2} \right) = O(n^{-\lambda}), \tag{5.24}$$

for some $\bar{\mu}$ sufficiently small. Hence, with a probability exceeding $1 - O(n^{-\lambda})$, the supremum Z_l will be attained on one of the intervals I_k with $\mu_k \geq \bar{\mu}$. Let

$$\{k_1, \ldots, k_{\rho_l}\} = \{k \in \mathcal{K}_l \mid \; \mu_k \geq \bar{\mu}\}.$$

(iii) *Decomposition of* $\bar{f}_h(x) - E\hat{f}_h(x)$

Let

$$\mathcal{J}_k = I_k \oplus supp\left(K\left(\frac{x-\cdot}{h}\right)\right) = [(2k_1-3)h, (2k_1+1)h) \times \ldots \times [(2k_d-3)h, (2k_d+1)h).$$

Further, let Z_i^k be the i-th variable of $Z_1, \ldots, Z_\nu$ that falls into $\mathcal{J}_k$, and let $\hat{\nu}_k = \#\{1 \leq i \leq \nu \mid \; Z_i \in \mathcal{J}_k\}$ be the number of them. Then

$$\bar{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{\hat{\nu}_k} K\left(\frac{x - Z_i^k}{h}\right).$$

Let $\nu_k = E\hat{\nu}_k = nP(Z_1 \in \mathcal{J}_k)$.

Now we have, for $x \in I_k$, that

$$\begin{aligned}
&\bar{f}_h(x) - E\widehat{f}_h(x) \\
&= \frac{1}{nh^d}\sum_{i=1}^{\widehat{\nu}_k}\left[K\left(\frac{x-Z_i^k}{h}\right) - \frac{1}{P(Z_1 \in \mathcal{J}_k)}\int K\left(\frac{x-z}{h}\right) f(z)\,dz\right] \\
&\quad + \frac{1}{nh^d}\frac{\widehat{\nu}_k - \nu_k}{P(Z_1 \in \mathcal{J}_k)}\int K\left(\frac{x-z}{h}\right) f(z)\,dz \\
&= \frac{1}{nh^d}\frac{\widehat{\nu}_k - \nu_k}{P(Z_1 \in \mathcal{J}_k)}\frac{1}{|I_k|}\int_{I_k}\int K\left(\frac{x-z}{h}\right) f(z)\,dz\,dx \\
&\quad + \frac{1}{nh^d}\sum_{i=1}^{[\nu_k]}\left[K\left(\frac{x-Z_i^k}{h}\right) - \frac{1}{P(Z_1 \in \mathcal{J}_k)}\int K\left(\frac{x-z}{h}\right) f(z)\,dz\right] \\
&\quad + \frac{1}{nh^d}\frac{\widehat{\nu}_k - \nu_k}{P(Z_1 \in \mathcal{J}_k)}\left\{\int K\left(\frac{x-z}{h}\right) f(z)\,dz - \frac{1}{|I_k|}\int_{I_k}\int K\left(\frac{x-z}{h}\right) f(z)\,dz\right\} \\
&\quad + \frac{1}{nh^d}\left\{\sum_{i=1}^{\widehat{\nu}_k}[\ldots] - \sum_{i=1}^{[\nu_k]}[\ldots]\right\} \\
&= T_{k1} + T_{k2}(x) + R_{k1}(x) + R_{k2}(x). \qquad (5.25)
\end{aligned}$$

The main purpose of this decomposition was to split $\bar{f}_h(x) - E\widehat{f}_h(x)$ into a term T_{k1} proportional to the Poisson variable $\widehat{\nu}_k \sim P_{\nu_k}$, a term $\{T_{k2}(x)\}_{x\in I_k}$ independent of T_{k1}, and two asymptotically negligible terms, $R_{k1}(x)$ and $R_{k2}(x)$.

(iv) *Proof of the assertion*

Next we show that

$$P\left(\sup_{k\in\mathcal{K}_l}\{T_{k1} + T_{k2}\} \in [c,d]\right) = O\left((d-c)(nh^d)^{1/2}\sqrt{\log(n)} + n^{-\lambda}\right), \quad (5.26)$$

where

$$T_{k2} = \sup_{x\in I_k}\{T_{k2}(x)\}.$$

We keep for a moment $\{T_{k2}\}_{k\in\mathcal{K}_l}$ fixed. Since the T_{k1}'s are independent of the T_{k2}'s, we obtain, by (5.20), that

$$
\begin{aligned}
&P\left(\sup_{k\in\mathcal{K}_l}\{T_{k1}+T_{k2}\}\in[c,d]\,\middle|\; T_{k_1,2},\ldots,T_{k_{\rho_l},2}\right)\\
&\leq P\left(T_{k_1,1}\in[c-T_{k_1,2},d-T_{k_1,2}]\right)\\
&\quad + P\left(T_{k_2,1}\in[c-T_{k_2,2},d-T_{k_2,2}];\; T_{k_1,1}<c-T_{k_1,2}\right)\\
&\quad +\ldots+ P\left(T_{k_{\rho_l},1}\in[c-T_{k_{\rho_l},2},d-T_{k_{\rho_l},2}];\; T_{k_1,1}<c-T_{k_1,2},\ldots,T_{k_{\rho_l-1},1}<c-T_{k_{\rho_l-1},2}\right)\\
&\leq (d-c)(nh^d)^{1/2}\sqrt{\log(n)}\left\{P\left(T_{k_1,1}>c-T_{k_1,2}\right)+\right.\\
&\qquad \left.+\ldots+P\left(T_{k_{\rho_l},1}>c-T_{k_{\rho_l},2};\; T_{k_1,1}\leq c-T_{k_1,2},\ldots,T_{k_{\rho_l-1},1}\leq c-T_{k_{\rho_l-1},2}\right)\right\}\\
&\quad + O(n^{-\lambda})\\
&= O\left((d-c)(nh^d)^{1/2}\sqrt{\log(n)}P\left(\sup_{k\in\mathcal{K}_l}\{T_{k1}+T_{k2}\}>c\,\middle|\; T_{k_1,2},\ldots,T_{k_{\rho_l},2}\right)\right)+O(n^{-\lambda})\\
&= O\left((d-c)(nh^d)^{1/2}\sqrt{\log(n)}+n^{-\lambda}\right). \qquad (5.27)
\end{aligned}
$$

Integrating over all realizations for $T_{k_1,2},\ldots,T_{k_{\rho_l},2}$, we get (5.26).
Since f is Lipschitz, we easily obtain that

$$\sup_{x\in I_k}\{|R_{k1}(x)|\}=\tilde{O}\left((nh^d)^{-1/2}h\sqrt{\log(n)},n^{-\lambda}\right). \qquad (5.28)$$

Because of $\widehat{\nu}_k-[\nu_k]=\tilde{O}((nh^d)^{1/2}(\log(n))^{1/2},n^{-\lambda})$, we can readily show that

$$\sup_{x\in I_k}\{|R_{k2}(x)|\}=\tilde{O}\left((nh^d)^{-3/4}(\log(n))^{3/4}+(nh^d)^{-1}\log(n),n^{-\lambda}\right). \qquad (5.29)$$

By (5.25), (5.26), (5.28) and (5.29) we obtain, with $\kappa_n=C[(nh^d)^{-1/2}h(\log(n))^{1/2}+(nh^d)^{-3/4}(\log(n))^{3/4}]$, that

$$
\begin{aligned}
&P\left(Z_l\in[c,d]\right)\\
&\leq P\left(\sup_{k\in\mathcal{K}_l}\{T_{k1}+T_{k2}\}\in[c-\kappa_n,d+\kappa_n]\right)+O(n^{-\lambda})\\
&= O\left((d-c)(nh^d)^{1/2}(\log(n))^{1/2}+h\log(n)+(nh^d)^{-1/4}(\log(n))^{5/4}\right). \qquad (5.30)
\end{aligned}
$$

By analogous considerations, where we only have to use (5.21) instead of (5.20), we obtain

$$
\begin{aligned}
&P\left(\inf_{k\in\mathcal{K}_l}\inf_{x\in I_k}\left\{\bar{f}_h(x)-E\widehat{f}_h(x)\right\}\in[-d,-c]\right)\\
&= O\left((d-c)(nh^d)^{1/2}(\log(n))^{1/2}+h\log(n)+(nh^d)^{-1/4}(\log(n))^{5/4}\right). \qquad (5.31)
\end{aligned}
$$

This implies

$$P\left(\sup_{x\in\mathbb{R}^d}\{|\bar{f}_h(x) - E\widehat{f}_h(x)|\} \in [c,d]\right)$$
$$\leq \sum_{l\in\{0,1\}^d} P\left(\sup_{k\in K_l}\sup_{x\in I_k}\{|\bar{f}_h(x) - E\widehat{f}_h(x)|\} \in [c,d]\right) + O(n^{-\lambda})$$
$$= O\left((d-c)(nh^d)^{1/2}(\log(n))^{1/2} + h\log(n) + (nh^d)^{-1/4}(\log(n))^{5/4}\right). \quad (5.32)$$

Using

$$\sup_{x\in\mathbb{R}^d}\{|\bar{f}_h(x) - \widehat{f}_h(x)|\} = \tilde{O}\left([n^{-1/2} + (nh^d)^{-1}]\log(n), n^{-\lambda}\right),$$

we obtain the assertion. □

Theorems 3.3, 3.4 and 3.5 are straightforward implications of Theorem 3.1 and Proposition 3.1. We omit these proofs.

Acknowledgment. Part of the research was carried out while the author was visiting the Institute of Statistics at the Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

References

Amosova, N. N. (1972). On limit theorems for probabilities of moderate deviations. *Vestnik Leningrad. Univ.* **13**, 5–14. (in Russian)

Ango Nze, P. (1992). Critères d'ergodicité de quelques modèles à représentation markovienne. *C. R. Acad. Sci. Paris, Ser. I* **315**, 1301–1304.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the derivation of density function estimators. *Ann. Statist.* **1**, 1071–1095.

Bühlmann, P. (1994). Blockwise bootstrapped empirical process for stationary sequences. *Ann. Statist.* **22**, 995-1012.

Dhompongsa, S. (1984). A note on the almost sure approximation of the empirical process of weakly dependent random vectors. *Yokohama Math. J.* **32**, 113–121.

Doukhan, P. (1994). *Mixing: Properties and Examples. Lecture Notes in Statistics* **85**, Springer, New York.

Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. Henri Poincaré* **31**, 393–427.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Faraway, J. J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85**, 1119–1122.

Friedman, J., Stuetzle, W. and Schroeder, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79**, 599–608.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series.* Springer, Berlin.

Hall, P. (1991). On convergence rates of suprema. *Probab. Theory Rel. Fields* **89**, 447–455.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *J. R. Statist. Soc.* B **55**, 291–304.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–1947.

Hart, J. D. (1995). Some automated methods of smoothing time-dependent data. *J. Nonpar. Statist.* **6**, 115-142.

Konakov, V., Läuter, H. and Liero, H. (1995). Comparison of the asymptotic power of tests based on L_2- and L_∞-norms under non-standard local alternatives. Discussion Paper No. 10/95, SFB 373, Humboldt University, Berlin.

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent rv's and the sample distribution function. *Z. Wahrsch. verw. Geb.* **32**, 111–131.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.

Masry, E. (1994). Probability density estimation from dependent observations using wavelet orthonormal bases. *Statist. Probab. Lett.* **21**, 181–194.

Mokkadem, A. (1990). Propriétés de mélange des processus autorégessifs polynomiaux. *Ann. Inst. Henri Poincaré* **26**, 219–260.

Neumann, M. H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23**, 1937–1959.

Neumann, M. H. and Kreiss, J.-P. (1996). Bootstrap confidence bands for the autoregression function, Preprint No. 263, Weierstrass Institute, Berlin.

Reiss, R.-D. (1993). *A Course on Point Processes.* Springer, New York.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* 4, 185–207.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Silverman, B. W. and Young, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika* 74, 469–479.

Spokoiny, V. G. (1996). Adaptive and spatially adaptive testing of a nonparametric hypothesis. Preprint No. 234, Weierstrass Institute, Berlin.

Recent publications of the Weierstraß–Institut für Angewandte Analysis und Stochastik

Preprints 1996

266. Daniela Peterhof, Björn Sandstede, Arnd Scheel: Exponential dichotomies for solitary–wave solutions of semilinear elliptic equations on infinite cylinders.

267. Andreas Rathsfeld: A wavelet algorithm for the solution of a singular integral equation over a smooth two–dimensional manifold.

268. Jörg Schmeling, Serge E. Troubetzkoy: Dimension and invertibility of hyperbolic endomorphisms with singularities.

269. Erwin Bolthausen, Dmitry Ioffe: Harmonic crystal on the wall: a microscopic approach.

270. Nikolai N. Nefedov, Klaus R. Schneider: Delayed exchange of stabilities in singularly perturbed systems.

271. Michael S. Ermakov: On large and moderate large deviations of empirical bootstrap measure.

272. Priscilla E. Greenwood, Jiaming Sun: On criticality for competing influences of boundary and external field in the Ising model.

273. Michael S. Ermakov: On distinguishability of two nonparametric sets of hypothesis.

274. Henri Schurz: Preservation of probabilistic laws through Euler methods for Ornstein–Uhlenbeck process.

275. Lev B. Ryashko, Henri Schurz: Mean square stability analysis of some linear stochastic systems.

276. Nikolaus Bubner, Jan Sokołowski, Jürgen Sprekels: Optimal boundary control problems for shape memory alloys under state constraints for stress and temperature.

277. Alison M. Etheridge, Klaus Fleischmann: Persistence of a two–dimensional super–Brownian motion in a catalytic medium.

278. Sergej Rjasanow, Wolfgang Wagner: Stochastic interacting particle systems as a numerical tool.

279. Vadim G. Bondarevsky: Energetic systems and global attractors for the 3D Navier–Stokes equations.

280. Henri Schurz: The invariance of asymptotic laws of stochastic systems under discretization.

281. Michael Nussbaum: The Pinsker bound: a review.

282. Pierluigi Colli, Maurizio Grasselli, Jürgen Sprekels: Automatic control via thermostats of a hyperbolic Stefan problem with memory.

283. Anton Bovier, Véronique Gayrard: An almost sure central limit theorem for the Hopfield model.

284. Peter Mathé: Efficient mixing of product walks on product groups.

285. Günter Albinus: Convex analysis of the energy model of semiconductor devices.

286. Ilja Schmelzer: Postrelativity – a paradigm for quantization with preferred Newtonian frame.

287. Evgenii Ya. Khruslov, Holger Stephan: Splitting of some nonlocalized solutions of the Korteweg–de Vries equation into solitons.

288. Björn Sandstede, Arnd Scheel, Claudia Wulff: Dynamics of spiral waves on unbounded domains using center–manifold reductions.

289. Ion Grama, Michael Nussbaum: Asymptotic equivalence for nonparametric generalized linear models.

290. Pierluigi Colli, Jürgen Sprekels: Weak solution to some Penrose–Fife phase–field systems with temperature–dependent memory.

291. Vladimir G. Spokoiny: Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice.

292. Peter E. Kloeden, Eckhard Platen, Henri Schurz, Michael Sørensen: On effects of discretization on estimators of drift parameters for diffusion processes.

293. Erlend Arge, Angela Kunoth: An efficient ADI–solver for scattered data problems with global smoothing.

294. Alfred Liemant, Ludwig Brehmer: A mean field approximation for hopping transport in disordered materials.