Econometric Theory, **24**, 2008, 726–748. Printed in the United States of America.
doi: 10.1017/S0266466608080304

UNIFORM CONVERGENCE RATES FOR KERNEL ESTIMATION WITH DEPENDENT DATA

BRUCE E. HANSEN
University of Wisconsin

This paper presents a set of rate of uniform consistency results for kernel estimators of density functions and regressions functions. We generalize the existing literature by allowing for stationary strong mixing multivariate data with infinite support, kernels with unbounded support, and general bandwidth sequences. These results are useful for semiparametric estimation based on a first-stage nonparametric estimator.

1. INTRODUCTION

This paper presents a set of rate of uniform consistency results for kernel estimators of density functions and regressions functions. We generalize the existing literature by allowing for stationary strong mixing multivariate data with infinite support, kernels with unbounded support, and general bandwidth sequences.

Kernel estimators were first introduced by Rosenblatt (1956) for density estimation and by Nadaraya (1964) and Watson (1964) for regression estimation. The local linear estimator was introduced by Stone (1977) and came into prominence through the work of Fan (1992, 1993).

Uniform convergence for kernel averages has been previously considered in a number of papers, including Peligrad (1991), Newey (1994), Andrews (1995), Liebscher (1996), Masry (1996), Bosq (1998), Fan and Yao (2003), and Ango Nze and Doukhan (2004).

In this paper we provide a general set of results with broad applicability. Our main results are the weak and strong uniform convergence of a sample average functional. The conditions imposed on the functional are general. The data are assumed to be a stationary strong mixing time series. The support for the data is allowed to be infinite, and our convergence is uniform over compact sets, expanding sets, or unrestricted euclidean space. We do not require the regression function or its derivatives to be bounded, and we allow for kernels with

This research was supported by the National Science Foundation. I thank three referees and Oliver Linton for helpful comments. Address correspondence to Bruce E. Hansen, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706-1393, USA; e-mail: bhansen@ssc.wisc.edu.

  0266-4666/08 $15.00

unbounded support. The rate of decay for the bandwidth is flexible and includes the optimal convergence rate as a special case. Our applications include estimation of multivariate densities and their derivatives, Nadaraya–Watson regression estimates, and local linear regression estimates. We do not consider local polynomial regression, although our main results could be applied to this application also.

These features are useful generalizations of the existing literature. Most papers assume that the kernel function has truncated support, which excludes the popular Gaussian kernel. It is also typical to demonstrate uniform convergence only over fixed compact sets, which is sufficient for many estimation purposes but is insufficient for many semiparametric applications. Some papers assume that the regression function, or certain derivatives of the regression function, is bounded. This may appear innocent when convergence is limited to fixed compact sets but is unsatisfactory when convergence is extended to expanding or unbounded sets. Some papers only present convergence rates using optimal bandwidth rates. This is inappropriate for many semiparametric applications where the bandwidth sequences may not satisfy these conditions. Our paper avoids these deficiencies.

Our proof method is a generalization of those in Liebscher (1996) and Bosq (1998).

Section 2 presents results for a general class of functions, including a variance bound, weak uniform convergence, strong uniform convergence, and convergence over unbounded sets. Section 3 presents applications to density estimation, Nadaraya–Watson regression, and local linear regression. The proofs are in the Appendix.

Regarding notation, for $x = (x_1, \ldots, x_d) \in R^d$ we set $\|x\| = \max(|x_1|, \ldots, |x_d|)$.

2. GENERAL RESULTS

2.1. Kernel Averages and a Variance Bound

Let $\{Y_i, X_i\} \in R \times R^d$ be a sequence of random vectors. The vector X_i may include lagged values of Y_i, e.g., $X_i = (Y_{i-1}, \ldots, Y_{i-d})$. Consider averages of the form

$$\hat{\Psi}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $h = o(1)$ is a bandwidth and $K(u): R^d \to R$ is a kernel-like function. Most kernel-based nonparametric estimators can be written as functions of averages of this form. By suitable choice of $K(u)$ and Y_i this includes kernel estimators of density functions, Nadaraya–Watson estimators of the regression

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

function, local polynomial estimators, and estimators of derivatives of density and regression functions.

We require that the function $K(u)$ is bounded and integrable:

Assumption 1. $|K(u)| \leq \bar{K} < \infty$ and $\int_{R^d} |K(u)|du \leq \mu < \infty$.

We assume that $\{Y_i, X_i\}$ is weakly dependent. We require the following regularity conditions.

Assumption 2. The sequence $\{Y_i, X_i\}$ is strictly stationary and strong mixing with mixing coefficients α_m that satisfy

$$\alpha_m \leq Am^{-\beta}, \tag{2}$$

where $A < \infty$ and for some $s > 2$

$$\mathrm{E}|Y_0|^s < \infty \tag{3}$$

and

$$\beta > \frac{2s-2}{s-2}. \tag{4}$$

Furthermore, X_i has marginal density $f(x)$ such that

$$\sup_x f(x) \leq B_0 < \infty \tag{5}$$

and

$$\sup_x \mathrm{E}(|Y_0|^s | X_0 = x) f(x) \leq B_1 < \infty. \tag{6}$$

Also, there is some $j^* < \infty$ such that for all $j \geq j^*$

$$\sup_{x_0, x_j} \mathrm{E}(|Y_0 Y_j| | X_0 = x_0, X_j = x_j) f_j(x_0, x_j) \leq B_2 < \infty, \tag{7}$$

where $f_j(x_0, x_j)$ denotes the joint density of $\{X_0, X_j\}$.

Assumption 2 specifies that the serial dependence in the data is strong mixing, and equations (2)–(4) specify a required decay rate. Condition (5) specifies that the density $f(x)$ is bounded, and (6) controls the tail behavior of the conditional expectation $\mathrm{E}(|Y_0|^s | X_0 = x)$. The latter can increase to infinity in the tails but not faster than $f(x)^{-1}$. Condition (7) places a similar bound on the joint density and conditional expectation. If the data are independent or m-dependent, then (7) is immediately satisfied under (6) with $B_2 = B_1^2$.

In many applications (such as density estimation) Y_i is bounded. In this case we can take $s = \infty$, (4) simplifies to $\beta > 2$, (6) is redundant with (5), and (7) is equivalent to $f_j(x_0, x_j) \leq B_2$ for all $j \geq j^*$.

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

The bound (7) requires that $\{X_0, X_j\}$ have a bounded joint density $f_j(x_0, x_j)$ for sufficient large j, but the joint density does not need to exist for small j. This distinction allows X_i to consist of multiple lags of Y_i. For example, if $X_i = (Y_{i-1}, Y_{i-2}, \ldots, Y_{i-d})$ for $d \geq 2$ then $f_j(x_0, x_j)$ is unbounded for $j < d$ because the components of X_0 and X_j overlap.

THEOREM 1. *Under Assumptions 1 and 2 there is a $\Theta < \infty$ such that for n sufficiently large*

$$\operatorname{Var}(\hat{\Psi}(x)) \leq \frac{\Theta}{nh^d}. \tag{8}$$

An expression for Θ is given in equation (A.5) in the Appendix.

Although Theorem 1 is elementary for independent observations, it is nontrivial for dependent data because of the presence of nonzero covariances. Our proof builds on the strategy of Fan and Yao (2003, pp. 262–263) by separately bounding covariances of short, medium, and long lag lengths.

2.2. Weak Uniform Convergence

Theorem 1 implies that $|\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = O_p((nh^d)^{-1/2})$ pointwise in $x \in R^d$. We are now interested in uniform rates. We start by considering uniformity over values of x in expanding sets of the form $\{x : \|x\| \leq c_n\}$ for sequences c_n that are either bounded or diverging slowly to infinity. To establish uniform convergence, we need the function $K(u)$ to be smooth. We require that K either has truncated support and is Lipschitz or that it has a bounded derivative with an integrable tail.

Assumption 3. For some $\Lambda_1 < \infty$ and $L < \infty$, either $K(u) = 0$ for $\|u\| > L$ and for all $u, u' \in R^d$

$$|K(u) - K(u')| \leq \Lambda_1 \|u - u'\|, \tag{9}$$

or $K(u)$ is differentiable, $|(\partial/\partial u)K(u)| \leq \Lambda_1$, and for some $\nu > 1$, $|(\partial/\partial u)K(u)| \leq \Lambda_1 \|u\|^{-\nu}$ for $\|u\| > L$.

Assumption 3 allows for most commonly used kernels, including the polynomial kernel class $c_p(1 - x^2)^p$, the higher order polynomial kernels of Müller (1984) and Granovsky and Müller (1991), the normal kernel, and the higher order Gaussian kernels of Wand and Schucany (1990) and Marron and Wand (1992). Assumption 3 excludes, however, the uniform kernel. It is unlikely that this is a necessary exclusion, as Tran (1994) established uniform convergence

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

of a histogram density estimator. Assumption 3 also excludes the Dirichlet kernel $K(x) = \sin(x)/(\pi x)$.

THEOREM 2. *Suppose that Assumptions 1–3 hold and for some $q > 0$ the mixing exponent β satisfies*

$$\beta > \frac{1 + (s-1)\left(1 + \frac{d}{q} + d\right)}{s-2} \tag{10}$$

and for

$$\theta = \frac{\beta - 1 - d - \frac{d}{q} - (1+\beta)/(s-1)}{\beta + 3 - d - (1+\beta)/(s-1)} \tag{11}$$

the bandwidth satisfies

$$\frac{\ln n}{n^{\theta} h^{d}} = o(1). \tag{12}$$

Then for

$$c_n = O((\ln n)^{1/d} n^{1/2q}) \tag{13}$$

and

$$a_n = \left(\frac{\ln n}{nh^{d}}\right)^{1/2}, \tag{14}$$

$$\sup_{\|x\| \le c_n} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = O_p(a_n). \tag{15}$$

Theorem 2 establishes the rate for uniform convergence in probability. Using (10) and (11) we can calculate that $\theta \in (0,1]$ and thus (12) is a strengthening of the conventional requirement that $nh^{d} \to \infty$. Also note that (10) is a strict strengthening of (4). If Y_i is bounded, we can take $s = \infty$, and then (10) and (11) simplify to $\beta > 1 + (d/q) + d$ and $\theta = (\beta - 1 - d - (d/q))/(\beta + 3 - d)$. If $q = \infty$ and $d = 1$ then this simplifies further to $\beta > 2$ and $\theta = (\beta - 2)/(\beta + 2)$, which is weaker than the conditions of Fan and Yao (2003, Lem. 6.1). If the mixing coefficients have geometric decay ($\beta = \infty$) then $\theta = 1$ and (15) holds for all q.

It is also constructive to compare Theorem 2 with Lemma B.1 of Newey (1994). Newey's convergence rate is identical to (15), but his result is restricted to independent observations, kernel functions K with bounded support, and bounded c_n.

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

2.3. Almost Sure Uniform Convergence

In this section we strengthen the result of the previous section to almost sure convergence.

THEOREM 3. *Define* $\phi_n = (\ln \ln n)^2 \ln n$. *Suppose that Assumptions 1–3 hold and for some* $q > 0$ *the mixing exponent* β *satisfies*

$$\beta > \frac{2 + s\left(3 + \frac{d}{q} + d\right)}{s - 2} \tag{16}$$

and for

$$\theta = \frac{\beta\left(1 - \frac{2}{s}\right) - \frac{2}{s} - 3 - \frac{d}{q} - d}{\beta + 3 - d} \tag{17}$$

the bandwidth satisfies

$$\frac{\phi_n^2}{n^\theta h^d} = O(1). \tag{18}$$

Then for

$$c_n = O(\phi_n^{1/d} n^{1/2q}), \tag{19}$$

$$\sup_{\|x\| \le c_n} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = O(a_n) \tag{20}$$

almost surely, where a_n *is defined in (14).*

The primary difference between Theorems 2 and 3 is the condition on the strong mixing coefficients.

2.4. Uniform Convergence over Unbounded Sets

The previous sections considered uniform convergence over bounded or slowly expanding sets. We now consider uniform convergence over unrestricted euclidean space. This requires additional moment bounds on the conditioning variables and polynomial tail decay for the function $K(u)$.

THEOREM 4. *Suppose the assumptions of Theorem 2 hold with* $h = O(1)$ *and* $q \ge d$. *Furthermore,*

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

$$\sup_x \|x\|^q \mathrm{E}(|Y_0||X_0 = x) f(x) \leq B_3 < \infty, \tag{21}$$

and for $\|u\| \geq L$

$$|K(u)| \leq \Lambda_2 \|u\|^{-q} \tag{22}$$

for some $\Lambda_2 < \infty$. *Then*

$$\sup_{x \in R^d} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = O_p(a_n).$$

THEOREM 5. *Suppose the assumptions of Theorem 3 hold with* $h = O(1)$ *and* $q \geq d$. *Furthermore, (21), (22), and* $\mathrm{E}\|X_0\|^{2q} < \infty$ *hold. Then*

$$\sup_{x \in R^d} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = O(a_n)$$

almost surely.

Theorems 4 and 5 show that the extension to uniformity over unrestricted euclidean space can be made with minimal additional assumptions. Equation (21) is a mild tail restriction on the conditional mean and density function. The kernel tail restriction (22) is satisfied by the kernels discussed in Section 2.2 for all $q > 0$.

3. APPLICATIONS

3.1. Density Estimation

Let $X_i \in R^d$ be a strictly stationary time series with density $f(x)$. Consider the estimation of $f(x)$ and its derivatives $f^{(r)}(x)$. Let $k(u) : R^d \to R$ denote a multivariate pth-order kernel function for which $k^{(r)}(u)$ satisfies Assumption 1 and $\int |u|^{p+r} |k(u)| du < \infty$. The Rosenblatt (1956) estimator of the rth derivative $f^{(r)}(x)$ is

$$\hat{f}^{(r)}(x) = \frac{1}{nh^{d+r}} \sum_{i=1}^{n} k^{(r)}\left(\frac{x - X_i}{h}\right),$$

where h is a bandwidth.

We first consider uniform convergence in probability.

THEOREM 6. *Suppose that for some* $q > 0$, *the strong mixing coefficients satisfy (2) with*

$$\beta > 1 + \frac{d}{q} + d, \tag{23}$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

$h = o(1)$, and (12) holds with

$$\theta = \frac{\beta - 1 - \frac{d}{q} - d}{\beta + 3 - d}. \tag{24}$$

Suppose that $\sup_x f(x) < \infty$ *and there is some* $j^* < \infty$ *such that for all* $j \geq j^*$, $\sup_{x_0, x_j} f_j(x_0, x_j) < \infty$ *where* $f_j(x_0, x_j)$ *denotes the joint density of* $\{X_0, X_j\}$. *Assume that the pth derivative of* $f^{(r)}(x)$ *is uniformly continuous. Then for any sequence* c_n *satisfying (13),*

$$\sup_{\|x\| \leq c_n} |\hat{f}^{(r)}(x) - f^{(r)}(x)| = O_p\left(\left(\frac{\ln n}{nh^{d+2r}}\right)^{1/2} + h^p\right). \tag{25}$$

The optimal convergence rate (by selecting the bandwidth h optimally) can be obtained when

$$\beta > 1 + d + \frac{d}{q} + \frac{d}{p + r}\left(2 + \frac{d}{2q}\right) \tag{26}$$

and is

$$\sup_{\|x\| \leq c_n} |\hat{f}^{(r)}(x) - f^{(r)}(x)| = O_p\left(\left(\frac{\ln n}{n}\right)^{p/(d+2p+2r)}\right). \tag{27}$$

Furthermore, if in addition $\sup_x \|x\|^q f(x) < \infty$ *and* $|k^{(r)}(u)| \leq \Lambda_2 \|u\|^{-q}$ *for* $\|u\|$ *large, then the supremum in (25) or (27) may be taken over* $x \in R^d$.

Take the simple case of estimation of the density ($r = 0$), second-order kernel ($p = 2$), and bounded c_n ($q = \infty$). In this case the requirements state that $\beta > 1 + d$ is sufficient for (25) and $\beta > 1 + 2d$ is sufficient for the optimal convergence rate (27). This is an improvement upon the work of Fan and Yao (2003, Thm. 5.3), who (for $d = 1$) require $\beta > \frac{5}{2}$ and $\beta > \frac{15}{4}$ for these two results.

An alternative uniform weak convergence rate has been provided by Andrews (1995, Thm. 1(a)). His result is more general in allowing for near-epoch-dependent arrays, but he obtains a slower rate of convergence.

We now consider uniform almost sure convergence.

THEOREM 7. *Under the assumptions of Theorem 6, if* $\beta > 3 + (d/q) + d$ *and (18) and (19) hold with*

$$\theta = \frac{\beta - 3 - \frac{d}{q} - d}{\beta + 3 - d},$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

then

$$\sup_{\|x\|\le c_n} |\hat{f}^{(r)}(x) - f^{(r)}(x)| = O\left(\left(\frac{\ln n}{nh^{d+2r}}\right)^{1/2} + h^p\right)$$

almost surely. The optimal convergence rate when

$$\beta > 3 + d + \frac{d}{q} + \frac{d}{p+r}\left(3 + \frac{d}{2q}\right)$$

is

$$\sup_{\|x\|\le c_n} |\hat{f}^{(r)}(x) - f^{(r)}(x)| = O\left(\left(\frac{\ln n}{n}\right)^{p/(d+2p+2r)}\right) \tag{28}$$

almost surely.

Alternative results for strong uniform convergence for kernel density estimates have been provided by Peligrad (1991), Liebscher (1996, Thms. 4.2 and 4.3), Bosq (1998, Thm. 2.2 and Cor. 2.2), and Ango Nze and Doukhan (2004). Theorem 6 contains Liebscher's result as the special case $r = 0$ and $q = \infty$, and he restricts attention to kernels with bounded support. Peligrad imposes ρ-mixing and bounded c_n. Bosq restricts attention to geometric strong mixing.

3.2. Nadaraya–Watson Regression

Consider the estimation of the conditional mean

$$m(x) = \mathrm{E}(Y_i | X_i = x).$$

Let $k(u): R^d \to R$ denote a multivariate symmetric kernel function that satisfies Assumptions 1 and 3 and let $\int |u|^2 |k(u)|\,du < \infty$. The Nadaraya–Watson estimator of $m(x)$ is

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i k\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n k\left(\frac{x - X_i}{h}\right)},$$

where h is a bandwidth.

THEOREM 8. *Suppose that Assumption 2 and equations (10)–(13) hold and the second derivatives of $f(x)$ and $f(x)m(x)$ are uniformly continuous and bounded. If*

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

$$\delta_n = \inf_{|x| \le c_n} f(x) > 0,$$

$h = o(1)$, *and* $\delta_n^{-1} a_n^* \to 0$ *where*

$$a_n^* = \left(\frac{\log n}{nh^d}\right)^{1/2} + h^2, \tag{29}$$

then

$$\sup_{|x| \le c_n} |\hat{m}(x) - m(x)| = O_p(\delta_n^{-1} a_n^*). \tag{30}$$

The optimal convergence rate when β *is sufficiently large is*

$$\sup_{|x| \le c_n} |\hat{m}(x) - m(x)| = O_p\left(\delta_n^{-1}\left(\frac{\ln n}{n}\right)^{2/(d+4)}\right). \tag{31}$$

THEOREM 9. *Suppose that the assumptions of Theorem 8 hold and equations (16)–(19) hold instead of (10)–(13). Then (30) and (31) can be strengthened to almost sure convergence.*

If c_n is a constant then the convergence rate is a_n, and the optimal rate is $(n^{-1} \ln n)^{2/(d+4)}$, which is the Stone (1982) optimal rate for independent and identically distributed (i.i.d.) data. Theorems 8 and 9 show that the uniform convergence rate is not penalized for dependent data under the strong mixing assumption.

For semiparametric applications, it is frequently useful to require $c_n \to \infty$ so that the entire function $m(x)$ is consistently estimated. From (30) we see that this induces the additional penalty term δ_n^{-1}.

Alternative results for the uniform rate of convergence for the Nadaraya–Watson estimator have been provided by Andrews (1995, Thm. 1(b)) and Bosq (1998, Thms. 3.2 and 3.3). Andrews allows for near-epoch-dependent arrays but obtains a slower rate of convergence. Bosq requires geometric strong mixing, a much stronger moment bound, and a specific choice for the bandwidth parameter.

3.3. Local Linear Regression

The local linear estimator of $m(x) = \mathrm{E}(Y_i | X_i = x)$ and its derivative $m^{(1)}(x)$ are obtained from a weighted regression of Y_i on $X_i - x_i$. Letting $k_i = k((x - X_i)/h)$ and $\xi_i = X_i - x$, the local linear estimator can be written as

$$\begin{pmatrix} \tilde{m}(x) \\ \tilde{m}^{(1)}(x) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n k_i & \sum_{i=1}^n \xi_i' k_i \\ \sum_{i=1}^n \xi_i k_i & \sum_{i=1}^n \xi_i \xi_i' k_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n k_i Y_i \\ \sum_{i=1}^n \xi_i k_i Y_i \end{pmatrix}.$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Let $k(u)$ be a multivariate symmetric kernel function for which $\int |u|^4 |k(u)| du < \infty$ and the functions $k(u)$, $uk(u)$, and $uu'k(u)$ satisfy Assumptions 1 and 3.

THEOREM 10. *Under the conditions of Theorem 8 and* $\delta_n^{-2} a_n^* \to 0$ *where* a_n^* *is defined in (29) then*

$$\sup_{|x| \le c_n} |\widetilde{m}(x) - m(x)| = O_p(\delta_n^{-2} a_n^*).$$

THEOREM 11. *Under the conditions of Theorem 9 and* $\delta_n^{-2} a_n^* \to 0$ *where* a_n^* *is defined in (29) then*

$$\sup_{|x| \le c_n} |\widetilde{m}(x) - m(x)| = O(\delta_n^{-2} a_n^*)$$

almost surely.

These are the same rates as for the Nadaraya–Watson estimator, except the penalty term for expanding c_n has been strengthened to δ_n^{-2}. When c_n is fixed the convergence rate is Stone's optimal rate.

Alternative uniform convergence results for pth-order local polynomial estimators with fixed c_n have been provided by Masry (1996) and Fan and Yao (2003, Thm. 6.5). Fan and Yao restrict attention to $d = 1$. Masry allows $d \ge 1$ but assumes that $(p + 1)$ derivatives of $m(x)$ are uniformly bounded (second derivatives in the case of local linear estimation). Instead, we assume that the second derivatives of the product $f(x)m(x)$ are uniformly bounded, which is less restrictive for the case of local linear estimation.

REFERENCES

Andrews, D.W.K. (1995) Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11, 560–596.

Ango Nze, P. & P. Doukhan (2004) Weak dependence: Models and applications to econometrics. *Econometric Theory* 20, 995–1045.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, 2nd ed. Lecture Notes in Statistics 110. Springer-Verlag.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiency. *Annals of Statistics* 21, 196–216.

Fan, J. & Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag.

Granovsky, B.L. & H.-G. Müller (1991) Optimizing kernel methods: A unifying variational principle. *International Statistical Review* 59, 373–388.

Liebscher, E. (1996) Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications* 65, 69–80.

Mack, Y.P. & B.W. Silverman (1982) Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 61, 405–415.

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Marron, J.S. & M.P. Wand (1992) Exact mean integrated squared error. *Annals of Statistics* 20, 712–736.
Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17, 571–599.
Müller, H.-G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics* 12, 766–774.
Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and Its Applications* 9, 141–142.
Newey, W.K. (1994) Kernel estimation of partial means and a generalized variance estimator. *Econometric Theory* 10, 233–253.
Peligrad, M. (1991) Properties of uniform consistency of the kernel estimators of density and of regression functions under dependence conditions. *Stochastics and Stochastic Reports* 40, 147–168.
Rio, E. (1995) The functional law of the iterated logarithm for stationary strongly mixing sequences. *Annals of Probability* 23, 1188–1203.
Rosenblatt, M. (1956) Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics* 27, 832–837.
Stone, C.J. (1977) Consistent nonparametric regression. *Annals of Statistics* 5, 595–645.
Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10, 1040–1053.
Tran, L.T. (1994) Density estimation for time series by histograms. *Journal of Statistical Planning and Inference* 40, 61–79.
Wand, M.P. & W.R. Schucany (1990) Gaussian-based kernels. *Canadian Journal of Statistics* 18, 197–204.
Watson, G.S. (1964) Smooth regression analysis. *Sankya, Series A* 26, 359–372.

APPENDIX

Proof of Theorem 1. We start with some preliminary bounds. First note that Assumption 1 implies that for any $r \leq s$,

$$\int_{R^d} |K(u)|^r \, du \leq \bar{K}^{r-1}\mu \leq \bar{K}^{s-1}\mu. \tag{A.1}$$

Second, assuming without loss of generality that $B_0 \geq 1$ and $B_1 \geq 1$, note that the L^r inequality, (5), and (6) imply that for any $1 \leq r \leq s$

$$\begin{aligned} \mathrm{E}(|Y_0|^r | X_0 = x) f(x) &\leq (\mathrm{E}(|Y_0|^s | X_0 = x))^{r/s} f(x) \\ &= (\mathrm{E}(|Y_0|^s | X_0 = x) f(x))^{r/s} f(x)^{(s-r)/s} \\ &\leq B_1^{r/s} B_0^{(s-r)/s} \\ &\leq B_1 B_0. \end{aligned} \tag{A.2}$$

Third, for fixed x and h let

$$Z_i = K\left(\frac{x - X_i}{h}\right) Y_i.$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Then for any $1 \le r \le s$, by iterated expectations, (A.2), a change of variables, and (A.1)

$$\begin{aligned} h^{-d}\mathrm{E}|Z_0|^r &= h^{-d}\mathrm{E}\left(\mathrm{E}\left(\left|K\left(\frac{x-X_0}{h}\right)Y_0\right|^r|X_0\right)\right) \\ &= h^{-d}\int_{R^d}\left|K\left(\frac{x-u}{h}\right)\right|^r \mathrm{E}(|Y_0|^r|X_0=u)f(u)\,du \\ &= \int_{R^d}|K(u)|^r\mathrm{E}(|Y_0|^r|X_0=x-hu)f(x-hu)\,du \\ &\le \int_{R^d}|K(u)|^r\,du B_1 B_0 \\ &\le \bar{K}^{s-1}\mu B_1 B_0 \\ &\equiv \bar{\mu} < \infty. \end{aligned} \tag{A.3}$$

Finally, for $j \ge j^*$, by iterated expectations, (7), two changes of variables, and Assumption 1,

$$\begin{aligned} \mathrm{E}|Z_0 Z_j| &= \mathrm{E}\left(\mathrm{E}\left(\left|K\left(\frac{x-X_0}{h}\right)K\left(\frac{x-X_j}{h}\right)Y_0Y_j\right||X_0,X_j\right)\right) \\ &= \int_{R^d}\int_{R^d}\left|K\left(\frac{x-u_0}{h}\right)K\left(\frac{x-u_j}{h}\right)\right|\mathrm{E}(|Y_0Y_j||X_0=u_0,X_j=u_j) \\ &\quad \times f_j(u_0,u_j)\,du_0\,du_j \\ &= \int_{R^d}\int_{R^d}|K(u_0)K(u_j)|\mathrm{E}(|Y_0Y_j||X_0=x-hu_0,X_j=x-hu_j) \\ &\quad \times f_j(x-hu_0,x-hu_j)\,du_0\,du_j \\ &\le h^{2d}\int_{R^d}\int_{R^d}|K(u_0)K(u_j)|\,du_0\,du_j B_2 \\ &\le h^{2d}\mu^2 B_2. \end{aligned} \tag{A.4}$$

Define the covariances

$$C_j = \mathrm{E}((Z_0 - \mathrm{E}Z_0)(Z_j - \mathrm{E}Z_j)).$$

Assume that n is sufficiently large so that $h^{-d} \ge j^*$. We now bound the C_j separately for $j \le j^*$, $j^* < j \le h^{-d}$, and $h^{-d}+1 < j < \infty$.

First, for $j \le j^*$, by the Cauchy–Schwarz inequality and (A.3) with $r = 2$,

$$|C_j| \le \mathrm{E}(Z_0 - \mathrm{E}Z_0)^2 \le \mathrm{E}Z_0^2 \le \bar{\mu}h^d.$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Second, for $j^* < j \leq h^{-d}$, (A.4) and (A.3) for $r = 1$ combine to yield

$$|C_j| \leq \mathrm{E}|Z_0 Z_j| + (\mathrm{E}|Z_0|)^2 \leq (\mu^2 B_2 + \bar{\mu}^2) h^{2d}.$$

Third, for $j > h^{-d} + 1$, using Davydov's lemma, (2), and (A.3) with $r = s$ we obtain

$$\begin{aligned} |C_j| &\leq 6\alpha_j^{1-2/s} (\mathrm{E}|Z_i|^s)^{2/s} \\ &\leq 6Aj^{-\beta(1-2/s)} (\bar{\mu} h^d)^{2/s} \\ &\leq 6A\bar{\mu}^{2/s} j^{-(2-2/s)} h^{2d/s}, \end{aligned}$$

where the final inequality uses (4).

Using these three bounds, we calculate that

$$\begin{aligned} nh^d \mathrm{Var}(\hat{\Psi}(x)) &= \frac{1}{n} \mathrm{E}\left(\sum_{i=1}^{n} Z_i - \mathrm{E}Z_i \right)^2 \\ &\leq C_0 + 2 \sum_{j=1}^{j^*} |C_j| + 2 \sum_{j=j^*+1}^{h^{-d}} |C_j| + 2 \sum_{j=h^{-d}+1}^{\infty} |C_j| \\ &\leq (1 + 2j^*) \bar{\mu} h^d + 2 \sum_{j=j^*+1}^{h^{-d}} (\mu^2 B_2 + \bar{\mu}^2) h^{2d} \\ &\quad + 2 \sum_{j=h^{-d}+1}^{\infty} 6A\bar{\mu}^{2/s} j^{-(2-2/s)} h^{2d/s} \\ &\leq (1 + 2j^*) \bar{\mu} h^d + 2(\mu^2 B_2 + \bar{\mu}^2) h^d + \frac{12A\bar{\mu}^{2/s}}{(s-2)/s} h^d, \end{aligned}$$

where the final inequality uses the fact that for $\delta > 1$ and $k \geq 1$

$$\sum_{j=k+1}^{\infty} j^{-\delta} \leq \int_k^{\infty} x^{-\delta} \, dx = \frac{k^{1-\delta}}{(\delta - 1)}.$$

We have shown that (8) holds with

$$\Theta = \left((1 + 2j^*) \bar{\mu} + 2(\mu^2 B_2 + \bar{\mu}^2) + \frac{12A\bar{\mu}^{2/s} s}{s-2} \right), \tag{A.5}$$

completing the proof. ■

Before giving the proof of Theorem 2 we restate Theorem 2.1 of Liebscher (1996) for stationary processes, which is derived from Theorem 5 of Rio (1995).

LEMMA (Liebscher/Rio). *Let Z_i be a stationary zero-mean real-valued process such that $|Z_i| \leq b$, with strong mixing coefficients α_m. Then for each positive integer $m \leq n$ and ε such that $m < \varepsilon b/4$*

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \varepsilon\right) \le 4\exp\left(-\frac{\varepsilon^2}{64\,\dfrac{n\sigma_m^2}{m} + \dfrac{8}{3}\,\varepsilon m b}\right) + 4\,\frac{n}{m}\,\alpha_m,$$

where $\sigma_m^2 = \mathrm{E}(\sum_{i=1}^{m} Z_i)^2$.

Proof of Theorem 2. We first note that (10) implies that θ defined in (11) satisfies $\theta > 0$, so that (12) allows $h = o(1)$ as required.

The proof is organized as follows. First, we show that we can replace Y_i with the truncated process $Y_i 1(|Y_i| \le \tau_n)$ where $\tau_n = a_n^{-1/(s-1)}$. Second, we replace the the supremum in (15) with a maximization over a finite N-point grid. Third, we use the exponential inequality of the lemma to bound the remainder. The second and third steps are a modification of the strategy of Liebscher (1996, proof of Thm. 4.2).

The first step is to truncate Y_i. Define

$$\begin{aligned} R_n(x) &= \hat{\Psi}(x) - \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) 1(|Y_i| \le \tau_n) \\ &= \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) 1(|Y_i| > \tau_n). \end{aligned} \tag{A.6}$$

Then by a change of variables, using the region of integration, (6), and Assumption 1

$$\begin{aligned} |\mathrm{E}R_n(x)| &\le \frac{1}{h^d}\int_{R^d} \left|K\left(\frac{x-u}{h}\right)\right| \mathrm{E}(|Y_0| 1(|Y_0| > \tau_n)|X_0 = u) f(u)\,du \\ &= \int_{R^d} |K(u)|\mathrm{E}(|Y_0| 1(|Y_0| > \tau_n)|X_0 = x - hu) f(x - hu)\,du \\ &\le \int_{R^d} |K(u)|\mathrm{E}(|Y_0|^s \tau_n^{-(s-1)} 1(|Y_0| > \tau_n)|X_0 = x - hu) f(x - hu)\,du \\ &\le \tau_n^{-(s-1)} \int_{R^d} |K(u)|\mathrm{E}(|Y_0|^s|X_0 = x - hu) f(x - hu)\,du \\ &\le \tau_n^{-(s-1)}\,\mu B_1. \end{aligned} \tag{A.7}$$

By Markov's inequality and the definition of τ_n

$$|R_n(x) - \mathrm{E}R_n(x)| = O_p(\tau_n^{-(s-1)}) = O_p(a_n),$$

and therefore replacing Y_i with $Y_i 1(|Y_i| \le \tau_n)$ results in an error of order $O_p(a_n)$. For the remainder of the proof we simply assume that $|Y_i| \le \tau_n$.

For the second step we create a grid using regions of the form $A_j = \{x : \|x - x_j\| \le a_n h\}$. By selecting x_j to lay on a grid, the region $\{x : \|x\| \le c_n\}$ can be covered with $N \le c_n^d h^{-d} a_n^{-d}$ such regions A_j. Assumption 3 implies that for all $|x_1 - x_2| \le \delta \le L$,

$$|K(x_2) - K(x_1)| \le \delta K^*(x_1), \tag{A.8}$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

where $K^*(u)$ satisfies Assumption 1. Indeed, if $K(u)$ has compact support and is Lipschitz then $K^*(u) = \Lambda_1 1(\|u\| \leq 2L)$. On the other hand, if $K(u)$ satisfies the differentiability conditions of Assumption 3, then $K^*(u) = \Lambda_1(1(\|u\| \leq 2L) + \|u - L\|^{-\eta} 1(\|u\| > 2L))$. In both cases $K^*(u)$ is bounded and integrable and therefore satisfies Assumption 1.

Note that for any $x \in A_j$ then $\|x - x_j\|/h \leq a_n$, and equation (A.8) implies that if n is large enough so that $a_n \leq L$,

$$\left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x_j - X_i}{h}\right)\right| \leq a_n K^*\left(\frac{x_j - X_i}{h}\right).$$

Now define

$$\tilde{\Psi}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K^*\left(\frac{x - X_i}{h}\right), \tag{A.9}$$

which is a version of $\hat{\Psi}(x)$ with $K(u)$ replaced with $K^*(u)$. Note that

$$\mathrm{E}|\tilde{\Psi}(x)| \leq B_1 B_0 \int_{R^d} K^*(u)\, du < \infty.$$

Then

$$\begin{aligned}\sup_{x\in A_j}|\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| &\leq |\hat{\Psi}(x_j) - \mathrm{E}\hat{\Psi}(x_j)| + a_n[|\tilde{\Psi}(x_j)| + \mathrm{E}|\tilde{\Psi}(x_j)|] \\ &\leq |\hat{\Psi}(x_j) - \mathrm{E}\hat{\Psi}(x_j)| + a_n|\tilde{\Psi}(x_j) - \mathrm{E}\tilde{\Psi}(x_j)| + 2a_n\mathrm{E}|\tilde{\Psi}(x_j)| \\ &\leq |\hat{\Psi}(x_j) - \mathrm{E}\hat{\Psi}(x_j)| + |\tilde{\Psi}(x_j) - \mathrm{E}\tilde{\Psi}(x_j)| + 2a_n M,\end{aligned}$$

the final inequality because $a_n \leq 1$ for n sufficiently large and for any $M > \mathrm{E}|\tilde{\Psi}(x)|$.

We find that

$$\begin{aligned}&P\left(\sup_{\|x\|\leq c_n}|\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| > 3Ma_n\right) \\ &\quad\leq N\max_{1\leq j\leq N} P\left(\sup_{x\in A_j}|\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| > 3Ma_n\right) \\ &\quad\leq N\max_{1\leq j\leq N} P(|\hat{\Psi}(x_j) - \mathrm{E}\hat{\Psi}(x_j)| > M) \qquad \text{(A.10)} \\ &\qquad + N\max_{1\leq j\leq N} P(|\tilde{\Psi}(x_j) - \mathrm{E}\tilde{\Psi}(x_j)| > M). \qquad \text{(A.11)}\end{aligned}$$

We now bound (A.10) and (A.11) using the same argument, as both $K(u)$ and $K^*(u)$ satisfy Assumption 1, and this is the only property we will use.

Let $Z_i(x) = Y_i K((x - X_i)/h) - \mathrm{E}Y_i K((x - X_i)/h)$. Because $|Y_i| \leq \tau_n$ and $|K((x - X_i)/h)| \leq \bar{K}$ it follows that $|Z_i(x)| \leq 2\tau_n\bar{K} \equiv b_n$. Also from Theorem 1 we have (for n sufficiently large) the bound

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

$$\sup_x \mathrm{E}\left(\sum_{i=1}^{m} Z_i(x)\right)^2 \le \Theta m h^d.$$

Set $m = a_n^{-1}\tau_n^{-1}$ and note that $m < n$ and $m < \varepsilon b_n/4$ for $\varepsilon = Ma_n nh^d$ for n sufficiently large. Then by the lemma, for any x, and n sufficiently large,

$$\begin{aligned} P(|\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| > Ma_n) &= P\left(\left|\sum_{i=1}^{n} Z_i(x)\right| > Ma_n nh^d\right) \\ &\le 4\exp\left(-\frac{M^2 a_n^2 n^2 h^{2d}}{64\Theta nh^d + 6\bar{K}Mnh^d}\right) + 4\frac{n}{m}\alpha_m \\ &\le 4\exp\left(-\frac{M^2 \ln n}{64\Theta + 6\bar{K}M}\right) + 4Anm^{-1-\beta} \\ &\le 4n^{-M/(64+6\bar{K})} + 4Ana_n^{1+\beta}\tau_n^{1+\beta}, \end{aligned}$$

the second inequality using (2) and (14) and the last inequality taking $M > \Theta$. Recalling that $N \le c_n^d h^{-d} a_n^{-d}$, it follows from this and (A.10)–(A.11) that

$$P\left(\sup_{\|x\|\le c_n} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| > 3Ma_n\right) \le O(T_{1n}) + O(T_{2n}), \tag{A.12}$$

where

$$T_{1n} = c_n^d h^{-d} a_n^{-d} n^{-M/(64+6\bar{K})} \tag{A.13}$$

and

$$T_{2n} = c_n^d h^{-d} n a_n^{1+\beta-d} \tau_n^{1+\beta}. \tag{A.14}$$

Recall that $\tau_n = a_n^{-1/(s-1)}$ and $c_n = O((\ln n)^{1/d} n^{1/2q})$. Equation (12) implies that $(\ln n)h^{-d} = o(n^\theta)$ and thus $c_n^d h^{-d} = o(n^{d/2q+\theta})$. Also

$$a_n = ((\ln n)h^{-d}n^{-1})^{1/2} \le o(n^{-(1-\theta)/2}).$$

Thus

$$T_{1n} = o(n^{d/2q+\theta+d(1-\theta)/2-M/(64+6\bar{K})}) = o(1)$$

for sufficiently large M and

$$T_{2n} = o(n^{d/2q+\theta+1-(1-\theta)[1+\beta-d-(1+\beta)/(s-1)]/2}) = o(1)$$

by (11). Thus (A.12) is $o(1)$, which is sufficient for (15). ■

Proof of Theorem 3. We first note that (16) implies that θ defined in (17) satisfies $\theta > 0$, so that (18) allows $h = o(1)$ as required.

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

The proof is a modification of the proof of Theorem 2. Borrowing an argument from Mack and Silverman (1982), we first show that $R_n(x)$ defined in (A.6) is $O(a_n)$ when we set $\tau_n = (n\phi_n)^{1/s}$. Indeed, by (A.7) and $s > 2$,

$$|\mathrm{E}R_n(x)| \le \tau_n^{-(s-1)}\mu B_1 \le n^{-(s-1)s}\mu B_1 = O(a_n),$$

and because

$$\sum_{n=1}^{\infty} P(|Y_n| > \tau_n) \le \sum_{n=1}^{\infty} \tau_n^{-s}\mathrm{E}|Y_n|^s \le \mathrm{E}|Y_0|^s \sum_{n=1}^{\infty} (n\phi_n)^{-1} < \infty,$$

using the fact that $\sum_{n=1}^{\infty}(n\phi_n)^{-1} < \infty$, then for sufficiently large n, $|Y_n| \le \tau_n$ with probability one. Hence for sufficiently large n and all $i \le n$, $|Y_i| \le \tau_n$, and thus $R_n(x) = 0$ with probability one. We have shown that

$$|R_n(x) - \mathrm{E}R_n(x)| = O(a_n)$$

almost surely. Thus, as in the proof of Theorem 2 we can assume that $|Y_i| \le \tau_n$.

Equations (A.12)–(A.14) hold with $\tau_n = (n\phi_n)^{1/s}$ and $c_n = O(\phi_n^{1/d} n^{1/2q})$. Employing $h^{-d} = O(\phi_n^{-2} n^{\theta})$ and $r_n \le o(\phi_n^{-1/2} n^{-(1-\theta)/2})$ we find

$$\begin{aligned} T_{1n} &= c_n^d h^{-d} r_n^{-d} n^{-M/(64+6\bar{K})} \\ &= o(\phi_n^{-1} n^{d/2q+\theta+d(1-\theta)/2-M/(64+6\bar{K})}) \\ &= o((n\phi_n)^{-1}) \end{aligned}$$

for sufficiently large M and

$$\begin{aligned} T_{2n} &= c_n^d h^{-d} n a_n^{1+\beta-d} \tau_n^{1+\beta} \\ &= O(\phi_n^{-1-(1+\beta-d)/2+(1+\beta)/s} n^{d/2q+\theta+1-(1-\theta)(1+\beta-d)/2+(1+\beta)/s}) \\ &= O((n\phi_n)^{-1}) \end{aligned}$$

by (17) and the fact that $(1 + \beta)/s < (1 + \beta - d)/2$ is implied by (16). Thus

$$\sum_{n=1}^{\infty} (T_{1n} + T_{2n}) < \infty.$$

It follows from this and (A.12) that

$$\sum_{n=1}^{\infty} P\left(\sup_{\|x\| \le c_n} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| > 3Ma_n \right) < \infty,$$

and (20) follows by the Borel–Cantelli lemma. ■

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Proof of Theorem 4. Define $c_n = n^{1/2q}$ and

$$\tilde{\Psi}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) 1(\|X_i\| \le c_n). \tag{A.15}$$

Observe that $c_n^{-q} \le O(a_n)$. Using the region of integration, a change of variables, (21), and Assumption 1,

$$\begin{aligned}
|\mathrm{E}(\hat{\Psi}(x) - \tilde{\Psi}(x))| &\le h^{-d}\mathrm{E}\left(|Y_0|\left|K\left(\frac{x - X_0}{h}\right)\right| 1(\|X_0\| > c_n)\right) \\
&= h^{-d}\int_{\|u\|>c_n} \mathrm{E}(|Y_0||X_0 = u)\left|K\left(\frac{x-u}{h}\right)\right| f(u)\,du \\
&\le h^{-d}c_n^{-q}\int_{R^d} \|u\|^q \mathrm{E}(|Y_0||X_0 = u)\left|K\left(\frac{x-u}{h}\right)\right| f(u)\,du \\
&= c_n^{-q}\int_{R^d} \|x - hu\|^q \mathrm{E}(|Y_0||X_0 = x - hu) f(x - hu)|K(u)|\,du \\
&\le c_n^{-q} B_3 \mu \\
&= O(a_n).
\end{aligned} \tag{A.16}$$

By Markov's inequality

$$\sup_x |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| = \sup_x |\tilde{\Psi}(x) - \mathrm{E}\tilde{\Psi}(x)| + O_p(a_n). \tag{A.17}$$

This shows that the error in replacing $\hat{\Psi}(x)$ with $\tilde{\Psi}(x)$ is $O_p(a_n)$.

Suppose that $c_n > L$, $\|x\| > 2c_n$, and $\|X_i\| \le c_n$. Then $\|x - X_i\| \ge c_n$, and (22) and $q \ge d$ imply that

$$K\left(\frac{x - X_i}{h}\right) \le \Lambda_2 \left\|\frac{x - X_i}{h}\right\|^{-q} \le \Lambda_2 h^q c_n^{-q} \le \Lambda_2 h^d c_n^{-q}.$$

Therefore

$$\begin{aligned}
\sup_{\|x\|>2c_n} |\tilde{\Psi}(x)| &\le \frac{1}{nh^d}\sum_{i=1}^{n} |Y_i| \sup_{\|x\|>2c_n}\left|K\left(\frac{x - X_i}{h}\right)\right| 1(\|X_i\| \le c_n) \\
&\le \frac{1}{n}\sum_{i=1}^{n} |Y_i| \Lambda_2 c_n^{-q} \\
&= O(a_n)
\end{aligned}$$

and

$$\sup_{\|x\|>2c_n} |\tilde{\Psi}(x) - \mathrm{E}\tilde{\Psi}(x)| = O(a_n) \tag{A.18}$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

almost surely. Theorem 2 implies that

$$\sup_{\|x\| \le 2c_n} |\tilde{\Psi}(x) - \mathrm{E}\tilde{\Psi}(x)| = O_p(a_n). \tag{A.19}$$

Equations (A.17)–(A.19) together establish the result. ■

Proof of Theorem 5. Let $c_n = (n\phi_n)^{1/2q}$ and let $\tilde{\Psi}(x)$ be defined as in (A.15). Because $\mathrm{E}|X_i|^{2q} < \infty$, by the same argument as at the beginning of the proof of Theorem 3, for n sufficiently large $\hat{\Psi}(x) = \tilde{\Psi}(x)$ with probability one. This and (A.16) imply that the error in replacing $\hat{\Psi}(x)$ with $\tilde{\Psi}(x)$ is $O(c_n^{-q}) \le O(a_n)$.

Furthermore, equation (A.18) holds. Theorem 3 applies because $1/2q \le 1/d$ implies $c_n = O(\phi_n^{1/d} n^{1/2q})$. Thus

$$\sup_x |\tilde{\Psi}(x) - \mathrm{E}\tilde{\Psi}(x)| = O(a_n)$$

almost surely. Together, this completes the proof. ■

Proof of Theorem 6. In the notation of Section 2, $\hat{f}(x) = h^{-r}\hat{\Psi}(x)$ with $K(x) = k^{(r)}(x)$ and $Y_i = 1$. Assumptions 1–3 are satisfied with $s = \infty$; thus by Theorem 2

$$\begin{aligned}
\sup_{\|x\| \le c_n} |\hat{f}^{(r)}(x) - \mathrm{E}\hat{f}^{(r)}(x)| &= h^{-r} \sup_{\|x\| \le c_n} |\hat{\Psi}(x) - \mathrm{E}\hat{\Psi}(x)| \\
&= O_p\left(h^{-r}\left(\frac{\log n}{nh^d}\right)^{1/2}\right) \\
&= O_p\left(\left(\frac{\log n}{nh^{d+2r}}\right)^{1/2}\right).
\end{aligned}$$

By integration by parts and a change of variables,

$$\begin{aligned}
\mathrm{E}\hat{f}^{(r)}(x) &= \frac{1}{h^{d+r}} \mathrm{E}\left(k^{(r)}\left(\frac{x - X_i}{h}\right)\right) \\
&= \frac{1}{h^{d+r}} \int k^{(r)}\left(\frac{x-u}{h}\right) f(u)\, du \\
&= \frac{1}{h^d} \int k\left(\frac{x-u}{h}\right) f^{(r)}(u)\, du \\
&= \int k(u) f^{(r)}(x - hu)\, du \\
&= f(x) + O(h^p),
\end{aligned}$$

where the final equality is by a pth-order Taylor series expansion and using the assumed properties of the kernel and $f(x)$. Together we obtain (25). Equation (27) is obtained by setting $h = (\ln n/n)^{1/(d+2p+2r)}$, which is allowed when $\theta = d/(d + 2p + 2r)$. ■

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Proof of Theorem 7. The argument is the same as for Theorem 6, except that Theorem 3 is used so that the convergence holds almost surely. ■

Proof of Theorem 8. Set $g(x) = m(x)f(x)$, $\hat{g}(x) = (nh^d)^{-1}\sum_{i=1}^n Y_i k((x - X_i)/h)$, and $\hat{f}(x) = (nh^d)^{-1}\sum_{i=1}^n k((x - X_i)/h)$. We can write

$$\hat{m}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} = \frac{\hat{g}(x)/f(x)}{\hat{f}(x)/f(x)}. \tag{A.20}$$

We examine the numerator and denominator separately.

First, Theorem 6 shows that

$$\sup_{\|x\| \le c_n} |\hat{f}(x) - f(x)| = O_p(a_n^*)$$

and therefore

$$\sup_{\|x\| \le c_n} \left| \frac{\hat{f}(x)}{f(x)} - 1 \right| = \sup_{\|x\| \le c_n} \left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| \le \frac{O_p(a_n^*)}{\inf_{|x| \le c_n} f(x)} \le O_p(\delta_n^{-1} a_n^*).$$

Second, an application of Theorem 2 yields

$$\sup_{\|x\| \le c_n} |\hat{g}(x) - \mathrm{E}\hat{g}(x)| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2}\right).$$

We calculate that

$$\begin{aligned}
\mathrm{E}\hat{g}(x) &= \frac{1}{h^d}\mathrm{E}\left(\mathrm{E}(Y_0|X_0)k\left(\frac{x - X_0}{h}\right)\right) \\
&= \frac{1}{h^d}\int_{R^d} k\left(\frac{x - u}{h}\right) m(u)f(u)\,du \\
&= \int_{R^d} k(u)g(x - hu)\,du \\
&= g(x) + O(h^2)
\end{aligned}$$

and thus

$$\sup_{\|x\| \le c_n} |\hat{g}(x) - g(x)| = O_p(a_n^*).$$

This and $g(x) = m(x)f(x)$ imply that

$$\sup_{\|x\| \le c_n} \left| \frac{\hat{g}(x)}{f(x)} - m(x) \right| \le \frac{O_p(a_n)}{\inf_{|x| \le c_n} f(x)} \le O_p(\delta_n^{-1} a_n^*). \tag{A.21}$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

Together, (A.20) and (A.21) imply that uniformly over $\|x\| \leq c_n$

$$\hat{m}(x) = \frac{\hat{g}(x)/f(x)}{\hat{f}(x)/f(x)} = \frac{m(x) + O_p(\delta_n^{-1} a_n^*)}{1 + O_p(\delta_n^{-1} a_n^*)} = m(x) + O_p(\delta_n^{-1} a_n^*)$$

as claimed.

The optimal rate is obtained by setting $h = (\ln n/n)^{1/(d+4)}$, which is allowed when $\theta = d/(d+4)$, which is implied by (11) for sufficiently large β. ■

Proof of Theorem 9. The argument is the same as for Theorem 8, except that Theorems 3 and 7 are used so that the convergence holds almost surely. ■

Proof of Theorem 10. We can write

$$\tilde{m}(x) = \frac{\hat{g}(x) - S(x)'M(x)^{-1}N(x)}{\hat{f}(x) - S(x)'M(x)^{-1}S(x)},$$

where

$$S(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \left(\frac{x - X_i}{h}\right) k\left(\frac{x - X_i}{h}\right),$$

$$M(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \left(\frac{x - X_i}{h}\right)\left(\frac{x - X_i}{h}\right)' k\left(\frac{x - X_i}{h}\right),$$

$$N(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \left(\frac{x - X_i}{h}\right) k\left(\frac{x - X_i}{h}\right) Y_i.$$

Defining $\Omega = \int_{R^d} uu'k(u)\,du$, Theorem 2 and standard calculations imply that uniformly over $\|x\| \leq c_n$,

$$S(x) = h\Omega f^{(1)}(x) + O_p(a_n^*),$$

$$M(x) = \Omega f(x) + O_p(a_n^*),$$

$$N(x) = h\Omega g^{(1)}(x) + O_p(a_n^*).$$

Therefore because $f^{(1)}(x)$ and $g^{(2)}(x)$ are bounded, uniformly over $\|x\| \leq c_n$,

$$f(x)^{-1}S(x) = O_p(\delta_n^{-1}(h + a_n^*)),$$

$$f(x)^{-1}M(x) = \Omega + O_p(\delta_n^{-1} a_n^*),$$

$$f(x)^{-1}N(x) = O_p(\delta_n^{-1}(h + a_n^*)),$$

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press

and so

$$\frac{S(x)'M(x)^{-1}S(x)}{f(x)} = O_p(\delta_n^{-2}(h + a_n^*)^2) = O_p(\delta_n^{-2} a_n^*)$$

and

$$\frac{S(x)'M(x)^{-1}N(x)}{f(x)} = O_p(\delta_n^{-2} a_n^*).$$

Therefore

$$\tilde{m}(x) = \frac{\dfrac{\hat{g}(x) - S(x)'M(x)^{-1}N(x)}{f(x)}}{\dfrac{\hat{f}(x) - S(x)'M(x)^{-1}S(x)}{f(x)}} = m(x) + O_p(\delta_n^{-2} a_n^*)$$

uniformly over $\|x\| \leq c_n$. ■

Proof of Theorem 11. The argument is the same as for Theorem 10, except that Theorems 3 and 7 are used so that the convergence holds almost surely. ■

https://doi.org/10.1017/S0266466608080304 Published online by Cambridge University Press