Uniform Convergence Rates for Nonparametric Estimation

Bruce E. Hansen[*]
University of Wisconsin[†]

www.ssc.wisc.edu/~bhansen

October 2004

Preliminary and Incomplete

Abstract

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 and kernels with unbounded support.

[*]Research supported by the National Science Foundation.
[†]Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706

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 and kernels with unbounded support.

The kernel estimators that we examine were first introduced by Rosenblatt (1956) for density estimation, 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).

Andrews (1995) provides a comprehensive set of results concerning the uniform consistency of kernel estimators, but his rates are not sharp. Masry (1996) derived sharp rates for uniform almost sure convergence, but confined attention to the case of bounded regression support, and placed overly restrictive conditions on the regression functions. Fan and Yao (2003) also have a set of results, but are quite restrictive in application.

In this paper we attempt to provide a general set of results with broad applicability. Our main result is the uniform convergence of a sample average functional, which can be easily used for application to density and regression estimation. The conditions imposed on the functional are quite general, allowing for kernels with unbounded support (such as the standard normal), so long as they satisfy a Lipschitz condition. The data are allowed to be generated either from a random sample or from a stationary strong mixing time series. The support for the data is allowed to be infinite, and our convergence results include the entire sample space rather than being restricted to a compact subset. The rate of decay for the bandwidth is allowed considerable flexibility.

Our proof method is a generalization of that in Fan and Yao (2003), and like theirs is based on an exponential inequality from Bosq (1998). We also borrow the trimming argument of Andrews (1995) to allow for unbounded regression support.

Section 2 of the paper presents the main results: a variance bounded and the rate of convergence for the sample average functional. Section 3 provides an application to density estimation, and Section 4 to regression estimation. The proofs are in the Appendix.

2 Basic Results

Let $\{X_i, Y_i\} \in R^d \times R$ be a sequence of random vectors. We are interested in averages of the form

$$\hat{G}_h(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i G\left(\frac{x - X_i}{h}\right) \tag{1}$$

where $G(x) : R^d \to R$ and

$$h = cn^{-\gamma} \tag{2}$$

is a bandwidth, where $0 < c < \infty$ and $0 < \gamma < 1/d$. For typical applications, the function G is either a kernel or the product of a kernel with a polynomial.

Assumption 1 *For all $x, x' \in R^d$, some $\Lambda < \infty$ and $\eta > 1/\gamma$ then*

$$|G(x)| \leq \begin{cases} \Lambda & , \quad |x| \leq 1 \\ \Lambda\,|x|^{-\eta} & , \quad |x| > 1 \end{cases}, \tag{3}$$

$$\left|G(x) - G(x')\right| \leq \Lambda\left|x - x'\right|, \tag{4}$$

Equation (3) imposes boundedness and a tail condition related to the bandwidth h. Equation (4) is a Lipschitz condition.

We require the following moment and smoothness conditions on the observations. Let $f_{Y|X}(y \mid x)$ and $f_X(x)$ denote the conditional density of Y_i given X_i, and the marginal density of X_i, respectively, and for any $j \geq 1$ let $f_j(x_0, x_j)$ denote the joint density of (X_0, X_j).

Assumption 2 *$\{X_i, Y_i\}$ is strictly stationary and ergodic, with strong mixing coefficients $\alpha_m \leq am^{-\beta}$ for some $a < \infty$ and*

$$\beta > \frac{2s - 2}{s - 2} \tag{5}$$

for some $s > 2$. Furthermore, $E\,|Y_{i+j}|^s < \infty$, and for all $j \geq 0$ and all $t \leq s$

$$\sup_x E\left(|Y_{i+j}|^t \mid X_i = x\right) f_X(x) \leq \Psi_1 < \infty, \tag{6}$$

for $j \geq d + 1$,

$$\sup_{x_1, x_2} f_j\left(x_1, x_2\right) \leq \Psi_2 < \infty, \tag{7}$$

and for $|x|$ large and some $1 < \mu \leq \infty$

$$E\left(|Y_i| \mid X_i = x\right) f_X(x) \leq \Psi_3\,|x|^{-\mu}. \tag{8}$$

Note that equations (6) and (8) involve the product of the conditional mean and the marginal density, and thus are not very restrictive. For independent data, $f_j(x_0, x_j) = f_X(x_0) f_X(x_j)$, so (7) holds when the density $f_X(x)$ is bounded. If the support of X_i is bounded, then we can take $\mu = \infty$ in (8)

We first describe a uniform bound on the variance of $\hat{G}_h(x)$.

Theorem 1 *Under Assumptions 1 and 2, there is $J < \infty$ such that*

$$Var\left(\hat{G}_h(x)\right) \leq \frac{Jh^d}{n}. \tag{9}$$

We now present our main result concerning the rate of convergence for $\hat{G}_h(x)$.

Theorem 2 *Under Assumptions 1 and 2, if in addition*

$$\beta > \frac{3}{s-2} + \left(d\frac{\mu}{\mu - 1}(1+\gamma d) + \frac{3+7\gamma d}{2}\right)\left(\frac{s}{(s-2)(1-\gamma d)}\right) \tag{10}$$

then

$$\sup_{x \in R^d}\left|\hat{G}_h(x) - E\hat{G}_h(x)\right| = O_p\left(h^d r_n\right) \tag{11}$$

where

$$r_n = \left(\frac{\log n}{h^d n}\right)^{1/2}.$$

The rate for the bandwidth is controlled by (10) which is satisfied for sufficiently large β. In particular, in the case of independent data or exponential decay for the mixing coefficients, we have $\beta = \infty$ and (10) is automatically satisfied.

The right side of equation (10) is increasing in γ, so the inequality is least restrictive by selecting γ close to zero (so the bandwidth h declines to zero slowly). The limiting case ($\gamma = 0$) is

$$\beta > \frac{3}{s-2} + \left(d\frac{\mu}{\mu-1} + \frac{3}{2}\right)\left(\frac{s}{s-2}\right)$$

Furthermore, if Y_i has all moments finite and the support of X_i is bounded, then $s = \infty$ and $\mu = \infty$, and this simplies to

$$\beta > d + \frac{3}{2}.$$

In particular, for $d = 1$, then the inequality is $\beta > 5/2$, the restriction used by Fan and Yao (2003), Lemma 6.1. Thus Theorem 2 generalizes their result to the case of mutlivariate data with unbounded support.

3 Density Estimation

Consider the estimation of $f_X(x)$, the density of X_i. Let $k(u) : R \to R$ denote a kernel function and let

$$K(x) = \prod_{j=1}^{d} k(x_j)$$

be a product kernel. Let h be a bandwidth. The kernel density estimator of $f_X(x)$ is

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

It is asymptotically optimal to set $h = cn^{-\gamma}$ with $\gamma = 1/(d+4)$, which we now assume.

Assumption 3 *(a)* $\int_R k(u)du = 1$; *(b)* $k(-u) = k(u)$; $k(u)$ *satisfies (3) and (4) for some* $\eta > d+4$. *The bandwidth satisfies* $h = cn^{-1/(d+4)}$.

We can use Theorem 2 to obtain the uniform rate of convergence for $f_X(x)$.

Theorem 3 *If Assumption* 2 *holds with* $Y_i = 1$ *and* $s = \infty$, *Assumption 3 holds, and*

$$\beta > \frac{1}{4}\left(d\frac{\mu}{\mu-1}(2d+4) + 6 + 5d\right) \tag{12}$$

then

$$\sup_{x \in R^d} \left| \hat{f}_X(x) - f_X(x) \right| = O_p(r_n)$$

where

$$r_n = n^{-2/(d+4)} \log^{1/2} n$$

Alternative results for the uniform rate of convergence for kernel density estimates have been provided by Andrews (1995, Theorem 1) and Fan and Yao (2003, Theorem 5.3). Andrews' result is more general in allowing for near-epoch-dependent arrays but obtains a slower rate of convergence. Fan and Yao obtained the same rate of convergence, but their result is restricted to univariate data, compact support for X_i, and kernels with compact support.

4 Nadaraya-Watson Estimator

Consider the estimation of the conditional mean

$$m(x) = E\left(Y_i \mid X_i = x\right).$$

Let the multivariate kernel K and bandwidth h be defined as in the previous section.

The Nadaraya-Watson estimator of $m(x)$ is

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

The local linear (LL) estimator of $m(x)$ is obtained from a weighted regression of Y_i on $x - X_i$. Letting

$$z_i = \begin{pmatrix} 1 \\ \frac{x-X_i}{h} \end{pmatrix}$$

then the LL estimator is

$$\tilde{m}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right) - \sum_{i=1}^n z_i' K\left(\frac{x-X_i}{h}\right)\left(\sum_{i=1}^n z_i z_i K\left(\frac{x-X_i}{h}\right)\right)^{-1} \sum_{i=1}^n z_i Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) - \sum_{i=1}^n z_i' K\left(\frac{x-X_i}{h}\right)\left(\sum_{i=1}^n z_i z_i K\left(\frac{x-X_i}{h}\right)\right)^{-1} \sum_{i=1}^n z_i K\left(\frac{x-X_i}{h}\right)}$$

We introduce the following smoothness condition.

Assumption 4 *For some $\delta > 0$,*

$$\sup_{|x_1 - x_2| \le \delta} \left|\nabla m(x_1)' \nabla f_X(x_2)\right| < \infty$$

$$\sup_{|x_1 - x_2| \le \delta} \left|\nabla^2 m(x_1) f_X(x_2)\right| < \infty$$

Observe that Assumption 4 does not require the regression function and its derivatives to be bounded. Rather, $\nabla m(x)$ and $\nabla^2 m(x)$ are required to not diverge faster than $\nabla f_X(x)$ and $f_X(x)$ decline to zero in the tails.

For any positive sequence δ_n define

$$A_n = \left\{x \in R^d : f(x) \ge \delta_n\right\}.$$

Theorem 4 *Under Assumptions 2, 3, 4, and (10),*

$$\sup_{x \in A_n} \left|\hat{m}(x) - m(x)\right| = O_p\left(\delta_n^{-2} r_n\right).$$

Alternative results for the uniform rate of convergence for the Nadaraya-Watson estimator have been provided by Andrews (1995, Theorem 1) . His results allow for near-epoch-dependent arrays but obtains a slower rate of convergence.

Theorem 5 *Under [conditions]*

$$\sup_{x \in A_n} |\tilde{m}(x) - m(x)| = O_p\left(\delta_n^{-2} r_n\right).$$

The conditions and proof are incomplete.

This result complements that of Masry (1996). Masry obtains almost sure convergence over compact sets, but imposed stronger conditions on the regression function.

References

[1] Andrews, Donald W.K. (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11, 560-596.

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

[3] Fan, Jianqing (1992): "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998-1004.

[4] Fan, Jianqing (1993): "Local linear regression smoothers and their minimax efficiency," *Annals of Statistics*, 21, 196-216.

[5] Fan, Jianqing and Qiwei Yao (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag.

[6] Masry, Elias (1996): "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571-599.

[7] Nadaraya, E. A. (1964): "On estimating regression," *Theory of Probability and Its Applications*, 9, 141-142.

[8] Rosenblatt, M. (1956): "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, 27, 832-837.

[9] Stone, C.J. (1977): "Consistent nonparametric regression," *Annals of Statistics*, 5, 595-645.

[10] Watson, G.S. (1964): "Smooth regression analysis," *Sankya Series A*, 26, 359-372.

5 Appendix

Proof of Theorem 1. WLOG assume that $\Lambda \geq 1$ and $\Psi \geq 1$. From (3) we observe that

$$\int_{R^d} |G(x)|\, dx \leq \Lambda \int_{|x|\leq 1} dx + \Lambda \int_{|x|>1} |x|^{-\eta}\, dx \leq \frac{\Lambda V_d \eta}{\eta - 1}$$

where V_d is the volumn of the unit sphere in R^d. Since $|G(x)| \leq \Lambda$ It follows that for any $1 \leq t \leq s$

$$\int_{R^d} |G(x)|^t\, dx \leq \Lambda^{t-1} \frac{\Lambda V_d \eta}{\eta - 1} \leq \frac{\Lambda^s V_d \eta}{\eta - 1} \equiv \Theta. \tag{13}$$

Let $U_{ni}(x) = Y_i G\left(\frac{x - X_i}{h}\right)$. By a change of variables, for any $1 \leq t \leq s$, using (13) and (6),

$$\begin{aligned} E\,|U_{ni}(x)|^t &\leq \int\int |y|^t \left| G\left(\frac{x-u}{h}\right)\right|^t f_{Y|X}(y \mid u)\, f_X(u)\, du dy \\ &= h^d \int |G(u)|^t \left(\int |y|^t f_{Y|X}(y \mid x - hu)\, dy\right) f_X(x - hu)\, du \\ &\leq \Theta \Psi_1 h^d. \end{aligned} \tag{14}$$

We now develop several alternative bounds on the covariances $|Cov\,(U_{n0}(x), U_{nj}(x))|$.

First, by the Cauchy-Schwarz inequality and (14) with $t = 2$,

$$\begin{aligned} |Cov\,(U_{n0}(x), U_{nj}(x))| &\leq Var\,(U_{n0}(x)) \\ &\leq 2EU_{n0}(x)^2 \\ &\leq 2\Theta\Psi_1 h^d. \end{aligned} \tag{15}$$

Second, take $j \geq d + 1$ and observe that

$$\begin{aligned} |Cov\,(U_{n0}(x), U_{nj}(x))| \leq\; & E\,|Y_0 Y_j - EY_0 EY_j| \left| G\left(\frac{x - X_0}{h}\right) G\left(\frac{x - X_j}{h}\right)\right| \\ & + (EY_0)^2\, E\left| G\left(\frac{x - X_0}{h}\right) G\left(\frac{x - X_j}{h}\right)\right| + (E\,|U_{n0}(x)|)^2. \end{aligned} \tag{16}$$

We examine the terms on the right-hand-side of (16). By change of variables, (13), and (7),

$$\begin{aligned} E\left| G\left(\frac{x - X_0}{h}\right) G\left(\frac{x - X_j}{h}\right)\right| &= \int\int G\left(\frac{x - u_0}{h}\right) G\left(\frac{x - u_j}{h}\right) f_j(u_0, u_j)\, du_0 du_j \\ &= h^{2d} \int\int G(u_0)\, G(u_j)\, f_j(x - hu_0, x - hu_j)\, du_0 du_j \\ &\leq h^{2d}\Theta^2 \sup_{x_1, x_2} f_j(x_1, x_2) \\ &\leq h^{2d}\Theta^2\Psi_2. \end{aligned} \tag{17}$$

Let $g_j(y_0, y_j \mid x)$ denote the conditional density of (Y_0, Y_j) given $X_0 = x$. Using (3), a change of variables, conditioning, (13), Davydov's Lemma, (6), and the assumption $\alpha_j \leq aj^{-\beta}$, we find

$$\begin{aligned}
& E\left(|Y_0Y_j - EY_0EY_j| \left|G\left(\frac{x-X_0}{h}\right)G\left(\frac{x-X_j}{h}\right)\right|\right) \\
\leq\ & \Lambda E\left(|Y_0Y_j - EY_0EY_j| \left|G\left(\frac{x-X_0}{h}\right)\right|\right) \\
=\ & \Lambda \int\int |y_0y_j - EY_0EY_j| \left|\int G\left(\frac{x-u}{h}\right)\right| g_j(y_0, y_j \mid u) f_X(u) du dy_0 dy_j \\
=\ & h^d \Lambda \int\int |y_0y_j - EY_0EY_j| \int |G(u)| g_j(y_0, y_j \mid x - hu) f_X(x-hu) du dy_0 dy_j \\
\leq\ & h^d \Lambda\Theta \sup_x E(|Y_0Y_j - EY_0EY_j| \mid X_0 = x) f_X(x) \\
\leq\ & h^d \Lambda\Theta \alpha_j^{1-2/s} \sup_x (E(|Y_0|^s \mid X_0 = x) f_X(x))^{1/s} (E(|Y_j|^s \mid X_0 = x) f_X(x))^{1/s} \\
\leq\ & h^d a\Lambda\Theta\Psi_1 j^{-\beta[1-2/s]}.
\end{aligned} \tag{18}$$

Equations (16)-(18) combine to show that for $j \geq d+1$

$$|Cov(U_{n0}(x), U_{nj}(x))| \leq a\Lambda\Theta\Psi_1 j^{-\beta[1-2/s]} h^d + \left((EY_0)^2 \Psi_2 + \Psi_1^2\right)\Theta^2 h^{2d}. \tag{19}$$

Third, using Davydov's Lemma, (14) with $t = s$, and the assumption $\alpha_j \leq aj^{-\beta}$,

$$\begin{aligned}
|Cov(U_{n0}(x), U_{nj}(x))| &\leq 16\alpha_j^{1-2/s} (E|U_{ni}(x)|^s)^{2/s} \\
&\leq 16aj^{-\beta(1-2/s)}\Theta\Psi_1 h^{2d/s} \\
&\leq 16aj^{-\beta(1-2/s)}\Theta\Psi_1 h^{d[2-\beta(1-2/s)]}
\end{aligned} \tag{20}$$

where the final inequality holds since $2/s \geq 2 - \beta(1-2/s)$ under (5).

The bounds (15), (19) and (20) show that

$$\begin{aligned}
nVar\left(\hat{G}_h(x)\right) &= \frac{1}{n} E\left(\sum_{i=1}^n U_{ni}(x) - EU_{ni}(x)\right)^2 \\
&\leq Var(U_{n0}(x)) + 2\sum_{j=1}^d |Cov(U_{n0}(x), U_{nj}(x))| \\
&\quad +2\sum_{j=d+1}^{h^{-d}} |Cov(U_{n0}(x), U_{nj}(x))| \\
&\quad +2\sum_{j=h^{-d}}^{\infty} |Cov(U_{n0}(x), U_{nj}(x))|
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\Theta\Psi_1 h^d (1+2d) \\
&\quad +2\sum_{j=d+1}^{h^{-d}} \left[a\Lambda\Theta\Psi_1 j^{-\beta(1-2/s)}h^d + \left((EY_0)^2\Psi_2+\Psi_1^2\right)\Theta^2 h^{2d}\right] \\
&\quad +2\sum_{j=h^{-d}}^{\infty} 16a\Theta\Psi_1 j^{-\beta(1-2/s)} h^{d[2-\beta(1-2/s)]} \\
&\leq 2\Theta\Psi_1 (1+2d) h^d \\
&\quad +\frac{2a\Lambda\Theta\Psi_1}{\beta\left(1-\frac{2}{s}\right)-1}h^d + 2\left((EY_0)^2\Psi_2+\Psi_1^2\right)\Theta^2 h^d \\
&\quad +\frac{32a\Theta\Psi_1}{\beta\left(1-\frac{2}{s}\right)-1}h^d \\
&\leq Jh^d
\end{aligned}
$$

which is (9) with

$$
J = 2\Theta\Psi_1(1+2d) + \frac{2a\Lambda\Theta\Psi_1}{\beta\left(1-\frac{2}{s}\right)-1} + 2\left((EY_0)^2\Psi_2+\Psi_1^2\right)\Theta^2 + \frac{32a\Theta\Psi_1}{\beta\left(1-\frac{2}{s}\right)-1} \tag{21}
$$

This is (9). For the final inequality we have used the fact that for $\delta > 0$ and $k \geq 1$

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

This completes the proof. ■

Proof of Theorem 2. We start by introducing some notation. First, define

$$
V_{ni}(x) = Y_i G\left(\frac{x-X_i}{h}\right) 1\left(|Y_i| \leq \frac{r_n^{-2/s}}{2}\right)
$$

and

$$
\hat{V}_h(x) = \frac{1}{n}\sum_{i=1}^{n} V_{ni}(x).
$$

$V_{ni}(x)$ has been truncated so that $|V_{ni}(x) - EV_{ni}(x)| \leq r_n^{-2/s}$.

Second, for μ defined in Assumption 2 let $\tau_n = 2^{-1}\left(h^d r_n\right)^{-1/(\mu-1)}$ and set $A = \{x : |x| \leq \tau_n\}$. The region A can be covered with

$$
N = \left(\frac{\tau_n}{h^{1+d}r_n}\right)^d = \frac{h^{-d(1+d\mu/2(\mu-1))}}{2^d}\left(\frac{n}{\log n}\right)^{d\mu/2(\mu-1)} \tag{22}
$$

hyperspheres of the form $A_j = \left\{x : |x - x_j| \leq h^{1+d}r_n\right\}$, which are centered at x_j and have radius $h^{1+d}r_n$. Let $A^c = \{x : |x| > \tau_n\}$.

We show below that

$$\sup_x \left|\hat{G}_h(x) - E\hat{G}_h(x)\right| = \sup_x \left|\hat{V}_h(x) - E\hat{V}_h(x)\right| + O_p(h^d r_n). \tag{23}$$

$$\sup_{x\in A} \left|\hat{V}_h(x) - E\hat{V}_h(x)\right| = \max_{1\le j\le N} \left|\hat{V}_h(x_j) - E\hat{V}_h(x_j)\right| + O_p(h^d r_n) \tag{24}$$

$$\sup_{x\in A^c} \left|\hat{V}_h(x) - E\hat{V}_h(x)\right| = O_p(h^d r_n) \tag{25}$$

$$\max_{1\le j\le N} \left|\hat{V}_h(x_j) - E\hat{V}_h(x_j)\right| = O_p(h^d r_n). \tag{26}$$

Together, these establish that

$$\sup_x \left|\hat{G}_h(x) - E\hat{G}_h(x)\right| = O_p(h^d r_n)$$

as desired.

It remains to show (23)-(26), which we take sequentially.

Proof of (23): First,

$$\begin{aligned} P\left(\left(h^d r_n\right)^{-1} \sup_x \left|\hat{G}_h(x) - \hat{V}_h(x)\right| > 0\right) &\le P\left(\max_{1\le i\le n} |Y_i| > \frac{r_n^{-2/s}}{2}\right) \\ &\le nP\left(|Y_i|^s > \frac{1}{2^s r_n^2}\right) \\ &\le 2^s E\left(|Y_i|^s 1\left(|Y_i|^s > \frac{1}{2^s r_n^2}\right)\right) \\ &\to 0 \end{aligned}$$

using Markov's inequality since $r_n \to 0$ as $n \to \infty$. Hence

$$\sup_x \left|\hat{G}_h(x) - \hat{V}_h(x)\right| = o_p(h^d r_n). \tag{27}$$

Second, by a change of variables and using (13)

$$\begin{aligned} \sup_x E\left|\hat{G}_h(x) - \hat{V}_h(x)\right| &\le \sup_x \int\int_{|y|\ge r_n^{-2/s}/2} \left|G\left(\frac{x-u}{h}\right)\right| |y|\, f_{Y|X}(y \mid u)\, f_X(u)\, dy du \\ &\le h^d \sup_x \int\int_{|y|\ge r_n^{-2/s}/2} |G(u)|\, |y|\, f_{Y|X}(y \mid x-hu)\, f_X(x-hu)\, dy du \\ &\le h^d \int |G(u)|\, du \sup_x \left(\int_{|y|\ge r_n^{-2/s}/2} |y|\, f_{Y|X}(y \mid x)\, dy f_X(x)\right) \\ &\le h^d \Theta \left(\frac{r_n^{-2/s}}{2}\right)^{1-s} \sup_x \left(\int |y|^s f_{Y|X}(y \mid x)\, dy f_X(x)\right) \\ &\le 2^s \Theta \Psi_1 h^d r_n, \end{aligned} \tag{28}$$

the final inequality using (6) and the fact that for $s > 2$ and n large

$$r_n^{\frac{2(s-1)}{s}} \leq r_n.$$

Equation (23) follows from (27) and (28).

Proof of (24): The Lipschitz condition (4) and the radius of A_j imply that for all $x \in A_j$

$$\left| G\left(\frac{x - X_i}{h}\right) - G\left(\frac{x_j - X_i}{h}\right) \right| \leq \Lambda h^{-1} |x - x_j| \leq \Lambda h^d r_n$$

and thus for all $x \in A_j$

$$\begin{aligned}
\left| \hat{V}_h(x) - \hat{V}_h(x_j) \right| &\leq \frac{1}{n} \sum_{i=1}^{n} |Y_i| \left| G\left(\frac{x - X_i}{h}\right) - G\left(\frac{x_j - X_i}{h}\right) \right| \\
&\leq \hat{\mu}_n \Lambda h^d r_n
\end{aligned}$$

where

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} |Y_i| = O_p(1).$$

Similarly

$$\left| E\hat{V}_h(x) - E\hat{V}_h(x_j) \right| \leq E \left| \hat{V}_h(x) - \hat{V}_h(x_j) \right| \leq E |Y_i| \Lambda h^d r_n.$$

Therefore

$$\begin{aligned}
\sup_{x \in A} \left| \hat{V}_h(x) - E\hat{V}_h(x) \right| &= \max_{1 \leq j \leq N} \sup_{x \in A_j} \left| \hat{V}_h(x) - E\hat{V}_h(x) \right| \\
&= \max_{1 \leq j \leq N} \left| \hat{V}_h(x_j) - E\hat{V}_h(x_j) \right| \\
&\quad + \max_{1 \leq j \leq N} \sup_{x \in A_j} \left(\left| \hat{V}_h(x) - \hat{V}_h(x_j) \right| + \left| E\hat{V}_h(x_j) - E\hat{V}_h(x) \right| \right) \\
&\leq \max_{1 \leq j \leq N} \left| \hat{V}_h(x_j) - E\hat{V}_h(x_j) \right| + \left(\hat{\mu}_n + E |Y_i| \right) \Lambda h^d r_n
\end{aligned}$$

which implies (24).

Proof of (25): For η defined in Assumption 1, let $\delta_n = \left(h^d r_n\right)^{-1/\eta}$. Observe that since $\eta > 1/\gamma$ then $\left(1 + \gamma d\right)/2\eta < \gamma$ and

$$h\delta_n = h \left(\frac{n}{h^d \log n}\right)^{1/2\eta} \leq O\left(n^{-\gamma + (1+\gamma d)/2\eta}\right) = o(1). \tag{29}$$

If $|X_i| \leq \tau_n - h\delta_n$ and $x \in A^c = \{x : |x| > \tau_n\}$, then $\left|\frac{x - X_i}{h}\right| \geq \delta_n$ and

$$\left| G\left(\frac{x - X_i}{h}\right) \right| \leq |G(\delta_n)| \leq \Lambda \delta_n^{-\eta} = h^d r_n$$

by (3). Hence

$$E\left(|Y_i|\sup_{x\in A^c}\left|G\left(\frac{x-X_i}{h}\right)\right|1\left(|X_i|\le\tau_n-h\delta_n\right)\right)\le\Lambda h^d r_n E\left|Y_i\right|.$$

Furthermore using (8),

$$\begin{aligned}E\left(|Y_i|\sup_{x\in A^c}\left|G\left(\frac{x-X_i}{h}\right)\right|1\left(|X_i|>\tau_n-h\delta_n\right)\right) &\le E\left(|Y_i|\,1\left(|X_i|>\tau_n-h\delta_n\right)\right)\\ &= \int_{|x|>\tau_n-h\delta_n}E\left(|Y_i|\mid X_i=x\right)f_X(x)dx\\ &\le 2\Psi_3\int_{\tau_n-h\delta_n}^{\infty}|x|^{-\mu}\,dx\\ &\le O\left((\tau_n-h\delta_n)^{1-\mu}\right)\\ &\le O\left(\tau_n^{1-\mu}\right)\\ &\le O(h^d r_n)\end{aligned}$$

the second-to-last inequality using (29) and the final inequality by the definition of τ_n. Thus

$$E\sup_{x\in A^c}\left|\hat{V}_h(x)-E\hat{V}_h(x)\right|\le 2E\left(|Y_i|\sup_{x\in A^c}\left|G\left(\frac{x-X_i}{h}\right)\right|\right)=O(h^d r_n).$$

(25) follows by application of Markov's inequality.

Proof of (26): Define

$$\sigma_m^2(x)=E\left(\sum_{i=1}^{m}\left(V_{ni}(x)-EV_{ni}(x)\right)\right)^2.$$

By Theorem 1 and n sufficiently large,

$$\sigma_m^2(x)\le mJh^d$$

(observe that (5) holds under (10)). Since $|V_{ni}-EV_{ni}|\le r_n^{-2/s}$, by Theorem 1.3 of Bosq (1998), for all x, $q\in(0,1]$ and $\varepsilon>0$

$$\begin{aligned}P\left(\left|\hat{V}_h(x)-E\hat{V}_h(x)\right|>\varepsilon\right) &\le 4\exp\left(-\frac{\varepsilon^2 n}{32q\sigma_{[1/q]}^2(x)+8q^{-1}\varepsilon r_n^{-2/s}}\right)+11nq\alpha_{[1/q]}\left(1+\frac{4}{\varepsilon r_n^{2/s}}\right)^{1/2}\\ &\le 4\exp\left(-\frac{\varepsilon^2 n}{32Jh^d+8q^{-1}\varepsilon r_n^{-2/s}}\right)+12anq^{1+\beta}\varepsilon^{-1/2}r_n^{-1/s}\end{aligned}$$

where the second inequality holds for n sufficiently large and the assumption on the mixing coefficients.

Set $q = r_n^{1-2/s}$ and $\varepsilon = Mr_nh^d$ for $M > J$. This gives the bound

$$\begin{aligned} & P\left(\left|\hat{V}_h(x) - E\hat{V}_h(x)\right| > Mr_nh^d\right) \\ \leq\ & 4\exp\left(-\frac{M^2r_n^2h^{2d}n}{32Jh^d + 8Mh^d}\right) + 12aM^{-1/2}nr_n^{(1+\beta)(1-2/s)-1/2-1/s}h^{-d/2} \\ =\ & 4n^{-M/40} + 12aM^{-1/2}h^{-d(3/2+\beta(1-2/s)-3/s)/2}n^{(3/2-\beta(1-2/s)+3/s)/2}(\log n)^{\lambda} \end{aligned}$$

where $\lambda = (\beta(1-2/s) + 1/2 - 3/s)/2$. Hence

$$\begin{aligned} P\left(\max_{1\leq j\leq N}\left|\hat{V}_h(x_j) - E\hat{V}_h(x_j)\right| > Mh^dr_n\right) &\leq N\,P\left(\left|\hat{V}_h(x_j) - E\hat{V}_h(x_j)\right| > Mh^dr_n\right) \\ &\leq 4Nn^{-M/40} + 12a\frac{Nn^{(3/2-\beta(1-2/s)+3/s)/2}(\log n)^{\lambda}}{M^{1/2}h^{d(3/2+\beta(1-2/s)-3/s)/2}}. \end{aligned} \tag{30}$$

We now show that the right-hand-side of (30) is $o(1)$ for M sufficiently large, which implies (26). Indeed, using definition (22)

$$\begin{aligned} Nn^{-M/40} &= \frac{h^{-d(1+d\mu/2(\mu-1))}}{2^d}\left(\frac{n}{\log n}\right)^{d\mu/2(\mu-1)}n^{-M/40} \\ &\leq n^{\gamma d(1+d\mu/2(\mu-1))+d\mu/2(\mu-1)-M/40} \end{aligned}$$

which is $= o(1)$ when $M > 40\gamma d(1 + d\mu/2(\mu-1)) + 20d\mu/(\mu-1)$. Furthermore,

$$\begin{aligned} \frac{Nn^{(3/2-\beta(1-2/s)+3/s)/2}(\log n)^{\lambda}}{h^{d(3/2+\beta(1-2/s)-3/s)/2}} &= \left(\frac{n^{d\mu/(\mu-1)+3/2-\beta(1-2/s)+3/s}}{h^{d(7/2+d\mu/(\mu-1)+\beta(1-2/s)-3/s)}}\right)^{1/2}(\log n)^{\lambda-d\mu/2(\mu-1)} \\ &= \left(n^{\gamma d(7/2+d\mu/(\mu-1)+\beta(1-2/s)-3/s)+d\mu/(\mu-1)+3/2-\beta(1-2/s)+3/s}\right)^{1/2}(\log n)^{\lambda-d\mu/2} \\ &= o(1) \end{aligned}$$

since

$$\gamma d\left(\frac{7}{2} + d\frac{\mu}{\mu-1} + \beta\left(1-\frac{2}{s}\right) - \frac{3}{s}\right) + d\frac{\mu}{\mu-1} + \frac{3}{2} - \beta\left(1-\frac{2}{s}\right) + \frac{3}{s} < 0$$

under (10). Thus (30) is $o(1)$ which establishes (26) as desired.

We have shown (23)-(26), which completes the proof. ■

Proof of Theorem 3. In the notation of Theorem 2, $\hat{f}_X(x) = h^{-d}\hat{G}_h(x)$ with $G(x) = K(x)$. Since $k(x)$ satisfies Assumption 1, so does the product $K(x)$. Equation (12) is (10) substituting $s = \infty$ and $\gamma = 1/(d+4)$. Then by Theorem 2

$$\sup_{x\in R^d}\left|\hat{f}_X(x) - E\hat{f}_X(x)\right| = h^{-d}\sup_{x\in R^d}\left|\hat{G}_h(x) - E\hat{G}_h(x)\right| = O_p(r_n)$$

where

$$r_n = \left(\frac{\log n}{n^{-d/(d+4)} n} \right)^{1/2} = \frac{\log^{1/2} n}{n^{2/(d+4)}}.$$

It is well known that under these assumptions

$$E\hat{f}_X(x) - f_X(x) = O\left(h^2\right) = O\left(n^{-2/(d+4)}\right) \leq O\left(r_n\right).$$

Together we obtain

$$\sup_{x \in R^d} \left| \hat{f}_X(x) - f_X(x) \right| = O_p\left(r_n\right).$$

■

Proof of Theorem 4. Let $e_i = Y_i - m(X_i)$. Then

$$\begin{aligned}
\hat{m}(x) - m(x) &= \frac{\sum_{i=1}^n \left(Y_i - m(x)\right) K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} \\
&= \frac{\sum_{i=1}^n \left(e_i + m(X_i) - m(x)\right) K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} \\
&= \hat{f}_X(x)^{-1} \left(\frac{1}{nh^d} \sum_{i=1}^n e_i K\left(\frac{x-X_i}{h}\right) + \frac{1}{nh^d} \sum_{i=1}^n \left(m(X_i) - m(x)\right) K\left(\frac{x-X_i}{h}\right) \right).
\end{aligned}$$

Now consider the terms on the right hand side.

First, by a Taylor expansion, Theorem 3, and the definition of A_n

$$\sup_{x \in A_n} \left| \hat{f}_X(x)^{-1} - f_X(x)^{-1} \right| \leq \sup_{x \in A_n} f_X(x)^{-2} \sup_{x \in A_n} \left| \hat{f}_X(x) - f_X(x) \right| \leq \delta_n^{-2} O_p\left(r_n\right),$$

Second, by Theorem 2 and the fact that $E\left(e_i \mid X_i\right) = 0$

$$\sup_{x \in A_n} \left| \frac{1}{nh^d} \sum_{i=1}^n e_i K\left(\frac{x-X_i}{h}\right) \right| = O_p\left(r_n\right).$$

Third, by Theorem 2

$$\sup_{x \in A_n} \left| \frac{1}{nh^d} \sum_{i=1}^n \left(\left(m(X_i) - m(x)\right) K\left(\frac{x-X_i}{h}\right) - E\left(m(X_i) - m(x)\right) K\left(\frac{x-X_i}{h}\right) \right) \right| = O_p\left(r_n\right).$$

Now, letting x^* denote a point intermediate between x and $x - hu$,

$$\begin{aligned}
& \frac{1}{h^d} E\left(m(X_i) - m(x)\right) K\left(\frac{x - X_i}{h}\right) \\
= \; & \frac{1}{h^d} \int \left(m(u) - m(x)\right) K\left(\frac{x-u}{h}\right) f_X(u) du \\
= \; & \int \left(m(x - hu) - m(x)\right) K\left(u\right) f_X(x - hu) du \\
= \; & h^2 \nabla m(x)' \int uu' \, K\left(u\right) \nabla f_X(x^*) du + \frac{h^2}{2} \operatorname{tr}\left(\int \nabla^2 m(x^*) uu' \, K\left(u\right) f_X(x - hu) du \right).
\end{aligned}$$

Thus for h sufficiently small

$$\begin{aligned}
& \left| \frac{1}{h^d} E\left(m(X_i) - m(x)\right) K\left(\frac{x - X_i}{h}\right) \right| \\
\leq \; & h^2 \left| \int \int uu' \, K\left(u\right) du \right| \left(\sup_{|x_1 - x_2| \leq \delta} \left| \nabla m(x_1)' \nabla f_X(x_2) \right| + \sup_{|x_1 - x_2| \leq \delta} \left| \nabla^2 m(x_1) f_X(x_2) \right| \right) \\
= \; & O\left(h^2\right) \leq O_p\left(r_n\right)
\end{aligned}$$

Together, these complete the proof.