A Uniform Strong Law of Large Numbers for Sample Kernel-Weighted Moments

Andrew E. Schulman* David Ruppert†

November 5, 1998

Abstract

With probability 1, the difference between the sample and ordinary expectations of kernel-weighted moments $h^{-1}\left[K^{(q)}\left(\frac{X-x}{h}\right)\right]^r\left(\frac{X-x}{h}\right)^s g(X,Y)$ has order $(nh)^{-1/2}(\log n)^{1/2}$ as $n \to \infty$, uniformly over intervals of x and h, for well-behaved K and g. Schulman and Ruppert (1998) prove the same theorem, but here we supply details of the proof which are omitted from that paper.

1 Introduction

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be i.i.d. pairs of random variables. We prove a uniform strong law of large numbers (Proposition 1) for quantities of the form $(nh)^{-1}\sum_{i=1}^n\left[K^{(q)}\left(\frac{X_i-x}{h}\right)\right]^r\left(\frac{X_i-x}{h}\right)^s g(X_i,Y_i)$, where the uniformity holds over intervals of h and x. The proof relies on a general uniform strong law of large numbers of Pollard (1984).

Section 2 provides background on uniform strong laws, most of it taken from Pollard (1984). Section 3 applies the theory of Section 2 to prove our Proposition.

*Mathematical Statistician, Biostatistics Branch, National Institute of Neurological Disorders and Stroke, Federal Building Room 7C08, Bethesda, Maryland 20892 (e-mail: Andrew_Schulman@nih.gov). Schulman's research was supported by NIEHS Training Grant ESO7261.

†Professor, School of Operations Research & Industrial Engineering,Cornell University, Ithaca, New York 14853 (e-mail: davidr@orie.cornell.edu). Ruppert's research was supported by NSA Grant MDA 904-95-H-1025 and NSF Grant DMS-9306196.

2 Uniform strong laws of large numbers

Strong laws of large numbers may be shown to hold uniformly over classes of functions that are not too large. Here "large" refers to L^1 covering numbers, that is, the number of functions required to approximate all members of a class to less than any $\epsilon > 0$ in the L^1 sense.

Definition 1. Let $\mathcal{F}$ be a class of functions on a set S, and let $\epsilon > 0$. The (L^1) *covering number* $N(\epsilon, Q, \mathcal{F})$ is the smallest number m for which there exist functions $g_1, \ldots, g_m$ on S (not necessarily in $\mathcal{F}$) such that $\min_i \int |f - g_i| \, dQ \leqslant \epsilon$ for each $f \in \mathcal{F}$.

Although $g_1, \ldots, g_m$ do not have to belong to $\mathcal{F}$, Pollard (1984) observes that we can require them to do so at the cost of doubling ϵ, by replacing g_i by $f_i \in \mathcal{F}$ satisfying $\int |g_i - f_i| \, dQ \leqslant \epsilon$.

The conditions on covering numbers that lead to uniform strong laws are collected in the following definition.

Definition 2. A class $\mathcal{F}$ of functions on a set S is *well-covered* if:

(i) $\mathcal{F}$ is *permissible* in the sense of Pollard (1984, Appendix C);

(ii) $|f| \leqslant C$ for all $f \in \mathcal{F}$ and some $C > 0$;

(iii) there exist constants $A > 0$ and $W > 0$ such that $\sup_{Q \in \mathcal{Q}} N(\epsilon, Q, \mathcal{F}) \leqslant A\epsilon^{-W}$ for all $0 < \epsilon < 1$.

Here $\mathcal{Q}$ is the set of all probability measures on S with its σ-algebra. A sequence $\{\mathcal{F}_n\}$ of classes of functions on sets $\{S_n\}$ is *uniformly well-covered* if

(i) $\mathcal{F}_n$ is permissible for each n;

(ii) $|f| \leqslant C$ for all $f \in \mathcal{F}_n$ and some $C > 0$;

(iii) there exist constants $A > 0$ and $W > 0$ such that $\sup_n \sup_{Q \in \mathcal{Q}_n} N(\epsilon, Q, \mathcal{F}_n) \leqslant A\epsilon^{-W}$ for all $0 < \epsilon < 1$.

Here $\mathcal{Q}_n$ is the set of all probability measures on S_n with its σ-algebra.

Permissibility is a guarantee that suprema of measurable functions over uncountable index sets are measurable. All the classes of functions considered here are permissible, so this condition will not be discussed further. For details the reader is referred to Appendix C of Pollard (1984).

Note that well-covering is a special case of uniform well-covering, in which $\mathcal{F}_n = \mathcal{F}$ for all n. Thus any results about uniform well-covering include ordinary well-covering as a special case. Also, if $\mathcal{F}$ is a well-covered class, then any sequence of subclasses of $\mathcal{F}$ is uniformly well-covered.

With these definitions, we now state a uniform strong law proved by Pollard (1984). Let E_n stand for empirical expectation, so that for a function f and a random variable ξ, $E_n f(\xi) = n^{-1} \sum_{i=1}^{n} f(\xi_i)$ for some i.i.d. random variables $\xi_1, \ldots, \xi_n$ with the same distribution as ξ.

Theorem 1 (Pollard (1984) II.37). *Let $\mathcal{F}_n = \{f_s : s \in S_n\}$ be a uniformly well-covered sequence of classes of functions, and $\{\alpha_n\}$ a non-increasing sequence of positive numbers. If $Ef_s(\xi)^2 \leqslant \delta_n^2$ for each $s \in S_n$, and $n\delta_n^2\alpha_n^2/\log n \to \infty$, then*

$$\sup_{s \in S_n} |E_n f_s(\xi) - Ef_s(\xi)| = o(\delta_n^2 \alpha_n) \quad \textit{almost surely.}$$

In order to invoke Theorem 1, we need to show uniform well-covering of the classes $\mathcal{F}_n$. In particular, we need uniform bounds of the form $A\epsilon^{-W}$ on covering numbers. The easiest way to get such bounds is by comparison with the covering numbers of related classes. The next lemma contains results of this type. The lemma is stated in terms of uniform well-covering, but as was noted above, it includes ordinary well-covering as a special case by setting $\mathcal{F}_{1n} = \mathcal{F}_1$ and $\mathcal{F}_{2n} = \mathcal{F}_2$.

Lemma 1. *Let $\{\mathcal{F}_{1n}\}$ and $\{\mathcal{F}_{2n}\}$ be two uniformly well-covered sequences of classes of functions on sets $\{S_{1n}\}$ and $\{S_{2n}\}$, respectively. Then the sequence $\{\mathcal{G}_n\}$ is uniformly well-covered, for each of the following $\mathcal{G}_n$:*

(a) $\mathcal{G}_n = \{a_1 f_1(s) + a_2 f_2(s) : f_1 \in \mathcal{F}_{1n}, f_2 \in \mathcal{F}_{2n}\}$ *on* $S_{1n} \cap S_{2n}$*, for each fixed* $a_1, a_2 \in \mathbb{R}$*;*

(b) $\mathcal{G}_n = \{a_1 f_1(s) + a_2 f_2(t) : f_1 \in \mathcal{F}_{1n}, f_2 \in \mathcal{F}_{2n}\}$ *on* $S_{1n} \times S_{2n}$*, for each fixed* $a_1, a_2 \in \mathbb{R}$*;*

(c) $\mathcal{G}_n = \{f_1(s) f_2(s) : f_1 \in \mathcal{F}_{1n}, f_2 \in \mathcal{F}_{2n}\}$ *on* $S_{1n} \cap S_{2n}$*;*

(d) $\mathcal{G}_n = \{f_1(s) f_2(t) : f_1 \in \mathcal{F}_{1n}, f_2 \in \mathcal{F}_{2n}\}$ *on* $S_{1n} \times S_{2n}$*. (Opsomer, 1994, Lemma A.5)*

Proof. All the classes are uniformly bounded, so check the covering numbers. For convenience write $Qf = \int f\, dQ$.

(a) Fix $\epsilon > 0$ and n, and let Q be a probability on $S_{1n} \cap S_{2n}$. If $a_1 \neq 0$ and $a_2 \neq 0$, then choose $f_{11}, \ldots, f_{1m_1}$ and $f_{21}, \ldots, f_{2m_2}$ that approximate all $f_1 \in \mathcal{F}_{1n}$ and $f_2 \in \mathcal{F}_{2n}$ to within $\frac{\epsilon}{2|a_1|}$ and $\frac{\epsilon}{2|a_2|}$, respectively, where $m_i \leqslant A_i(\frac{\epsilon}{2|a_i|})^{-W_i}$, $i = 1, 2$, and A_i and W_i are independent of Q and n. If either a_i is zero, then $\mathcal{F}_{in}$ may be ignored so choose $m_i = 1$ and $f_{i1} \equiv 0$. Now the functions $a_1 f_{1i}(\cdot) + a_2 f_{2j}(\cdot)$ approximate all $a_1 f_1(\cdot) + a_2 f_2(\cdot)$:

$$\begin{aligned}
&\min_{i,j} Q|a_1 f_1 + a_2 f_2 - (a_1 f_{1i} + a_2 f_{2j})| \\
&\leqslant \min_{i,j} \big[|a_1| Q|f_1 - f_{1i}| + |a_2| Q|f_2 - f_{2j}|\big] \\
&= |a_1| \min_i Q|f_1 - f_{1i}| + |a_2| \min_j Q|f_2 - f_{2j}| \leqslant \epsilon,
\end{aligned}$$

and there are at most $m_1 m_2 \leqslant \left[A_1 A_2 (2|a_1|)^{W_1} (2|a_2|)^{W_2}\right] \epsilon^{-(W_1+W_2)}$ such functions.

(b) Define two new sequences of classes of functions on $S_{1n} \times S_{2n}$, by $\mathcal{G}_{1n} = \{g_1(s,t) = f_1(s) : f_1 \in \mathcal{F}_{1n}\}$ and $\mathcal{G}_{2n} = \{g_2(s,t) = f_2(t) : f_2 \in \mathcal{F}_{2n}\}$. It suffices to show that $\mathcal{G}_{1n}$ and $\mathcal{G}_{2n}$ are uniformly well-covered on $S_{1n} \times S_{2n}$, since then by part (a) (with $S_{1n} \times S_{2n}$ in place of S_{1n} and S_{2n}) the sequence of classes of all $a_1 g_1(s,t) + a_2 g_2(s,t) = a_1 f_1(s) + a_2 f_2(t)$ is uniformly well-covered.

Fix $\epsilon > 0$ and n, and let Q be a probability on $S_{1n} \times S_{2n}$. For $\mathcal{G}_{1n}$, define a probability Q_1 on S_{1n} by $Q_1(A) = Q(A \times S_{2n})$, for measurable $A \subset S_{1n}$. Choose $f_{11}, \ldots, f_{1m_1}$ such that $\min_i Q_1|f_1 - f_{1i}| < \epsilon$ for all $f_1 \in \mathcal{F}_{1n}$, where $m_1 \leqslant A\epsilon^{-W}$ and A and W are independent of n and Q_1, and hence of Q. Now the functions $g_{1i}(s,t) := f_{1i}(s)$ approximate all $g_1(s,t) = f_1(s)$ for $f_1 \in \mathcal{F}_{1n}$, since $\min_i Q|g_1 - g_{1i}| = \min_i Q_1|f_1 - f_{1i}| < \epsilon$. Hence $\mathcal{G}_{1n}$ is uniformly well-covered. An analogous argument applies to $\mathcal{G}_2$.

(c) The proof is the same as (a), but with the inequality

$$Q|f_1 f_2 - f_{1i} f_{2j}| \leqslant \sup|f_2| \cdot Q|f_1 - f_{1i}| + \sup|f_{1i}| \cdot Q|f_2 - f_{2j}|.$$

Since the classes $\{\mathcal{F}_{1n}\}$ are uniformly bounded, we can without loss of generality choose all f_{1i} to satisfy the same uniform bound.

(d) The proof is the same as (b), with $f_1 f_2$ in place of $a_1 f_1 + a_2 f_2$. ■

To apply Lemma 1, we need first to bound the covering numbers of some initial class. Pollard develops bounds for classes which are, again, not too big, where now the size of $\mathcal{F}$ depends on how well the graphs of functions in $\mathcal{F}$ can divide arbitrary sets of points in $S \times \mathbb{R}$. The *graph* of a function f on a set S is the set $\{(s,t) : 0 \leqslant t \leqslant f(s) \text{ or } f(s) \leqslant t \leqslant 0\}$ in $S \times \mathbb{R}$.

Definition 3. A class $\mathcal{D}$ of subsets of S is a *polynomial class* or has *polynomial discrimination* if there exists a polynomial $\rho(\cdot)$ such that, from every set S_0 of N points in S, there are at most $\rho(N)$ distinct sets of the form $S_0 \cap D$ with D in $\mathcal{D}$.

Since there are 2^N subsets of any set of size N, a class that picks out only polynomially many such subsets must eventually fail to pick them all out, for large enough N. Such a class is in some sense small, and for the graphs of a class $\mathcal{F}$ of uniformly bounded functions, this turns out to be sufficient to bound the covering numbers of $\mathcal{F}$. The next lemma is trivially different from the version given by Pollard.

Lemma 2 (Pollard (1984) II.25). *If $\mathcal{F}$ is a class of uniformly bounded functions, and the graphs of functions in $\mathcal{F}$ form a polynomial class of sets, then $\mathcal{F}$ is well-covered.*

The next problem, then, is to show polynomial discrimination of the graphs of some initial class of functions. The definition is difficult to use, although Opsomer (1994, Lemma A.4) does so. Fortunately, Pollard provides some tools for showing polynomial discrimination, of which the two in the next lemma will be sufficient for our purposes.

Lemma 3 (Pollard (1984), II.18 and II.15).

(a) If $\mathcal{G}$ is a finite-dimensional vector space of real functions on S, then the class of sets of the form $\{g \geqslant 0\}$, for $g \in \mathcal{G}$, has polynomial discrimination in $S \times \mathbb{R}$.

(b) If classes $\mathcal{C}$ and $\mathcal{D}$ have polynomial discrimination, then so do each of:

- *(i)* $\{D^c : D \in \mathcal{D}\}$
- *(ii)* $\{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$
- *(iii)* $\{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$

In summary, our approach is as follows. First, use Lemma 3 to show polynomial discrimination of the graphs of functions in some initial class, and conclude from Lemma 2 that the class is well-covered. Extend the well-covering to other classes of interest, using Lemma 1 or similar arguments, and then apply Theorem 1 to get uniform a.s. rates of convergence of sample expectations of functions in the well-covered classes.

3 Kernel-Weighted Moments

Following the theory of Section 2, we now use polynomial discrimination to find a well-covered class of functions (Lemma 4), extend the well-covering to classes of kernel-weighted monomials (Lemmas 5 and 6), and use the well-covering to prove a uniform strong law for sample kernel-weighted moments (Proposition 1).

Let $\mathcal{X}$ be a measurable subset of $\mathbb{R}$, and let $\mathcal{H}_n = [\underline{\alpha} n^{\underline{\gamma}}, \overline{\alpha} n^{\overline{\gamma}}]$ be a sequence of bandwidth intervals, where $-1 < \underline{\gamma} \leqslant \overline{\gamma} < 0$ and $\underline{\alpha}, \overline{\alpha} > 0$. We use the following assumptions.

(A1) The kernel function K is supported on $[-1,1]$, and D-times differentiable on $\mathbb{R}$, for some $D \in \{0, 1, \dots\}$. $K^{(D)}$ has bounded variation on $[-1,1]$ (write $K^{(D)} \in BV[-1,1]$).

(A2) $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. observations from the distribution of a random pair (X, Y), where X has a bounded density f_X.

Lemma 4. *If $g : [c,d] \to \mathbb{R}$ has bounded variation and $-\infty < c \leqslant d < \infty$, then the location-scale class $\mathcal{G} = \{g(\frac{\cdot - t}{h}) : t \in \mathbb{R}, h > 0\}$ is well-covered.*

Proof. (After Pollard (1984), example II.26) First suppose that g is non-negative and strictly increasing on $[c,d]$. The graph of $g(\frac{\cdot - t}{h})$ is the set

$$\begin{aligned}
&\{(x,y) : 0 \leqslant y \leqslant g(\tfrac{x-t}{h}),\ c \leqslant \tfrac{x-t}{h} \leqslant d\} \\
&= \{(x,y) : 0 \leqslant y,\ g^{-1}(y) \leqslant \tfrac{x-t}{h},\ c \leqslant \tfrac{x-t}{h} \leqslant d\} \\
&= (\mathbb{R} \times [0,\infty)) \cap \{(x,y) : x - hg^{-1}(y) - t \geqslant 0\} \\
&\qquad \cap \{(x,y) : x - hc - t \geqslant 0\} \cap \{(x,y) : -x + hd + t \geqslant 0\}.
\end{aligned} \tag{3.1}$$

Consider the second set of (3.1). Define a vector space of functions $\{k_{aht}(x,y) = ax - hg^{-1}(y) - t : a, h, t \in \mathbb{R}\}$ on $\mathbb{R}^2$. By Lemma 3(a), the sets $\{(x,y) : ax - hg^{-1}(y) - t \geqslant 0\}$ form a polynomial class; hence the smaller class with $a = 1$ and $h > 0$ must also be a polynomial class. By a similar argument, the classes of sets $\{(x,y) : x - hc - t \geqslant 0\}$ and $\{(x,y) : -x + hd + t \geqslant 0\}$ are also polynomial classes. Therefore by Lemma 3(b), the class $\mathcal{G}$ of sets in (3.1) is a polynomial class. Also since g has bounded variation, g is bounded; therefore $\mathcal{G}$ is uniformly bounded and hence well-covered, by Lemma 2.

Now consider arbitrary $g \in BV[c,d]$. It is a fact (see e.g. **?**, Theorem 5.2.5 and its proof) that any function of bounded variation on $[c,d]$ may be written as a difference of two bounded, strictly increasing, non-negative functions on $[c,d]$. Thus write $g(\frac{\cdot - t}{h}) = g_1(\frac{\cdot - t}{h}) - g_2(\frac{\cdot - t}{h})$, for bounded, non-negative, strictly increasing g_1 and g_2. As shown above, the location-scale classes of g_1 and g_2 are well-covered; so the location-scale class of $g_1 - g_2$ is also well-covered, by Lemma 1(a). ∎

Lemma 5. *The classes* $\{[K^{(q)}(\frac{\cdot - t}{h})]^r (\frac{\cdot - t}{h})^s : t \in \mathbb{R}, h > 0\}$ *and* $\{\left|K^{(q)}(\frac{\cdot - t}{h})\right|^r \left|\frac{\cdot - t}{h}\right|^s : t \in \mathbb{R}, h > 0\}$ *are well-covered, for each* $q = 0, \dots, D$, $r = 1, 2, \dots,$ *and* $s = 0, 1, \dots$.

Proof. By Lemma 4, it suffices to show that $[K^{(q)}(\cdot)]^r(\cdot)^s$ and $|K^{(q)}(\cdot)|^r|\cdot|^s$ have bounded variation on $[-1,1]$.

By assumption (A1), $K^{(D)} \in BV[-1,1]$. In particular $K^{(D)}$ is bounded, so for any partition $-1 = x_1 < \cdots < x_N = 1$, $\sum |K^{(D-1)}(x_{i+1}) - K^{(D-1)}(x_i)| = \sum \left| \int_{x_i}^{x_{i+1}} K^{(D)}(t)\, dt \right| \leqslant \int_{-1}^{1} |K^{(D)}(t)|\, dt < \infty$; therefore $K^{(D-1)} \in BV[-1,1]$. Repeat the argument to show that all of $K, \dots, K^{(D)}$ have bounded variation on $[-1,1]$.

For any f, $f \in BV[-1,1]$ implies also $f^r \in BV[-1,1]$, since

$$\begin{aligned} |f(x)^r - f(y)^r| &= |(f(x) - f(y)) \textstyle\sum_{j=0}^{r-1} f(x)^j f(y)^{r-1-j}| \\ &\leqslant r \sup |f|^{r-1} \cdot |f(x) - f(y)| \quad \text{for all } x, y. \end{aligned}$$

Also clearly $(\cdot)^s \in BV[-1,1]$. If $f, g \in BV[-1,1]$, then $fg \in BV[-1,1]$: f and g are bounded, so $|f(x)g(x) - f(y)g(y)| \leqslant |f(x) - f(y)| \sup |g| + |g(x) - g(y)| \sup |f|$ implies $fg \in BV[-1,1]$. Therefore $[K^{(q)}(\cdot)]^r(\cdot)^s \in BV[-1,1]$, for each q, r, and s.

Finally, $f \in BV[-1,1]$ implies $|f| \in BV[-1,1]$, since $\big||f(x)| - |f(y)|\big| \leqslant |f(x) - f(y)|$. ∎

We can use Lemma 5 to show uniform well-covering of a related class of functions, the scaled monomials $h^{-1/2}[K^{(q)}(\frac{\cdot-t}{h})]^r(\frac{\cdot-t}{h})^s$, as long as h is bounded away from zero for each n. For our purposes it is good enough to assume a little more, that $h \in \mathcal{H}_n = [\underline{\alpha} n^{\underline{\gamma}}, \overline{\alpha} n^{\overline{\gamma}}]$.

Lemma 6. *The classes* $\mathcal{E}_n = \{n^{\underline{\gamma}/2} h^{-1/2}[K^{(q)}(\frac{\cdot-t}{h})]^r(\frac{\cdot-t}{h})^s : t \in \mathbb{R}, h \in \mathcal{H}_n\}$ *and* $\mathcal{F}_n = \{n^{\underline{\gamma}/2} h^{-1/2} \left|K^{(q)}(\frac{\cdot-t}{h})\right|^r \left|\frac{\cdot-t}{h}\right|^s : t \in \mathbb{R}, h \in \mathcal{H}_n\}$ *are uniformly well-covered, for each* $q = 0, \dots, D$, $r = 1, 2, \dots,$ *and* $s = 0, 1, \dots$.

Proof. The sequence $\mathcal{G}_n = \{n^{\underline{\gamma}/2} h^{-1/2} : h \in \mathcal{H}_n\}$ of classes of constant functions on $\mathbb{R}$ is uniformly bounded above by $\underline{\alpha}^{-1/2}$. Therefore the constant functions $\{\epsilon, 2\epsilon, \dots, J\epsilon \approx \underline{\alpha}^{-1/2}\}$, where $J = \lfloor \underline{\alpha}^{-1/2}/\epsilon \rfloor$, approximate all $n^{\underline{\gamma}/2} h^{-1/2}$ to within ϵ, and there are at most $\underline{\alpha}^{-1/2}\epsilon^{-1}$ such functions. Thus the $\mathcal{G}_n$ are uniformly well-covered, so by Lemmas 5 and 1(c), the $\mathcal{E}_n$ and $\mathcal{F}_n$ are also uniformly well-covered. ∎

Lemma 6 allows us to apply Theorem 1 to prove our main result. The following uniform order notation is convenient: write $A_n(x,h) = O_U(B_n(x,h))$ if $\sup_{n,x,h} |A_n(x,h)/B_n(x,h)| < \infty$, and $A_n(x,h) = o_U(B_n(x,h))$ if $\sup_{x,h} |A_n(x,h)/B_n(x,h)| \to 0$ as $n \to \infty$. Here and below, suprema are taken over $x \in \mathcal{X}$ and $h \in \mathcal{H}_n$. We also write $O_U(B_n)$ to stand for some function A_n which satisfies $A_n = O_U(B_n)$.

Proposition 1. *Let $g(x,y)$ be a measurable function on $\mathcal{X} \times \mathbb{R}$ such that: $P\big(|g(X,Y)| \geqslant z\big) \leqslant Cz^{-\alpha}$ for all z and some $C > 0$ and $\alpha > 2/(1+\underline{\gamma})$; and $\eta(\cdot) := E(g(X,Y)^2|X=\cdot)$ is bounded. Then with probability 1,*

$$\begin{aligned} &E_n h^{-1}\big[K^{(q)}\big(\tfrac{X-x}{h}\big)\big]^r \big(\tfrac{X-x}{h}\big)^s g(X,Y) \\ &= E h^{-1}\big[K^{(q)}\big(\tfrac{X-x}{h}\big)\big]^r \big(\tfrac{X-x}{h}\big)^s g(X,Y) + O_U((nh)^{-1/2}(\log n)^{1/2}) \end{aligned}$$

for each $q = 0, \dots, D$, $r = 1, 2, \dots,$ *and* $s = 0, 1, \dots$; *and similarly for* $h^{-1}\left|K^{(q)}\big(\frac{X-x}{h}\big)\right|^r \left|\frac{X-x}{h}\right|^s |g(X,Y)|$.

Proof. The proofs with and without the absolute values are identical, so only the latter is given. If A is any set, the indicator function of A will also be denoted A. The empirical probability of A is denoted $P_n A := E_n A$.

Write $L = K^{(q)}$ and

$$\begin{aligned}
R_{nxh} &= n^{\underline{\gamma}/2} h^{-1/2} L(\tfrac{X-x}{h})^r (\tfrac{X-x}{h})^s g(X,Y) \\
S_{nxh} &= n^{\underline{\gamma}/2} h^{-1/2} L(\tfrac{X-x}{h})^r (\tfrac{X-x}{h})^s g(X,Y)\{|g(X,Y)| \leqslant \Delta_n\} \\
T_{nxh} &= n^{\underline{\gamma}/2} h^{-1/2} L(\tfrac{X-x}{h})^r (\tfrac{X-x}{h})^s g(X,Y)\{|g(X,Y)| > \Delta_n\}
\end{aligned}$$

for all $x \in \mathcal{X}$, $h \in \mathcal{H}_n$, and some sequence $\Delta_n \to \infty$. Theorem 1 applies to the bounded piece, S_{nxh}, as long as Δ_n is not too large; we look for the largest possible Δ_n, in order to make T_{nxh} as small as possible.

(a) Convergence of $E_n S_{nxh}$. The classes $\{\Delta_n^{-1} g(\cdot\,,\cdot)\{|g(\cdot\,,\cdot)| \leqslant \Delta_n\}\}$ are trivially uniformly well-covered (each set has only one element), so by Lemmas 6 and 1(d), the classes $\{\Delta_n^{-1} S_{nxh} : x \in \mathcal{X}, h \in \mathcal{H}_n\}$ (where S_{nxh} is treated as a function of X and Y) are uniformly well-covered. Also

$$\begin{aligned}
E(\Delta_n^{-1} S_{nxh})^2 &\leqslant \Delta_n^{-2} n^{\underline{\gamma}} h^{-1} \int L(\tfrac{u-x}{h})^{2r} (\tfrac{u-x}{h})^{2s} \eta(u) f_X(u)\, du \\
&= \Delta_n^{-2} n^{\underline{\gamma}} \int L(t)^{2r} t^{2s} \eta(x+ht) f_X(x+ht)\, dt \\
&= O_U(\Delta_n^{-2} n^{\underline{\gamma}})
\end{aligned}$$

since f_X and η are bounded, and L is bounded with compact support. So for Theorem 1 choose $\delta_n^2 = \Delta_n^{-2} n^{\underline{\gamma}} \tau_n$ for any $\tau_n \to \infty$, and to satisfy $n\delta_n^2/\log n \to \infty$, let $\Delta_n = n^{(1+\underline{\gamma})/2}/(\log n)^{1/2}$. Then Theorem 1 with $\alpha_n \equiv 1$ gives

$$\begin{aligned}
\Delta_n^{-1} E_n S_{nxh} &= \Delta_n^{-1} E S_{nxh} + o_U(\Delta_n^{-2} n^{\underline{\gamma}} \tau_n) \quad \text{a.s.} \\
\text{or} \qquad E_n S_{nxh} &= E S_{nxh} + o_U(n^{(\underline{\gamma}-1)/2} (\log n)^{1/2} \tau_n) \quad \text{a.s.}
\end{aligned}$$

Since this holds for all $\tau_n \to \infty$,

$$E_n S_{nxh} = E S_{nxh} + O_U(n^{(\underline{\gamma}-1)/2} (\log n)^{1/2}) \quad \text{a.s.} \tag{3.2}$$

(b) Convergence of $E_n T_{nxh}$. Let $Z = g(X,Y)$. By hypothesis, $P(|Z| \geqslant z) \leqslant C z^{-\alpha}$ for all z and some $\alpha > 2$. Let $p = \alpha - \epsilon$ for some small $\epsilon > 0$; then $E|Z|^p = p \int z^{p-1} P(|Z| \geqslant z)\, dz \leqslant$

$pC\int z^{\alpha-\epsilon-1}z^{-\alpha}\,dz<\infty$. Use Hölder's inequality:

$$\begin{aligned}
&\sup_{x,h}|E_nT_{nxh}-ET_{nxh}|\\
&\leqslant E_n\sup_{x,h}|T_{nxh}|+E\sup_{x,h}|T_{nxh}|\\
&=(E_n+E)n^{\underline{\gamma}/2}\sup_{x,h}h^{-1/2}|L(\tfrac{X-x}{h})|^r|\tfrac{X-x}{h}|^s|Z|\{|Z|>\Delta_n\}\\
&\leqslant \underline{\alpha}^{-1/2}(\sup|L|^r)\,(E_n+E)|Z|\{|Z|>\Delta_n\}\\
&\leqslant \underline{\alpha}^{-1/2}(\sup|L|^r)\,\big[(E_n+E)|Z|^p\big]^{\frac{1}{p}}\cdot\big[(P_n+P)\{|Z|>\Delta_n\}\big]^{\frac{p-1}{p}}\\
&\leqslant N\big[(P_n+P)\{|Z|>\Delta_n\}\big]^{\frac{p-1}{p}}\quad\text{a.s.}\\
&\leqslant N\big[\big|(P_n-P)\{|Z|>\Delta_n\}\big|+2P\{|Z|>\Delta_n\}\big]^{\frac{p-1}{p}}
\end{aligned}\tag{3.3}$$

for large n and some $N>0$, since $E_n|Z|^p\to E|Z|^p<\infty$ a.s.

For Theorem 1, $E(\{|Z|>\Delta_n\}^2)=P\{|Z|>\Delta_n\}=O(\Delta_n^{-\alpha})=O(n^{(1+\underline{\gamma})/2}/\log n)^{-\alpha}$, but we cannot choose $\delta_n^2=(n^{(1+\underline{\gamma})/2}/\log n)^{-\alpha}$, since then $\alpha>2/(1+\underline{\gamma})$ implies $n\delta_n^2/\log n\to 0$. The smallest allowable δ_n is $\delta_n^2=\tau_n n^{-1}\log n$, for any $\tau_n\to\infty$. This and $\alpha_n\equiv 1$ in Theorem 1 give $|(P_n-P)\{|Z|>\Delta_n\}|=o(\tau_n n^{-1}\log^2 n)$ a.s. for any $\tau_n\to\infty$, or $|(P_n-P)\{|Z|>\Delta_n\}|=O(n^{-1}\log^2 n)$ a.s. Insert this expression into (3.3) and simplify to get

$$\sup_{x,h}|E_nT_{nxh}-ET_{nxh}|=O(n^{-1}\log^2 n)^{1-1/\alpha+O(\epsilon)}\quad\text{a.s.}\tag{3.4}$$

(c) Convergence of E_nR_{nxh}. Comparing (3.2) and (3.4), we can show that the error bound for T_{nxh} is smaller than that of S_{nxh} if and only if $\alpha>2/(1+\underline{\gamma})$. This is true by hypothesis, so $\sup_{x,h}|(E_n-E)R_{nxh}|=O(\sup_{x,h}|(E_n-E)S_{nxh}|)=O(n^{(\underline{\gamma}-1)/2}(\log n)^{1/2})$ a.s. Multiply by $n^{-\underline{\gamma}/2}h^{-1/2}$ to get the result. ∎

Acknowledgements. The authors thank Naomi Altman, who reviewed this material as a member of Andrew Schulman's doctoral committee.

References

Opsomer, J.-D. (1994), *Optimal Bandwidth Selection for Fitting an Additive Model by Local Polynomial Regression*, Ph.D. dissertation, Cornell University.

Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer-Verlag.

Schulman, A. E. and Ruppert, D. (1998), "Rates of Convergence of Three Local Bandwidth Selectors for Local Polynomial Regression," manuscript to be submitted *very soon* to *Annals of Statistics*.