The Annals of Statistics
1987, Vol. 15, No. 2, 780–799

U-PROCESSES: RATES OF CONVERGENCE[1]

BY DEBORAH NOLAN AND DAVID POLLARD

Yale University

This paper introduces a new stochastic process, a collection of *U*-statistics indexed by a family of symmetric kernels. Conditions are found for the uniform almost-sure convergence of a sequence of such processes. Rates of convergence are obtained. An application to cross-validation in density estimation is given. The proofs adapt methods from the theory of empirical processes.

1. Introduction. The theory of *U*-statistics parallels the theory for sums of independent random variables. Hoeffding (1948) proved a central limit theorem and [(1961), according to Serfling (1980), page 191] a strong law of large numbers for these statistics. In this paper we extend the strong-law results to families of *U*-statistics, obtaining uniform limit theorems analogous to those for empirical processes. In a subsequent paper [Nolan and Pollard (1988)], we will prove analogues of the central limit theorem for empirical processes.

Let $\xi_1, \xi_2, \ldots$ be independent observations taken from a distribution P on a set $\mathscr{X}$, and $\mathscr{F}$ be a class of real-valued symmetric functions on $\mathscr{X} \otimes \mathscr{X}$. Define

$$S_n(f) = \sum_{1 \le i \ne j \le n} f(\xi_i, \xi_j), \quad \text{for } f \text{ in } \mathscr{F}.$$

With a $[n(n-1)]^{-1}$ standardization, $S_n(f)$ would become a *U*-statistic in the sense adopted by Serfling (1980, Chapter 5). We treat S_n as a stochastic process indexed by $\mathscr{F}$. We call it the *U*-process.

Our initial motivation for studying these *U*-processes was the paper of Stone (1984) on cross-validation in density estimation, which generalized and improved upon results of Hall (1983). The kernel density estimator

$$\hat{p}_\sigma(x) = (n\sigma)^{-1} \sum_{j=1}^{n} K\left(\frac{x - \xi_j}{\sigma}\right)$$

depends upon the choice of the smoothing parameter σ. Cross-validation gives methods for determining a suitable σ from the data. Stone's analysis of one cross-validatory estimator involves sums such as

$$n^{-1} \sum_{i=1}^{n} \hat{p}_\sigma(\xi_i) = (n^2\sigma)^{-1} \sum_{i,j} K\left(\frac{\xi_i - \xi_j}{\sigma}\right).$$

This is a deterministic function, $K(0)/n\sigma$, plus a standardized *U*-process indexed by σ.

Received July 1985; revised July 1986.

[1]Supported in part by NSF Grant DMS-8503347.

AMS 1980 *subject classifications.* Primary 60F15; secondary 62G99, 60G20

Key words and phrases. *U*-statistics, empirical processes, rates of convergence, cross-validation, reversed submartingale, maximal inequality, kernel density estimation.

Hall (1982) had extracted a similar process, for which he obtained rates of pointwise convergence, from an analysis of cross-validation via maximization of a pseudolikelihood. Hall (1983) measured the distance between $\hat{p}_\sigma$ and the true density $p(\cdot)$ by the integrated squared error,

$$L_n(\sigma) = \int (\hat{p}_\sigma - p)^2.$$

The function

$$M_n(\sigma) = \int \hat{p}_\sigma^2 - 2[\sigma n(n-1)]^{-1} \sum_{i \neq j} K\left(\frac{\xi_i - \xi_j}{\sigma}\right) + \int p^2$$

is a plausible estimate of the unknown function $L_n(\sigma)$. Hall showed that the value σ_M that minimized M_n also came close to minimizing the mean integrated squared error $\mathbb{P}L_n$. (Throughout this paper we use linear functional notation; the expected value of a random variable X is denoted $\mathbb{P}X$.) He applied a strong approximation theorem for empirical processes to prove a uniform convergence result for $L_n - M_n$, which translated into information about the locations of the minima.

Stone (1984) recognized that $L_n - M_n$ could be decomposed into a sum of an empirical process and what we are calling a U-process. Much of his proof was devoted to establishing a rate of convergence for the U-process, by means of a delicate Poissonization argument. The U-process took the form $S_n(\tilde{f})$, where

$$\tilde{f}(x, y) = f(x, y) - Pf(x, \cdot) - Pf(\cdot, y) + P \otimes P(f).$$

The centering ensures that $P\tilde{f}(x, \cdot) = 0$ for every x; that is, $\tilde{f}$ is P-degenerate. This is a variation on the projection method of Hoeffding (1948), which is useful for eliminating cross-product terms in the calculation of moments of U-statistics. Stone bounded tail probabilities using high-order moments of U-statistics.

In this paper we generalize these almost-sure limit theorems to cover U-processes indexed by a class of symmetric functions. By means of the projection method, we reduce the problem to a search for probabilistic bounds on $\sup_{\mathscr{F}} |S_n f|$ for classes of P-degenerate functions.

Our approach parallels the approach to the theory of empirical processes exposited in Chapters II and VII of Pollard (1984). However, there are significant differences between the two theories.

Empirical processes are constructed from sums of independent random variables, whereas U-processes are like quadratic forms. This difference shows up in the behavior of tail probabilities. For the increments of empirical processes, tail probabilities decrease like the Gaussian distribution; for the increments of U-processes, they are more like the exponential distribution. Another difference shows up in a symmetrization argument based on the familiar empirical process technique. To exploit P-degeneracy of U-processes, we symmetrize to get inequalities for expectations rather than inequalities for tail probabilities.

These inequalities, which are presented in Section 2, are key to the limit theorems proved in Sections 3 and 4. Section 3 gives sufficient conditions for the uniform almost-sure convergence of U-processes. Section 4 strengthens these to

find rates of convergence, which include Stone's result as an application. Section 5 summarizes some results about covering numbers, which we need in the earlier sections.

Throughout the paper we ignore measurability difficulties. For a discussion of the appropriate precautions, consult Chapter 10 of Dudley (1984).

2. Inequalities. Observations $\xi_1, \xi_2, \ldots$ are drawn from a fixed distribution P on a space $\mathscr{X}$. Let $\mathscr{F}$ be a class of real-valued, symmetric functions on $\mathscr{X} \otimes \mathscr{X}$, with a nonnegative envelope F:

$$F(\cdot, \cdot) \geq |f(\cdot, \cdot)|, \quad \text{if } f \in \mathscr{F}.$$

For each f in $\mathscr{F}$ we define the symmetric sum

$$S_n(f) = \sum_{i \neq j} f(\xi_i, \xi_j).$$

Here, and throughout the paper, the range of summation is the set of all $n(n-1)$ pairs with $1 \leq i \neq j \leq n$.

We will reduce most calculations with U-statistics to the degenerate case. We will call f degenerate, or P-degenerate if there is any doubt about the underlying distribution, if $Pf(x, \cdot) = 0$ for all x.

The norm $\|\cdot\|$ will indicate a supremum over $\mathscr{F}$. For example, $\|S_n\|$ stands for $\sup_{\mathscr{F}}|S_n(f)|$. Where the advantages of clarity outweigh the disadvantages of mild abuse of notation we will also make free use of expressions like $\|S_n(f)\|$ or $\|S_n(f^2)\|$.

The steps towards our main inequality (Theorem 6) all have empirical process analogues.

Symmetrization inequality. As with empirical processes, probabilistic bounds on $\|S_n\|$ can be obtained by first transforming to a symmetrized version of the process. Independently take a double sample $x_1, \ldots, x_{2n}$ from P and a sample $\sigma_1, \ldots, \sigma_n$ from the distribution that gives each of $+1$ and -1 probability $\frac{1}{2}$. If $\sigma_i = +1$ define $\xi_i = x_{2i}$ and $\eta_i = x_{2i-1}$; if $\sigma_i = -1$ define $\xi_i = x_{2i-1}$ and $\eta_i = x_{2i}$. Both $\{\xi_i\}$ and $\{\eta_i\}$ are independent samples from P. Define

$$S_n'(f) = \sum_{i \neq j} f(\xi_i, \eta_j),$$

$$S_n''(f) = \sum_{i \neq j} f(\eta_i, \eta_j),$$

$$T_n^0(f) = S_n(f) - 2S_n'(f) + S_n''(f) = \sum_{i \neq j} \sigma_i \sigma_j f_{ij},$$

where

$$f_{ij} = f(x_{2i}, x_{2j}) - f(x_{2i}, x_{2j-1}) - f(x_{2i-1}, x_{2j}) + f(x_{2i-1}, x_{2j-1}).$$

Write T_n for the measure that places mass one at each of the $4n(n-1)$ pairs (x_α, x_β) appearing in the definition of the f_{ij}. Clearly, $|T_n^0 f| \leq T_n|f|$ for every f; the measure T_n puts mass 1 at each support point of T_n^0. Notice that with the

addition of $2n$ extra support points T_n would be built up to a measure with the same distribution as S_{2n}.

LEMMA 1. *For each class $\mathcal{F}$ of P-degenerate, $P \otimes P$-integrable functions,*

$$\mathbb{P}\|S_n\| \le \mathbb{P}\|T_n^0\|.$$

PROOF. Denote expectations over the $\{\eta_i\}$, with the $\{\xi_i\}$ held fixed, by $\mathbb{P}_\eta$. For a fixed P-degenerate f,

$$|S_n f| = \left|S_n(f) - 2\mathbb{P}_\eta S_n'(f) + \mathbb{P}_\eta S_n''(f)\right| \le \mathbb{P}_\eta\left|T_n^0(f)\right|.$$

Take the supremum over $\mathcal{F}$ then average out over the $\{\xi_i\}$. □

Covering numbers. We will bound $\mathbb{P}\|T_n^0\|$ by working conditionally on the double sample $\{x_\alpha\}$. As with empirical processes, the argument will build up approximations to $\|T_n^0\|$ from the values of $T_n^0 f$ on finite subclasses of $\mathcal{F}$. Because $|T_n^0 f| \le T_n|f|$, the subclasses will be chosen to give good approximations in the sense of the $\mathcal{L}^1(T_n)$ or $\mathcal{L}^2(T_n)$ norms. The success of the argument will depend upon the existence of good bounds for the size of the approximating subclasses, that is, bounds on random covering numbers.

DEFINITION 2. Let S be a set equipped with a pseudometric d. The covering number $N(\varepsilon, d, S)$ is defined as the smallest value of N for which there exist N closed balls of radius ε, and centers in S, whose union covers S.

The concept is of interest only when $N(\varepsilon, d, S)$ is finite. It requires that there is a subset S^* of S, with cardinality $N(\varepsilon, d, S)$, for which

$$\min_{S^*} d(s, s^*) \le \varepsilon, \quad \text{for each } s \text{ in } S.$$

Even if S happens to be contained in some larger set, the definition insists that S^* be a subset of S. In the literature this constraint is not always imposed.

If the class $\mathcal{F}$ has an envelope F, and Q is a measure on $\mathcal{X} \otimes \mathcal{X}$ for which $0 < Q(F^p) < \infty$, we define the distance $d_{Q,p,F}$ on $\mathcal{F}$ by

$$d_{Q,p,F}(f, g) = \left[Q(|f - g|^p)/Q(F^p)\right]^{1/p}.$$

We write $N_p(\varepsilon, Q, \mathcal{F}, F)$ for the covering number $N(\varepsilon, d_{Q,p,F}, \mathcal{F})$. Thus $N_p(\varepsilon, Q, \mathcal{F}, F)$ is the smallest cardinality for a subclass $\mathcal{F}^*$ of $\mathcal{F}$ such that

$$\min_{\mathcal{F}^*} Q|f - f^*|^p \le \varepsilon^p Q(F^p), \quad \text{for each } f \text{ in } \mathcal{F}.$$

Notice that $N_p(\varepsilon, Q, \mathcal{F}, F)$ is unchanged if Q is replaced by a constant multiple of Q. We will use N_p covering numbers only for the cases $p = 1$ and $p = 2$. Section 5 summarizes some of the results for covering numbers that will be needed in Sections 3 and 4.

Exponential inequality. Application of an approximation argument based on random covering numbers depends upon the existence of good probabilistic

bounds for the increments of T_n^0, conditional on the $\{x_\alpha\}$. For empirical process, the Hoeffding inequality gives an exponential bound on the conditional tail probabilities of a symmetrized empirical process. A similar exponential inequality is available for T_n^0. (We thank G. Pisier and J. Zinn for the proof of the next lemma.)

LEMMA 3. *For each real symmetric matrix* $A = [a_{ij}]$ *with* $\sum_{i \neq j} a_{ij}^2 \leq (4\pi^2)^{-1}$,

$$\mathbb{P} \exp\Big(\sum_{i \neq j} \sigma_i \sigma_j a_{ij} \Big) \leq \exp\Big(\tfrac{1}{2}\pi^2 \sum_{i \neq j} a_{ij}^2 \Big).$$

PROOF. Let $\{g_i\}$ be independent $N(0,1)$ random variables that are also independent of the $\{\sigma_i\}$. Write V for the constant $2/\pi$, the square of the expected value $\mathbb{P}|N(0,1)|$. The left-hand side of the asserted inequality equals

$$\mathbb{P} \exp\Big(\sum_{i \neq j} \sigma_i \sigma_j V^{-1} \mathbb{P}|g_i g_j| a_{ij} \Big).$$

Because $\sigma_i |g_i|$ has the same distribution as g_i, Jensen's inequality bounds this expectation by

$$\mathbb{P} \exp\Big(V^{-1} \sum_{i \neq j} g_i g_j a_{ij} \Big).$$

We may assume that the diagonal elements of A are zero. Rotate to a new coordinate frame in which A is represented by a diagonal matrix of eigenvalues $\{\lambda_i\}$. The $\{g_i\}$ are transformed by the matrix of eigenvectors into $\{h_i\}$, a new collection of independent $N(0,1)$ random variables. The expectation becomes

$$\mathbb{P} \exp\Big(\sum_i \lambda_i h_i^2 / V \Big) = \prod_i (1 - 2\lambda_i / V)^{-1/2}.$$

Observe that

$$\sum_i \lambda_i = \operatorname{trace}(A) = 0,$$

$$\sum_i \lambda_i^2 = \operatorname{trace}(A^2) = \sum_{i,j} a_{ij}^2 \leq V^2/16.$$

So, by virtue of the inequality $\log(1 - x) \geq -x - x^2$, for $|x| \leq \frac{1}{2}$, the expectation is bounded by $\exp[\frac{1}{2}\sum_i (2\lambda_i / V)^2]$, which equals the asserted upper bound. □

COROLLARY 4.

$$\mathbb{P}_\sigma \{ T_n^0 f \geq \varepsilon \} \leq 2 \exp\Big(-\tfrac{1}{13} \varepsilon \big(T_n f^2 \big)^{-1/2} \Big),$$

where $\mathbb{P}_\sigma$ *denotes expectation conditional on the double sample.*

PROOF. The left-hand side is less than $\mathbb{P}_\sigma \exp(-\varepsilon/c + T_n^0 f / c)$. Put $c = 13(T_n f^2)^{1/2}$; then bound $\sum_{i \neq j} f_{ij}^2$ by

$$4 T_n f^2 = 4 \sum_{i \neq j} f(x_{2i}, x_{2j})^2 + f(x_{2i}, x_{2j-1})^2 + f(x_{2i-1}, x_{2j})^2 + f(x_{2i-1}, x_{2j-1})^2.$$

The 2 is a convenient upper bound for $\exp(\frac{1}{2}\pi^2 4/169)$. □

Chaining inequality. For the proof in Section 3 of a uniform convergence result analogous to the uniform strong law of large numbers for empirical processes, the exponential inequality of Lemma 3 and a bound on the size of covering numbers will suffice. For results on rates of convergence [and for the analogues of the empirical central limit theorem in Nolan and Pollard (1988)], we will need a sharper bound, based on a chain of approximations built up from a sequence of $\mathscr{L}^2(T_n)$ approximating classes for $\mathscr{F}$. We use a variation on a result of Pisier (1983). We modify Pisier's proof slightly to accommodate a slight relaxation of the condition on his Ψ function, and to avoid the appeal to his subtle Lemma 1.7.

We state the theorem in terms of a stochastic process $\{Z(s): s \in S\}$, where S is an index class equipped with a pseudometric $d(\cdot, \cdot)$. To avoid some circumlocution regarding separable versions, we assume Z has continuous sample paths. [For the $\mathscr{L}^2(T_n)$ pseudometric on $\mathscr{F}$, the paths of T_n^0 are continuous.]

LEMMA 5. *Let Ψ be a convex, strictly increasing function on $[0, \infty)$ with $0 \le \Psi(0) \le 1$. Suppose Z satisfies*:

(i) *if $d(s, t) = 0$, then $Z(s) = Z(t)$ almost surely*;
(ii) *if $d(s, t) > 0$, then $\mathbb{P}\Psi(|Z(s) - Z(t)|/d(s, t)) \le 1$*;
(iii) *there exists a point s_0 in S for which $\sup_S d(s, s_0) < \infty$*;
(iv) *the sample paths of Z are continuous.*

Then

$$\mathbb{P}\sup_S |Z(s) - Z(s_0)| \le 8\int_0^\theta \Psi^{-1}(N(x, d, S))\, dx,$$

where θ equals one quarter of the supremum in (iii).

PROOF. The assumptions about Ψ enter the proof through a simple inequality. Suppose $X_1, \ldots, X_n$ are random variables, and Δ is a real number such that $\mathbb{P}\Psi(|X_i|/\Delta) \le 1$ for each i. Then

$$\mathbb{P}\max_i |X_i| \le \Delta\Psi^{-1}(N).$$

This follows from Jensen's inequality:

$$\Psi\mathbb{P}\Big(\max_i |X_i|/\Delta\Big) \le \sum_i \mathbb{P}\Psi(|X_i|/\Delta) \le N.$$

We will apply the bound to the increments of Z.

Define $\delta_i = \theta/2^{i-1}$ for $i = 0, 1, \ldots$. Construct maximal subsets $\{s_0\} = S_0 \subseteq S_1 \subseteq \cdots$ of S with the property

$$d(s, t) > 2\delta_i, \quad \text{if } s, t \in S_i \text{ and } s \ne t.$$

By the definition of maximality, there is a map γ_i from S into S_i for which $d(s, \gamma_i s) \le 2\delta_i$. We may assume that $\gamma_i s = s$ if s is in S_i. Also S_i can contain no more than $N_i = N(\delta_i, d, S)$ points: Otherwise some δ_i-ball of a covering family would have to contain two distinct points of S_i.

Choose a positive integer k. For the moment hold it fixed; at the end of the proof it will be sent off to infinity. For each s in S_k define a chain of points

$$s_k = s,\ s_{k-1} = \gamma_{k-1}s_k, \ldots, s_1 = \gamma_1 s_2, s_0.$$

By construction $d(s_i, s_{i-1}) \le 2\delta_{i-1} = 8(\delta_i - \delta_{i+1})$. Thus,

$$\begin{aligned}
\mathbb{P}\max_{S_k}|Z(s) - Z(s_0)| &\le \sum_{i=1}^{k} \mathbb{P}\max_{S_i}|Z(s_i) - Z(s_{i-1})| \\
&\le \sum_{i=1}^{\infty} 8(\delta_i - \delta_{i+1})\Psi^{-1}(N_i) \\
&\le 8\int_0^{\delta_1}\Psi^{-1}(N(x, d, S))\,dx.
\end{aligned}$$

Let k tend to infinity, invoking monotone convergence and the continuity of the sample paths, to complete the proof. □

Maximal inequality. To bound $\mathbb{P}\|S_n\|$, we apply Lemma 5 to the symmetrized process, conditioning at first on the double sample.

THEOREM 6. *If $\mathcal{F}$ is a P-degenerate class with envelope F, then there exists a universal constant C such that*

$$\mathbb{P}\|S_n\| \le C\mathbb{P}\big(\theta_n + \tau_n J_n(\theta_n/\tau_n)\big),$$

where

$$J_n(s) = \int_0^s \log N_2(x, T_n, \mathcal{F}, F)\,dx,$$

$$\tau_n = \big(T_n F^2\big)^{1/2}, \qquad \theta_n = \tfrac{1}{4}\sup_{\mathcal{F}}\big(T_n f^2\big)^{1/2}.$$

PROOF. Define

$$\begin{aligned}
Z(f) &= T_n^0(f)/\tau_n, \\
\Psi(x) &= \tfrac{1}{4}\exp\big(x/4\pi - \tfrac{1}{8}\big), \\
d(f, g)^2 &= T_n(f - g)^2/\tau_n^2.
\end{aligned}$$

Write $\mathbb{P}_\sigma$ to denote expectation over the $\{\sigma_i\}$, conditioning on the double sample $\{x_\alpha\}$. Check the nonobvious conditions, (ii) and (iii), of Lemma 5. Since $d(f, 0) \le 4\theta_n/\tau_n$, condition (iii) is satisfied. Check condition (ii) for $Z(g_1) - Z(g_2)$. Put $f = g_1 - g_2$. Bound $e^{|x|}$ by $e^x + e^{-x}$ then apply Lemma 3. As in Corollary 4,

$$\sum_{i\ne j} f_{ij}^2 \le 4T_n f^2 = 4\tau_n^2 d(g_1, g_2)^2,$$

so

$$\begin{aligned}
\mathbb{P}_\sigma\exp\big(|Z(g_1) - Z(g_2)|/4\pi d(g_1, g_2)\big) &= \mathbb{P}_\sigma\exp\Big(|T_n^0 f|/4\pi\big(T_n f^2\big)^{1/2}\Big) \\
&\le 4\exp\Big(\tfrac{1}{2}\pi^2\sum_{i\ne j} f_{ij}^2/16\pi^2 T_n f^2\Big) \\
&\le 4\exp\big(\tfrac{1}{8}\big).
\end{aligned}$$

The invariance property that we built into the definitions of $N_p(\varepsilon, Q, \mathscr{F}, F)$ ensures that the covering numbers for d are the same as $N_2(x, T_n, \mathscr{F}, F)$, which we abbreviate to $N_2(x)$. Apply Lemma 5.

$$\begin{aligned}\mathbb{P}_\sigma\|T_n^0/\tau_n\| &\le 8\int_0^{\theta_n/\tau_n}\Psi^{-1}(N_2(x))\,dx \\ &\le 8\int_0^{\theta_n/\tau_n}\tfrac{1}{2}\pi + 4\pi\log 4N_2(x)\,dx \\ &\le \pi(4 + 32\log 4)\theta_n/\tau_n + 32\pi J_n(\theta_n/\tau_n).\end{aligned}$$

Multiply through by τ_n, and then average out over the $\{x_\alpha\}$. An appeal to the symmetrization inequality of Lemma 1 completes the proof. □

3. Uniform almost-sure convergence. For a fixed $P \otimes P$-integrable f, the sequence $\{S_n(f)/n(n-1)\}$ is a reversed martingale [Serfling (1980), Lemma 5.1.5B]. It converges almost surely to its expected value $P \otimes Pf$. We will strengthen the result by giving sufficient conditions for the convergence to hold uniformly over $\mathscr{F}$.

The conditions imposed on $\mathscr{F}$ and the method of proof will be similar to the conditions and methods for the uniform strong law for empirical processes [Pollard (1984), Section II.5]. But there are a few extra difficulties to overcome, because the symmetrization inequality for U-processes involves expectations. The corresponding inequality for empirical processes involves tail probabilities.

THEOREM 7. *Let $\mathscr{F}$ be a class of symmetric functions with $P \otimes P$-integrable envelope F, and P_n be the empirical measure. If for each $\varepsilon > 0$,*

(i) $\log N_1(\varepsilon, T_n, \mathscr{F}, F) = o_p(n)$,
(ii) $\log N_1(\varepsilon, P_n \otimes P, \mathscr{F}, F) = o_p(n)$,
(iii) $N_1(\varepsilon, P \otimes P, \mathscr{F}, F) < \infty$,

then $\|(S_n/n(n-1)) - P \otimes P\| \to 0$ almost surely.

PROOF. For notational convenience write α_n for $n(n-1)$. As with empirical processes, the supremum involved in the definition of $\|\cdot\|$ and the reversed martingale property of S_n make

$$R_n = \|S_n/\alpha_n - P \otimes P\|$$

a reversed submartingale. It converges almost surely. To prove the asserted convergence it therefore suffices to prove, for each $\varepsilon > 0$, that

$$\limsup \mathbb{P}\{R_n > \varepsilon\} < \varepsilon.$$

For a suitably large value of M the contributions from functions $f\{F > M\}$ can be made small

$$\begin{aligned}&\mathbb{P}\{\|S_n f\{F > M\}/\alpha_n - P \otimes Pf\{F > M\}\| \ge \varepsilon\} \\ &\quad\le \varepsilon^{-1}[\mathbb{P}S_n F\{F > M\}/\alpha_n + P \otimes PF\{F > M\}] \\ &\quad= 2\varepsilon^{-1}P \otimes PF\{F > M\} \\ &\quad< \varepsilon, \quad \text{if } M \text{ is large enough.}\end{aligned}$$

So, with fixed $\varepsilon > 0$, we may as well assume that $f = 0$ outside $\{F \le M\}$, for each f in $\mathscr{F}$. The value of M will depend on ε. This convenient simplification leaves intact assumptions (i), (ii) and (iii).

We decompose $S_n/\alpha_n - P \otimes P$ into a sum of an empirical process plus a degenerate U-process. Define

$$\tilde{f}(x, y) = f(x, y) - Pf(x, \cdot) - Pf(\cdot, y) + P \otimes Pf.$$

Clearly $\tilde{f}$ is degenerate. The corresponding decomposition for the U-process is

$$S_n(f)/\alpha_n - P \otimes P(f) = S_n(\tilde{f})/\alpha_n + 2(P_n - P) \otimes P(f).$$

Write $P\mathscr{F}$ for the class of all functions on $\mathscr{X}$ of the form $Pf(x, \cdot)$ with f in $\mathscr{F}$. From assumption (ii) and Lemma 20 (proved in Section 5 at the end of this paper), we get

$$\log N_1(\varepsilon, P_n, P\mathscr{F}, PF) = o_p(n).$$

This implies [Theorem II.24 of Pollard (1984)] uniform convergence to zero of the contribution from the empirical measure. It remains to show that $\|S_n(\tilde{f})\|/\alpha_n$ converges in probability to zero.

We will prove that $\mathbb{P}\|S_n(\tilde{f})\|/\alpha_n \to 0$, by means of Lemma 1. Because $\tilde{f}$ is degenerate, $\mathbb{P}\|S_n\tilde{f}\| \le \mathbb{P}\|T_n^0\tilde{f}\|$. We bound the conditional expectation over the $\{\sigma_i\}$, with the $\{x_\alpha\}$ held fixed, by means of an approximation based on covering numbers for the measure

$$\mu_n = T_n + 2\alpha_n P_n \otimes P + \alpha_n P \otimes P.$$

The assumptions placed on the covering numbers for T_n, $P_n \otimes P$ and $P \otimes P$ imply that

$$\log N_1(\delta, \mu_n, \mathscr{F}, F) = o_p(n),$$

for each fixed $\delta > 0$. As Corollary 15 in Section 5 will show, this is a simple consequence of the definition of covering numbers. Let $\mathscr{F}^*$ be a subclass of $\mathscr{F}$, of size $N_1(\delta, \mu_n, \mathscr{F}, F)$, with the property: to each f in $\mathscr{F}$ there is an f^* in $\mathscr{F}^*$ for which

$$\mu_n|f - f^*| \le \delta\mu_n F.$$

Because

$$\left|T_n^0(\tilde{f} - \tilde{f}^*)\right| \le \mu_n|f - f^*|,$$

it follows that

$$\mathbb{P}_\sigma\|T_n^0\tilde{f}\| \le \delta\mu_n F + \mathbb{P}_\sigma\Big(\max_{\mathscr{F}^*}|T_n^0\tilde{f}|\Big).$$

For a fixed $\tilde{f}$ with $|\tilde{f}| \le 4M$, the exponential inequality of Lemma 3 gives

$$\mathbb{P}_\sigma\exp\big(|T_n^0\tilde{f}|/nD\big) \le 2\exp\Big(\tfrac{1}{2}\pi^2\sum_{i\ne j}\tilde{f}_{ij}^2/n^2D^2\Big) < 4,$$

if the constant D is chosen large enough ($D = 8\pi M$ would suffice). Apply this to

each $\tilde{f}$, for f in $\mathscr{F}^*$:

$$\mathbb{P}_\sigma \| T_n^0 \tilde{f} \| \le \delta \mu_n F + nD \log \sum_{\mathscr{F}^*} \mathbb{P}_\sigma \exp\left(|T_n^0 \tilde{f}| / nD \right)$$

$$\le \delta \mu_n F + nD \log\left(4 N_1(\delta, \mu_n, \mathscr{F}, F) \right).$$

Because $\mu_n F / \alpha_n \to 4P \otimes P(F)$ almost surely and $\log N_1(\delta, \mu_n, \mathscr{F}, F) = o_p(n)$, we deduce that

$$\mathbb{P}_\sigma \| T_n^0 \tilde{f} \| / \alpha_n \to 0, \quad \text{in probability.}$$

The left-hand side is uniformly bounded by $4M$, by virtue of our simplifying assumption that $f = 0$ outside $\{F \le M\}$. Thus

$$\mathbb{P} \| S_n \tilde{f} \| / \alpha_n \le \mathbb{P} \| T_n^0 \tilde{f} \| / \alpha_n \to 0$$

as asserted. □

4. Rates of convergence. Under more stringent conditions on $\mathscr{F}$, the uniform convergence result of Theorem 7 can be improved to give a rate for the uniform almost-sure convergence of the U-process. In this section we will assume $\mathscr{F}$ to be P-degenerate. The same decomposition as in the proof of Theorem 7 would reduce calculations for a general $\mathscr{F}$ to the degenerate case. We also impose stronger conditions on the covering numbers.

DEFINITION 8. Call a class of functions $\mathscr{F}$ Euclidean for the envelope F if there exist constants A and V such that

$$N_1(\varepsilon, Q, \mathscr{F}, F) \le A \varepsilon^{-V}, \quad \text{for } 0 < \varepsilon \le 1,$$

whenever $0 < QF < \infty$. Call A and V the Euclidean constants for F.

If $\mathscr{F}$ is Euclidean, then for each $p > 1$,

$$N_p(\varepsilon, Q, \mathscr{F}, F) \le A 2^{pV} \varepsilon^{-pV}, \quad \text{for } 0 < \varepsilon \le 1,$$

whenever $0 < QF^p < \infty$. This follows from the definition of $N_1(2(\varepsilon/2)^p, \mu, \mathscr{F}, F)$ for the measure $\mu(\cdot) = Q(\cdot (2F)^{p-1})$.

The name Euclidean hints at an analogy with the finite-dimensional space $\mathbb{R}^V$, where the number of closed balls of radius ε needed to cover a bounded set increases like ε^{-V}. For many purposes Euclidean classes do indeed behave somewhat like bounded subsets of finite-dimensional spaces. Simple criteria for identifying Euclidean classes will be discussed in Section 5. These will suffice for the application to cross-validation of kernel density estimators in Example 11 at the end of this section.

With a degenerate f, the random variables $n^{-1}S_n(f)$ have a nontrivial limiting distribution [Serfling (1980), Theorem 5.5.2]. We should expect therefore to have to downweight $S_n(f)$ by slightly more than n^{-1} to get almost-sure convergence to zero. The appropriate weighting factor will depend on f.

THEOREM 9. *Let $\mathscr{F}$ be a Euclidean class of P-degenerate functions with envelope* 1. *Let $W(n, x)$ be a bounded weight function that is decreasing in both arguments and satisfies*

$$\sum_{n=1}^{\infty} n^{-1} \int_0^1 W(n, x)(1 + \log(1/x))\, dx < \infty.$$

If $\nu(\cdot)$ is a function on $\mathscr{F}$ for which $\nu(f) \geq \sup_x P|f(x, \cdot)|$, then

$$n^{-1}\left\| W\left(n, \nu(f)^{1/2}\right) S_n(f) \right\| \to 0, \quad \textit{almost surely.}$$

Here is the idea behind the proof. The precise details will be given after a preliminary lemma. We will use the maximal inequality of Theorem 6 and a stratification argument. We will stratify $\mathscr{F}$ into subclasses on which $\nu(f)$ is almost constant; $\mathscr{F}_i$ will contain those f for which $\delta_{i+1} \leq \nu(f)^{1/2} < \delta_i$, with $\{\delta_i\}$ decreasing geometrically. We will write $\|\cdot\|_i$ instead of $\|\cdot\|$ when the supremum is restricted to the subclass $\mathscr{F}_i$. If w_{ni} is a bound for $W(n, \nu(f)^{1/2})$ on $\mathscr{F}_i$, then

$$\mathbb{P}\left\{ \left\| W\left(n, \nu(f)^{1/2}\right) S_n/n \right\|_i \geq \varepsilon \right\} \leq \varepsilon^{-1} w_{ni} \mathbb{P} \|S_n/n\|_i.$$

Because $\|S_n/n(n-1)\|_i$ is a reversed submartingale, an even stronger, maximal inequality will hold; the random variable in the left-hand side can be increased to a maximum over a block of n with only a doubling of the upper bound.

For a Euclidean class, the covering integral $\int_0^s \log N_2(x, T_n, \mathscr{F}, 1)\, dx$ is bounded by a constant multiple of

$$H(s) = s\left[1 + \log(1/s)\right].$$

For a suitably large constant C_1, the maximal inequality of Theorem 6 gives

$$\mathbb{P}\|S_n/n\|_i \leq C_1 \mathbb{P} H\left(\|T_n f^2\|_i^{1/2}/n\right).$$

Because $\mathbb{P}T_n f^2/n^2$ is less than δ_i^2 for each f in $\mathscr{F}_i$, optimistically one might hope to bound $H(\|T_n f^2\|_i^{1/2}/n)$ by some multiple of $H(\delta_i)$. The integral condition of the theorem controls the sum of $w_{ni} H(\delta_i)$ over the strata. The next lemma, an analogue of Le Cam's (1983) square-root trick, will justify the optimistic upper bound, but only when δ_i is bigger than $n^{-1}\log n$. We will be forced to stop the stratification at a $k(n)$, where $\delta_{k(n)}$ is approximately $n^{-1}\log n$. The remaining functions, for which $\nu(f)$ is too small, will form a single class $\mathscr{F}_{k(n)}$. On $\mathscr{F}_{k(n)}$ we must use a pessimistic lower bound: $\nu(f) \geq 0$. Boundedness of the weight function takes care of the contributions from the $\mathscr{F}_{k(n)}$ classes.

LEMMA 10. *Let $\mathscr{G}$ be a Euclidean class of functions, with $0 \leq g \leq 1$ for each g in $\mathscr{G}$. There exists a positive constant β, which depends only upon the Euclidean constants A and V of $\mathscr{G}$ (for the envelope* 1*), such that: If*

$$t \geq \max\left\{ \sup_{\mathscr{G}} \sup_{x} \left(Pg(x, \cdot)\right)^{1/2}, n^{-1}\log n \right\},$$

then

$$\mathbb{P}\Big\{\sup_{\mathcal{G}} S_n g > \beta^2 n^2 t^2\Big\} \le 2A\exp(-nt).$$

PROOF. Construct independent samples $\{\xi_i\}$ and $\{\eta_i\}$ as for the symmetrization inequality in Section 2. Define S_n' and S_n'' as before. For each fixed g,

$$\mathbb{P}_\eta\{S_n'(g) \le 2n^2t^2\} \ge 1 - \mathbb{P}_\eta S_n'(g)/2n^2t^2 \ge \tfrac{1}{2}.$$

On the set where $\{\sup_{\mathcal{G}} S_n g > \beta^2 n^2 t^2\}$, there is a G for which

$$(S_nG)^{1/2} > \beta nt.$$

From the lower bound on the conditional probability,

$$\begin{aligned}\mathbb{P}\Big\{\sup_{\mathcal{G}} S_n g > \beta^2 n^2 t^2\Big\} &\le \mathbb{P}\big[\{S_nG > \beta^2 n^2 t^2\}2\mathbb{P}_\eta\{S_n'G \le 2n^2t^2\}\big] \\ &= 2\mathbb{P}\{(S_nG)^{1/2} > \beta nt, (2S_n'G)^{1/2} \le 2nt\} \\ &\le 2\mathbb{P}\Big\{\sup_{\mathcal{G}}(S_ng + S_n''g)^{1/2} - (2S_n'g)^{1/2} > (\beta-2)nt\Big\}.\end{aligned}$$

Write $\zeta(g)$ for $(S_ng + S_n''g)^{1/2} - (2S_n'g)^{1/2}$.

Condition on the double sample $\{x_\alpha\}$. Because T_n depends only on $\{x_\alpha\}$, the Euclidean property of $\mathcal{G}$ implies that there is a subclass $\mathcal{G}^*$ of $\mathcal{G}$ with cardinality at most At^{-2V} such that: To each g in $\mathcal{G}$ there is a g^* in $\mathcal{G}^*$ with

$$T_n|g - g^*| \le t^2 T_n(1) \le 4n^2t^2.$$

From the inequalities

$$(S_ng + S_n''g)^{1/2} \le (S_ng^* + S_n''g^*)^{1/2} + (S_n|g-g^*|)^{1/2} + (S_n''|g-g^*|)^{1/2},$$

$$\big|(S_n'g)^{1/2} - (S_n'g^*)^{1/2}\big| \le (S_n'|g-g^*|)^{1/2},$$

it follows that

$$\begin{aligned}\zeta(g) &\le \zeta(g^*) + 2(S_n|g-g^*| + 2S_n'|g-g^*| + S_n''|g-g^*|)^{1/2} \\ &= \zeta(g^*) + 2(T_n|g-g^*|)^{1/2} \\ &\le \zeta(g^*) + 4nt.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}_\sigma\Big\{\sup_{\mathcal{G}}\zeta(g) > (\beta-2)nt\Big\} &\le \mathbb{P}_\sigma\Big\{\max_{\mathcal{G}^*}\zeta(g) > (\beta-6)nt\Big\} \\ &\le At^{-2V}\max_{\mathcal{G}^*}\mathbb{P}_\sigma\big\{T_n^0 g \ge (\beta-6)nt(T_ng)^{1/2}\big\},\end{aligned}$$

because $[(S_ng + S_n''g)^{1/2} + (2S_n'g)^{1/2}]\zeta(g) = T_n^0 g$, and the factor in square brackets is bigger than $(T_ng)^{1/2}$. Corollary 4 bounds the last conditional probability by

$$2At^{-2V}\exp\big[-\tfrac{1}{13}(\beta-6)nt\big(T_ng/T_ng^2\big)^{1/2}\big],$$

which is less than

$$2A\exp\bigl[2V\log(1/t) - \tfrac{1}{13}(\beta - 6)nt\bigr],$$

because $g^2 \le g$. With β large enough, the exponent is less than $-nt$ for all t greater than $n^{-1}\log n$. □

PROOF OF THEOREM 9. Break $\mathscr{F}$ into strata and group the n into blocks. Write $n(j)$ for 2^j and γ_n for $n^{-1}\log n$. Define $k(j)$ as the value of k for which

$$2\gamma_{n(j)} > 2^{-k} \ge \gamma_{n(j)},$$

then put $\delta_i = 2^{-i}$ for $i = 0, \ldots, k(j)$ and $\delta_{k(j)+1} = 0$. Define

$$\mathscr{F}_i = \bigl\{f \in \mathscr{F}\colon \delta_{i+1} < \nu(f)^{1/2} \le \delta_i\bigr\},$$
$$\mathscr{N}(j) = \{n\colon n(j) \le n < n(j+1)\}.$$

Abbreviate the supremum over $\mathscr{F}_i$ to $\|\cdot\|_i$.

Because $W(n, \nu(f)^{1/2}) \le W(n(j), \delta_{i+1})$ when $n \in \mathscr{N}(j)$ and $f \in \mathscr{F}_i$, and because $\{\|S_n/n(n-1)\|_i\}$ is a reversed submartingale,

$$\begin{aligned}
&\mathbb{P}\Bigl\{\max_{\mathscr{N}(j)}\bigl\|W\bigl(n, \nu(f)^{1/2}\bigr)S_n(f)/n\bigr\|_i \ge \varepsilon\Bigr\}\\
&\quad\le \varepsilon^{-1}W(n(j), \delta_{i+1})(n(j+1) - 2)\mathbb{P}\max_{\mathscr{N}(j)}\|S_n/n(n-1)\|_i\\
&\quad\le \varepsilon^{-1}W(n(j), \delta_{i+1})2(n(j) - 1)\mathbb{P}\bigl\|S_{n(j)}/n(j)(n(j) - 1)\bigr\|_i\\
&\quad\le w_{ji}\mathbb{P}\|S_{n(j)}\|_i/n(j),
\end{aligned}$$

where $w_{ji} = 2\varepsilon^{-1}W(n(j), \delta_{i+1})$. Notice that this is slightly different from the w_{ni} used earlier.

Write $H(s)$ for $s[1 + \log(1/s)]$. For some constant C_1, Theorem 6 gives

$$\mathbb{P}\|S_{n(j)}\|_i/n(j) \le C_1\mathbb{P}H\bigl[\|T_{n(j)}f^2\|_i^{1/2}/n(j)\bigr].$$

Break the range of integration according to whether $\|T_{n(j)}f^2\|_i^{1/2}/n(j)$ is greater than $2\beta\delta_i$ or not, where β is the constant defined in Lemma 10. Because $0 \le f^2 \le 1$, and $T_{n(j)}$ is less than a measure that has the same distribution as $S_{2n(j)}$, and $\delta_i \ge \gamma_{n(j)} \ge \gamma_{2n(j)}$ for $0 \le i \le k(j)$, the lemma gives

$$\mathbb{P}\bigl\{\|T_{n(j)}f^2\|_i^{1/2} > \beta 2n(j)\delta_i\bigr\} \le 2A\exp\bigl(-2n(j)\delta_i\bigr) \le 2An(j)^{-2}.$$

Thus,

$$\mathbb{P}H\bigl[\|T_{n(j)}f^2\|_i^{1/2}/n(j)\bigr] \le H(2\beta\delta_i) + 2An(j)^{-2}.$$

For $0 \le i \le k(j)$ the right-hand side is bounded by a multiple of $H(\delta_i)$. Sum over i to combine the contributions from each stratum.

$$\begin{aligned}
&\mathbb{P}\Bigl\{\max_{\mathscr{N}(j)}\bigl\|W\bigl(n, \nu(f)^{1/2}\bigr)S_n(f)/n\bigr\| \ge \varepsilon\Bigr\}\\
&\quad\le C_2\sum_{i=0}^{k(j)} w_{ji}H(\delta_i)\\
&\quad\le C_3W(n(j), 0)H\bigl(\delta_{k(j)}\bigr) + C_3\int_0^1 W(n(j), x)(1 + \log 1/x)\,dx.
\end{aligned}$$

The factor $H(\delta_{k(j)})$ can be bounded by a multiple of $j^2 2^{-j}$; because W is bounded, the first term is summable over j. Monotonicity of W implies

$$W(n(j), x) \leq \sum_{n(j-1) \leq n < n(j)} 2n^{-1} W(n, x).$$

With this cosmetic substitution we get, for some constants C_4 and C_5,

$$\sum_j \mathbb{P}\left\{ \max_{\mathcal{N}(j)} \left\| W(n, \nu(f)^{1/2}) S_n(f)/n \right\| \geq \varepsilon \right\}$$

$$\leq C_4 + C_5 \sum_n n^{-1} \int_0^1 W(n, x)(1 + \log 1/x)\, dx,$$

which is finite by assumption. The asserted almost-sure convergence now follows by the Borel–Cantelli lemma. □

EXAMPLE 11. In improving upon an optimality theorem of Hall (1983) for cross-validated kernel density estimators, Stone (1984) proved subtle results for the rate of convergence of a particular U-process. A slightly different form of Stone's key lemma can be deduced from our Theorem 9.

To keep the exposition simple we restrict ourselves to the case of a bounded density $p(\cdot)$ on the real line and smoothing by a nonnegative, symmetric density function $K(\cdot)$. The extension to higher dimensions and smoothing kernels taking both positive and negative values is straightforward.

In place of Stone's assumption that K satisfy a Hölder continuity condition and have compact support, we impose the comparable condition that K be of bounded variation.

The estimator of the unknown $p(\cdot)$ is defined by convolution of the empirical measure with a rescaled K:

$$\hat{p}_\sigma(x) = (n\sigma)^{-1} \sum_{i=1}^n K\left(\frac{x - \xi_i}{\sigma} \right).$$

Stone's goal was to minimize the integrated squared error loss function

$$L_n(\sigma) = \int (\hat{p}_\sigma - p)^2 = \int \hat{p}_\sigma^2 - 2\int p\hat{p}_\sigma + \int p^2.$$

Of course, direct minimization is not possible if $p(\cdot)$ is unknown. However, one can minize an estimate of the loss function

$$M_n(\sigma) = \int \hat{p}_\sigma^2 - 2[\sigma n(n-1)]^{-1} \sum_{i \neq j} K\left(\frac{\xi_i - \xi_j}{\sigma} \right) + \int p^2.$$

This estimate is suggested by the heuristic

$$\int p\hat{p}_\sigma = P\hat{p}_\sigma \approx P_n \hat{p}_\sigma = (n^2\sigma)^{-1} \sum_{i,j} K\left(\frac{\xi_i - \xi_j}{\sigma} \right).$$

The usual cross-validation adjustment removes the over-large contributions for $i = j$.

Stone showed that the σ_M that minimizes $M_n(\cdot)$ does almost as well as the σ_L that minimizes $L_n(\cdot)$, in the sense that

$$L_n(\sigma_M)/L_n(\sigma_L) \to 1 \quad \text{almost surely.}$$

This can be proved by comparing both L_n and M_n with the expected value $\mathbb{P}L_n(\sigma)$. Indeed the result would follow immediately from:

$$\sup_\sigma \left| \frac{L_n(\sigma)}{\mathbb{P}L_n(\sigma)} - 1 \right| \to 0 \quad \text{almost surely,} \tag{12}$$

$$\sup_\sigma \frac{|L_n(\sigma) - M_n(\sigma) + Z_n|}{\mathbb{P}L_n(\sigma)} \to 0 \quad \text{almost surely,} \tag{13}$$

where Z_n is a random variable that does not depend on σ. These two would imply that

$$\sup_\sigma \left| \frac{L_n(\sigma) - M_n(\sigma) + Z_n}{L_n(\sigma)} \right| \to 0 \quad \text{almost surely,}$$

whence

$$(1 - o(1))L_n(\sigma_M) = M_n(\sigma_M) - Z_n \le M_n(\sigma_L) - Z_n = (1 + o(1))L_n(\sigma_L).$$

To establish (12) we rearrange $L_n(\sigma) - \mathbb{P}L_n(\sigma)$ as

$$\int (\hat{p}_\sigma - p_\sigma)^2 - \mathbb{P}\int (\hat{p}_\sigma - p_\sigma)^2 + 2\int (\hat{p}_\sigma - p_\sigma)(p_\sigma - p),$$

where

$$p_\sigma(x) = \mathbb{P}\hat{p}_\sigma = \sigma^{-1}\int K\left(\frac{x-y}{\sigma}\right)p(y)\,dy.$$

The third term, the cross product, can be handled by empirical process methods. Nolan (1986) has proved an appropriate limit theorem, using methods analogous to those of this paper; Pollard (1988) will offer an alternative approach. The difference between the first integral and its expectation corresponds to a degenerate U-process. To see this, define

$$G_{t,\sigma}(x) = K\left(\frac{t-x}{\sigma}\right) - \int K\left(\frac{t-y}{\sigma}\right)p(y)\,dy,$$

$$\Gamma_\sigma(x, y) = \sigma^{-1}\int G_{t,\sigma}(x)G_{t,\sigma}(y)\,dt.$$

A little algebra shows that

$$\begin{aligned}\int (\hat{p}_\sigma - p_\sigma)^2 &= \int (n\sigma)^{-2}\sum_{i,j} G_{t,\sigma}(\xi_i)G_{t,\sigma}(\xi_j)\,dt \\ &= (n\sigma)^{-1}\Gamma_\sigma(0,0) + (n^2\sigma)^{-1}\sum_{i\ne j}\Gamma_\sigma(\xi_i, \xi_j).\end{aligned}$$

Each Γ_σ is P-degenerate; the class of all Γ_σ is a candidate for Theorem 9.

It is easy to check that there exists a constant C for which

$$\sup_{x,y,\sigma} |\Gamma_\sigma(x, y)| \le C,$$

$$\sup_x P|\Gamma_\sigma(x, \cdot)| \le C(1 \wedge \sigma), \quad \text{for all } \sigma > 0.$$

We can rescale to make C equal to 1.

The assumption of bounded variation, which we impose upon K, is sufficient to make the $\{\Gamma_\sigma\}$ a Euclidean class. We will sketch the proof of this at the end of Section 5, to illustrate how easy it is to establish the Euclidean property in particular cases.

As Stone showed, there is another constant C_1 for which

$$\mathbb{P}L_n(\sigma) \ge C_1\left[(n\sigma)^{-1} + (\sigma^4 \wedge 1)\right], \quad \text{for } \sigma > 0.$$

Thus,

$$\sup_\sigma \left| \frac{\int(\hat{p}_\sigma - p_\sigma)^2 - \mathbb{P}\int(\hat{p}_\sigma - p_\sigma)^2}{\mathbb{P}L_n(\sigma)} \right| \le C_2 \sup_\sigma \left| W\left(n, \nu(\Gamma_\sigma)^{1/2}\right) n^{-1} S_n(\Gamma_\sigma) \right|,$$

where $\nu(\Gamma_\sigma) = 1 \wedge \sigma$ and $W(n, x) = (1 + nx^{10})^{-1}$. Since W is bounded by 1 and

$$\int_0^1 W(n, x)(1 + \log(1/x))\, dx = O\left(n^{-1/10} \log n\right),$$

the conditions of Theorem 9 are satisfied. The uniform convergence to zero of the U-process contribution to $(L_n - \mathbb{P}L_n)/\mathbb{P}L_n$ is established.

A similar analysis takes care of (13). Indeed,

$$L_n(\sigma) - M_n(\sigma) = 2[\sigma n(n-1)]^{-1} \sum_{i \ne j} K\left(\frac{\xi_i - \xi_j}{\sigma}\right) - 2P\hat{p}_\sigma,$$

which breaks into an empirical process contribution, $2(P_n - P)p_\sigma$, plus a U-process constructed from the degenerate functions

$$f_\sigma(x, y) = K\left(\frac{x - y}{\sigma}\right) - \sigma p_\sigma(x) - \sigma p_\sigma(y) + \sigma P p_\sigma.$$

The empirical process methods cited above show that $(P_n - P)(p_\sigma - p)$ is small compared to $\mathbb{P}L_n(\sigma)$, so we take $Z_n = -2(P_n - P)p$. Full details of this alternative to Stone's argument have appeared in Nolan (1986).

5. Covering numbers. We collect together in this section useful facts about covering numbers, some of which have become part of the empirical process folklore.

Remember that $N(\varepsilon, d, S)$ denotes the smallest number of closed balls of radius ε, and centers in S, needed to cover S.

The first result is particularly handy for covering numbers calculated for $\mathcal{L}^p$ metrics, where the underlying measure is a sum of several other measures. Similar results were given in inequality (3.11) of Alexander (1984) and in Problem II.24 of Pollard (1984).

LEMMA 14. *Let d, γ and δ be pseudometrics on S for which $d \le \gamma + \delta$. Then*

$$N(4r, d, S) \le N(r, \gamma, S)N(r, \delta, S).$$

PROOF. Let $\Gamma_1, \ldots, \Gamma_m$ be closed balls of γ radius r that cover S, and $\Delta_1, \ldots, \Delta_n$ be closed balls of δ radius r that cover S. From each nonempty $\Gamma_i \cap \Delta_j$ choose a point s_{ij}. There are at most mn such s_{ij}. Every point in $\Gamma_i \cap \Delta_j$ is no further than $2r$ away from s_{ij}, in both γ and δ distances. □

COROLLARY 15. *If $\mathcal{F}$ is a space of functions with envelope F, and μ and ν are measures for which $\mu(F^p) < \infty$ and $\nu(F^p) < \infty$ for some $p \ge 1$, then*

$$N_p(4r, \mu + \nu, \mathcal{F}, F) \le N_p(r, \mu, \mathcal{F}, F)N_p(r, \nu, \mathcal{F}, F).$$

LEMMA 16. *If $\mathcal{F}$ and $\mathcal{G}$ are classes of functions with envelopes F and G, and μ is a measure with $\mu(F^p) < \infty$ and $\mu(G^p) < \infty$, then the class*

$$\mathcal{F} + \mathcal{G} = \{f + g: f \in \mathcal{F}, g \in \mathcal{G}\},$$

with envelope $F + G$ satisfies

$$N_p(2r + 2s, \mu, \mathcal{F} + \mathcal{G}, F + G) \le N_p(r, \mu, \mathcal{F}, F)N_p(s, \mu, \mathcal{G}, G).$$

PROOF. Find functions $f_1, \ldots, f_m$ and $g_1, \ldots, g_n$ for which

$$\min_i \mu|f - f_i|^p \le r^p \mu F^p, \quad \text{for } f \text{ in } \mathcal{F},$$

$$\min_j \mu|g - g_j|^p \le s^p \mu G^p, \quad \text{for } g \text{ in } \mathcal{G}.$$

Then, with the appropriate i and j,

$$\begin{aligned}\left(\mu|f + g - f_i - g_j|^p\right)^{1/p} &\le r(\mu F^p)^{1/p} + s(\mu G^p)^{1/p} \\ &\le 2(r + s)\left(\mu(F + G)^p\right)^{1/p}.\end{aligned}$$

□

COROLLARY 17. *If $\mathcal{F}$ is Euclidean for envelope F and $\mathcal{G}$ is Euclidean for envelope G, then $\mathcal{F} + \mathcal{G}$ is Euclidean for envelope $F + G$.*

A class $\mathcal{D}$ of subsets of a set S is said to be a polynomial class (or a Vapnik–Červonenkis class) if there exists a polynomial $p(\cdot)$ for which

$$\text{cardinality}\{D \cap F: D \in \mathcal{D}\} \le p(|F|),$$

for every finite subset F of S. There are several simple criteria for a class to have this property [Dudley (1984), Section 9, and Pollard (1984), Section II.4].

LEMMA 18.

(i) *If $\mathcal{D}_1$ and $\mathcal{D}_2$ are polynomial classes, then so is the collection of all sets $D_1 \cup D_2$, $D_1 \cap D_2$ and D_i^c, with $D_i \in \mathcal{D}_i$.*

(ii) *If $\mathcal{G}$ is a finite-dimensional vector space of real functions, then the collection of all sets of the form $\{g \ge 0\}$ or $\{g \le 0\}$ or $\{g > 0\}$ or $\{g < 0\}$ is a polynomial class.*

It is not hard to prove that a class of sets $\mathscr{D}$ is a polynomial class if and only if the corresponding class of indicator functions is Euclidean for the envelope 1.

If f is a real-valued function on a set $\mathscr{X}$, define

$$\text{graph}(f) = \{(x,t) \in \mathscr{X} \otimes \mathbb{R} : 0 < t < f(x) \text{ or } 0 > t > f(x)\}.$$

Set theoretic properties of the graphs translate into bounds on the covering numbers.

LEMMA 19 [Pollard (1984), Lemma II.25]. *If* $\{\text{graph}(f): f \in \mathscr{F}\}$ *is a polynomial class of sets, then* $\mathscr{F}$ *is Euclidean for the envelope* $\sup_{\mathscr{F}}|f|$.

LEMMA 20. *Let* $\mathscr{F}$ *be a collection of functions on* $\mathscr{X} \otimes \mathscr{X}$ *with constant envelope* C. *Let* μ *and* ν *be finite measures. Then for each* $p \geq 1$,

$$N_p(r, \mu, \nu\mathscr{F}, C\nu(\mathscr{X})) \leq N_p\left(r\,\nu(\mathscr{X})^{-1/p}, \mu \otimes \nu, \mathscr{F}, C\right),$$

where $\nu\mathscr{F} = \{\nu f(x,\cdot): f \in \mathscr{F}\}$.

PROOF. Choose functions $f_1, \ldots, f_n$, with $n = N_p(r\nu(\mathscr{X})^{-1/p}, \mu \otimes \nu, \mathscr{F}, C)$ for which

$$\min_i \mu \otimes \nu|f - f_i|^p \leq r^p \mu \otimes \nu(C^p).$$

Then, for the appropriate i,

$$\begin{aligned}\mu|\nu f(x,\cdot) - \nu f_i(x,\cdot)|^p &\leq \mu \otimes \nu|f - f_i|^p(\nu(\mathscr{X}))^{p-1}, \quad \text{by Jensen's inequality} \\ &\leq r^p \mu\left[(C\nu(\mathscr{X}))^p\right].\end{aligned}$$

□

COROLLARY 21. *Let* $\mathscr{F}$ *be a uniformly bounded Euclidean class of functions on* $\mathscr{X} \otimes \mathscr{X}$. *For each finite measure* ν, *the class* $\nu\mathscr{F}$ *is Euclidean.*

LEMMA 22.

(i) *Let* $\rho(\cdot)$ *be a real-valued function of bounded variation on* $\mathbb{R}^+$. *The class of all functions on* $\mathbb{R}^d$ *of the form* $x \to \rho(|Ax + b|)$, *with* A *ranging over all* $m \times d$ *matrices and* b *ranging over* $\mathbb{R}^m$, *is Euclidean for a constant envelope.*

(ii) *Let* $\lambda(\cdot)$ *be a real-valued function of bounded variation on* $\mathbb{R}$. *The class of all functions on* $\mathbb{R}^d$ *of the form* $x \to \lambda(\alpha' x + \beta)$, *with* α *ranging over* $\mathbb{R}^d$ *and* β *ranging over* $\mathbb{R}$, *is Euclidean for a constant envelope.*

PROOF. As the arguments for both assertions are similar, we will prove only (i). By virtue of Lemma 16, it is enough to treat the two monotone components of $\rho(\cdot)$ separately. Assume, without loss of generality, that $\rho(\cdot)$ is bounded and nondecreasing, with $\rho(0) = 0$. Define $\rho^{-1}(\cdot)$ as the usual left-continuous inverse of $\rho(\cdot)$ on the range $T = (0, \sup\rho)$. Partition T into regions T_1 and T_2 such that

$$\{z \in \mathbb{R}^+: \rho(z) > t\} = \begin{cases} \left(\rho^{-1}(t), \infty\right), & \text{if } t \in T_1, \\ \left[\rho^{-1}(t), \infty\right), & \text{if } t \in T_2. \end{cases}$$

Then the graph of $\rho(|Ax + b|)$ can be written

$$\{t \in T_1, |Ax + b| > \rho^{-1}(t)\} \cup \{t \in T_2, |Ax + b| \geq \rho^{-1}(t)\}.$$

Define $g_{A,b}(x, t) = |Ax + b|^2 - (\rho^{-1}(t))^2$. The functions $g_{A,b}(\cdot, \cdot)$ span a finite-dimensional vector space. By part (ii) of Lemma 18, the sets $\{g_{A,b} > 0\}$ and $\{g_{A,b} \geq 0\}$ all belong to a polynomial class. Augment the class by the two sets $\mathbb{R}^d \otimes T_1$ and $\mathbb{R}^d \otimes T_2$; then appeal to Lemma 18(i) and Lemma 19 to complete the proof. □

As an illustration of the way in which these lemmas may be applied, we will show that the class $\{\Gamma_\sigma: \sigma > 0\}$, from Example 11, is Euclidean. Let ν be the finite measure that has density K. Define functions $f_\sigma(\cdot, \cdot, \cdot)$ on $\mathbb{R}^3$, for $\sigma > 0$, by

$$f_\sigma(x, y, z) = K\left(\frac{x - y}{\sigma} + z\right).$$

Lemma 22(ii) shows that $\{f_\sigma: \sigma > 0\}$ is a (subclass of a) Euclidean class. A simple change of variables gives

$$\Gamma_\sigma(x, y) = \nu f_\sigma(x, y, \cdot) - P \otimes \nu f_\sigma(x, \cdot, \cdot) - P \otimes \nu f_\sigma(\cdot, y, \cdot) + P \otimes P \otimes \nu f_\sigma(\cdot, \cdot, \cdot).$$

An appeal to Corollary 21, or its extension to functions of three variables, and Lemma 16 complete the argument.

Similar arguments would work in higher dimensions if K were a linear combination of two multidimensional distribution functions, or if $K(x)$ were of the form $k(|x|)$ with $k(\cdot)$ a function of bounded variation on the real line.

Acknowledgments. We thank Steve Marron, Charles Stone and a very careful referee for helpful criticisms and corrections.

REFERENCES

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

DUDLEY, R. (1984). A course on empirical processes. *Ecole d'Été de Probabilités de Saint-Flour, XII–1982. Lecture Notes in Math.* **1097** 1–142. Springer, New York.

HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 383–390.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174. Springer, New York.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

HOEFFDING, W. (1961). Unpublished paper, cited by Serfling (1980).

LE CAM, L. (1983). A remark on empirical measures. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 305–327. Wadsworth, Belmont, Calif.

NOLAN, D. (1986). *U*-processes. Ph.D. dissertation, Yale Univ.

NOLAN, D. and POLLARD, D. (1988). *U*-processes: functional limit theorems. *Ann. Probab.* To appear.

PISIER, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Banach Spaces, Harmonic Analysis, and Probability Theory. Lecture Notes in Math.* **995** 123–154. Springer, New York.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

POLLARD, D. (1988). Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. *Ann. Statist.* To appear.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
YALE UNIVERISTY
BOX 2179 YALE STATION
NEW HAVEN, CONNECTICUT 06520