The Annals of Statistics
1992, Vol. 20, No. 1, 469–489

ON THE LAST TIME AND THE NUMBER OF TIMES AN ESTIMATOR IS MORE THAN ε FROM ITS TARGET VALUE

BY NILS LID HJORT AND GRETE FENSTAD

University of Oslo and Norwegian Computing Centre, and University of Oslo

Suppose $\hat{\theta}_n$ is a strongly consistent estimator for θ_0 in some i.i.d. situation. Let N_ε and Q_ε be, respectively, the last n and the total number of n for which $\hat{\theta}_n$ is at least ε away from θ_0. The limit distributions for $\varepsilon^2 N_\varepsilon$ and $\varepsilon^2 Q_\varepsilon$ as ε goes to zero are obtained under natural and weak conditions. The theory covers both parametric and nonparametric cases, multidimensional parameters and general distance functions. Our results are of probabilistic interest, and, on the statistical side, suggest ways in which competing estimators can be compared. In particular several new optimality properties for the maximum likelihood estimator sequence in parametric families are established. Another use of our results is ways of constructing sequential fixed-volume or shrinking-volume confidence sets, as well as sequential tests with power 1. The paper also includes limit distribution results for the last n and the number of n for which the supremum distance $\|F_n - F\| \geq \varepsilon$, where F_n is the empirical distribution function. Other results are reached for $\varepsilon^{5/2} N_\varepsilon$ and $\varepsilon^{5/2} Q_\varepsilon$ in the context of nonparametric density estimation, referring to the last time and the number of times where $|f_n(x) - f(x)| \geq \varepsilon$. Finally, it is shown that our results extend to several non-i.i.d. situations.

1. Introduction and summary. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (i.i.d.) variables and suppose $\hat{\theta}_n$ is an estimator based on the first n observations which is strongly consistent for some parameter θ_0 of interest, that is, $\hat{\theta}_n$ converges almost surely (a.s.) to θ_0. How large must n be in order for $\hat{\theta}_n$ to be very close to θ_0?

This natural question can be made precise in several different ways. (i) We can ask for an m such that

$$\Pr\left\{\left|\hat{\theta}_n - \theta_0\right| \leq \varepsilon\right\} \geq 0.95 \quad \text{for all } n \geq m. \tag{1.1}$$

An approximate answer to this question is readily given in the traditional cases where one has convergence in distribution of $\sqrt{n}\,(\hat{\theta}_n - \theta_0)$ to some appropriate $N(0, \sigma_0^2)$. Then $\sqrt{n}\,\varepsilon/\sigma_0 \geq 1.96$ suffices and we find $m \doteq 1.96^2\sigma_0^2/\varepsilon^2$ (and the assumption of strong consistency is not needed). (ii) We might ask for simulta-

Received May 1990; revised May 1991.

AMS 1980 *subject classifications*. Primary 60F05, 62F12; secondary 60G15.

Key words and phrases. Asymptotic optimality, asymptotic relative efficiency, density estimation, Glivenko–Cantelli, the number of ε-misses, sequential confidence regions, tests with power 1, the last n.

neous closeness for all large n, with high enough probability, that is,

$$(1.2)\quad \Pr\{|\hat{\theta}_n - \theta_0| \le \varepsilon \text{ for all } n \ge m\} = \Pr\Big\{\sup_{n \ge m} |\hat{\theta}_n - \theta_0| \le \varepsilon\Big\} \doteq 0.95,$$

which also can be thought of as a requirement for a sequential fixed-width confidence interval procedure. There is a finite m solving this problem since $\sup_{n \ge m}|\hat{\theta}_n - \theta_0| \to_p 0$ when $\hat{\theta}_n \to \theta_0$ a.s. (iii) We could study the random variable

$$(1.3)\qquad N_\varepsilon = \sup\{n \ge 1: |\hat{\theta}_n - \theta_0| \ge \varepsilon\},$$

which by the assumption of strong consistency is finite with probability 1. One has

$$(1.4)\quad \Pr\{\varepsilon^2 N_\varepsilon \ge y\} = \Pr\{N_\varepsilon \ge m\} = \Pr\Big\{\sqrt{m} \sup_{n \ge m} |\hat{\theta}_n - \theta_0| \ge \sqrt{y_0}\Big\},$$

in which $m = \langle y/\varepsilon^2 \rangle$ is the smallest integer greater than or equal to y/ε^2, and $y_0 = m\varepsilon^2$ is close to y; $y_0 - \varepsilon^2 < y \le y_0$. This shows that problems (ii) and (iii) are closely related; $\varepsilon^2 N_\varepsilon$ has a limiting distribution if $\sqrt{m}\, \sup_{n \ge m}|\hat{\theta}_n - \theta_0|$ has.

While problem (i) is well-studied and solved, problems (ii) and (iii) are virtually unstudied, presumably because the random variables they are concerned with depend upon the full sequence of estimators and as such cannot be observed. They have nevertheless some immediate probabilistic and statistical appeal and provide information about the speed of convergence of $\hat{\theta}_n$ to its target value. Stating or proving almost sure convergence touches basic chords in both probabilists and statisticians. Since $\hat{\theta}_n \to \theta_0$ a.s. means nothing but that N_ε is a.s. finite, it appears natural to inquire about its approximate size, for example via its approximate distribution and approximate expected value. Even Serfling's physician [(1980), page 49] is interested in (ii) and (iii). The last n viewpoint also invites two competing estimation methods to be compared in terms of the limit distributions of their respective $N_{\varepsilon,1}$ and $N_{\varepsilon,2}$. There are also natural connections to sequential testing and sequential confidence sets.

This paper provides general solutions to (ii) and (iii) and several related problems. The answer to question (1.2) turns out to be $m \doteq 2.241^2\sigma_0^2/\varepsilon^2$, for example. In Section 2, the limit distribution of $\varepsilon^2 N_\varepsilon$ is found in the i.i.d. case, under natural conditions, also in the more laborious p-dimensional case, where a result is reached for a general distance function $\|\hat{\theta}_n - \theta_0\|$. The limit distribution is that of the maximum of a certain squared mean zero Gaussian process. Section 3 demonstrates that these natural conditions are fulfilled in important classes of cases, including smooth functions of averages and maximum likelihood estimators. Comparing estimators in terms of limit distributions for their N_ε's is seen to lead to the familiar expression of asymptotic relative efficiency, that of a ratio of inverse variances, in the one-parameter case. Our arguments establish still another asymptotic optimality property for the parametric maximum likelihood estimator sequence, also in the p-parameter case: No other sequence will have its tail stochastically faster included in a given ε-neighbourhood, regardless of distance measure used.

In Section 4, a somewhat more involved problem is solved, that of obtaining a limit distribution theorem for the last n for which the supremum distance $\|F_n - F\|$ exceeds ε, where F_n is the empirical distribution function, that is, for the last n in the Glivenko–Cantelli theorem. A certain optimality property for F_n is established. Section 5 considers the N_ε problem in a different context, that of nonparametric density estimation. In this situation, $\varepsilon^2 N_\varepsilon$ goes to infinity; it is $\varepsilon^{5/2} N_\varepsilon$ which has a limiting distribution. Section 6 goes back to the situation of Sections 2 and 3, is technical and demonstrates the convergence of $\varepsilon^2 E N_\varepsilon$ to the appropriate limit, again under natural conditions.

In Section 7 the general methods of earlier sections are used to establish convergence results for other natural quantities related to the full estimator sequence, like Q_ε, the number of times the estimator misses with more than ε. Once more there is an asymptotic optimality property for the parametric maximum likelihood method: No other estimator sequence has stochastically fewer ε-misses. And a result obtained for nonparametric density estimation is that the best smoothing parameter for the kernel method, in the sense of leading to the fewest ε-misses as well as to the smallest last n, is equal to 1.008 times the traditional suggestion. Finally Section 8 contains a number of additional results and remarks.

These problems have only rarely been discussed in the literature. Bahadur (1968) considered a variable similar to our N_ε and indeed asked (page 307) "What else can be said about N_ε [than that it is a.s. finite], especially for very small ε?" He derived only a log-log law for N_ε, however, and failed to find what he [also] was searching for (page 308): a simple N_ε-related criterion for comparison of estimators that would be equivalent to the traditional measure of asymptotic relative efficiency. We find such a relation, however, as mentioned above; see (2.5) and (7.2). Robbins, Siegmund and Wendel (1968) found in effect the limit distribution of $\varepsilon^2 N_\varepsilon$ for a one-sided version of the problem, but only in the case of a simple average of zero mean unit variance variables. They phrased their result in the more probabilistic guise of the last exit time for sample sums S_n outside the linear $n\varepsilon$ boundary. Kao (1978) generalised some of their results and proved convergence of moments under minimal conditions, but still only in the probabilistic random walk case just mentioned, which in our statistical reformulation corresponds to estimators of the simple i.i.d.-average form. Some results of Müller (1968, 1970, 1972) also turn out to be related to some of ours, as explained in Remark (i) in Section 2. Finally, Stute (1983) proved a more statistically inspired result, similar to ours of Section 2, but only for certain M-estimators of a simple one-dimensional location parameter. We emphasize that our results cover general classes of multidimensional estimators (and even an infinite-dimensional case, in Section 4) as well as general distance measures.

Our results give natural asymptotic relative efficiency measures for comparison of estimators, like (7.2) in the p-dimensional case, but are of first order and cannot distinguish between competing sequences with the same first order limit distribution. Some second order results are in Hjort and Fenstad (1991) and Hjort and Khasminskii (1991). These lead to measures of asymptotic relative deficiency in cases where the asymptotic relative efficiency is 1, and make it possible to exhibit estimator sequences that in the second order sense have the smallest possible expected number of ε-errors; see also 8E.

2. Limit distribution of $\varepsilon^2 N_\varepsilon$.

A simple but fundamental lemma is the following:

Lemma. *Consider i.i.d. variables Z_i with mean zero and variance* 1 *and let $S_n = \Sigma_{i=1}^n Z_i$. Then*

$$(2.1) \qquad \sqrt{m} \sup_{n \ge m} |S_n/n| \to_d W_{\max} = \sup_{t \ge 1} |W(t)/t| =_d \max_{0 \le s \le 1} |W(s)|,$$

where $W(\cdot)$ is the Brownian motion process.

Proof. By Donsker's theorem, $S_{[mt]}/\sqrt{m}$ converges in distribution to Brownian motion $W(t)$, a Gaussian mean zero process with independent increments and covariance function $\min(s, t)$, in each of the function spaces $D[b, c]$; see, for example, Billingsley (1968). Hence $\sqrt{m}\, S_{[mt]}/[mt]$ tends to $W(t)/t$ in $D[1, c]$; from which it follows, by continuity of the supremum mapping, that

$$\sqrt{m} \sup_{m \le n \le cm} \left| \frac{S_n}{n} \right| = \sup_{1 \le t \le c} \sqrt{m} \left| \frac{S_{[mt]}}{[mt]} \right| \to_d \sup_{1 \le t \le c} \left| \frac{W(t)}{t} \right| =_d \sup_{c^{-1} \le s \le 1} |W(s)|$$

[employing the trick that $W^*(s) = sW(1/s)$ is a new Brownian motion]. The stronger statement of the lemma follows from this provided we can demonstrate

$$\gamma_c = \limsup_{m \to \infty} \Pr\left\{ \sqrt{m} \sup_{n \ge cm} |S_n/n| \ge \delta \right\} \to 0$$

as c grows to infinity, for each given positive δ; see, for example, Billingsley's (1968) Theorem 4.2. But $\gamma_c \le 6.75/c\delta^2$ as a consequence of a special case of inequality (6.4) stated and proved in Section 6. □

This proves useful. Assume that a one-dimensional $\hat{\theta}_n$ admits a representation of the type

$$(2.2) \qquad \hat{\theta}_n - \theta_0 = \sigma_0 \bar{Z}_n + R_n,$$

where $\bar{Z}_n = S_n/n$ is the average of Z_i's that are i.i.d. with mean zero and variance 1, σ_0 is the standard deviation of the limiting distribution and R_n is the residual noise, typically of size $O_p(1/n)$. Define N_ε as in (1.3), let $y > 0$ be given and let m and $y_0 \doteq y$ be as in (1.4). Then, when $\varepsilon \to 0$, which is the same as $m \to \infty$,

$$\begin{aligned}\Pr\{\varepsilon^2 N_\varepsilon \geq y\} &= \Pr\{N_\varepsilon \geq m\}\\ &= \Pr\Big\{\sqrt{m}\sup_{n \geq m}|\hat{\theta}_n - \theta_0| \geq \sqrt{y_0}\Big\}\\ &= \Pr\Big\{\sigma_0\sqrt{m}\sup_{n \geq m}|S_n/n + \sigma_0^{-1}R_n| \geq \sqrt{y_0}\Big\} \to \Pr\{\sigma_0 W_{\max} \geq \sqrt{y}\},\end{aligned}$$

provided the R_n's are small enough. What is required is precisely that the difference between $\sqrt{m}\ \sup_{n \geq m}|\sigma_0 S_n/n + R_n|$ and $\sigma_0\sqrt{m}\ \sup_{n \geq m}|S_n/n|$ goes to zero in probability as m tends to infinity. For this it suffices that

$$D_m = \sqrt{m}\sup_{n \geq m}|R_n| \to_p 0, \tag{2.3}$$

since the absolute value of the difference is dominated by D_m. [Note that the requirement for convergence of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ to $N(0, \sigma_0^2)$ is the weaker $\sqrt{n}R_n \to_p 0$.] Accordingly, we have the basic result

$$\varepsilon^2 N_\varepsilon \to_d \sigma_0^2 W_{\max}^2 \tag{2.4}$$

for any estimator that admits representation (2.2) under condition (2.3). The next section demonstrates that (2.2) with (2.3) hold for smooth functions of averages and for maximum likelihood type estimators.

Remarks. (i) The lemma was proved using just familiar $D[1, c]$-convergence and an extra inequality for γ_c to take care of $[c, \infty)$. An alternative and in some sense more elegant approach is to demonstrate convergence in distribution of $S_{[mt]}/\sqrt{m}$ to $W(t)$ on the full halfline $[0, \infty)$, in some appropriate metric space of functions, and then apply the $x(\cdot) \to \sup_{t \geq 1}|x(t)/t|$ mapping. One 'appropriate space' is that of all right-continuous functions $x(\cdot)$ with left-hand limits satisfying $x(0) = 0$ and $\lim_{t \to \infty} x(t)/t = 0$, equipped with the topology induced by the norm $\sup_{t \geq 0}|x(t)|/\max\{1, t\}$. Convergence can indeed be proved using the tail inequality for γ_c, and is also related to what is proved in Müller (1968); see also Müller (1970, 1972).

(ii) Note that the limiting distribution in (2.4) is only dependent upon σ_0, and that the competition criterion of achieving the stochastically smallest limit distribution for N_ε becomes equivalent to that of achieving the smallest possible limiting variance.

(iii) Another optimality property for the maximum likelihood estimator: Consider estimation in some given parametric model. From the previous remark it is clear that under traditional regularity conditions (see Section 3), no other sequence of estimators will have its tail $\{\tilde{\theta}_n\colon n \geq m\}$ included in a given neighbourhood stochastically faster than the sequence of maximum likelihood solutions. Of course the same is true for the rather wide class of estimator sequences that are asymptotically equivalent to these, like Bayes estimators.

(iv) Asymptotic relative efficiency: Let $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ be two estimator sequences, both strongly consistent for θ_0 and suppose $\sqrt{n}(\hat{\theta}_{n,j} - \theta_0)$ tends to $N(0, \sigma_j^2)$, for $j = 1, 2$. The traditional notion of asymptotic relative efficiency (a.r.e.) is the limiting ratio of sample sizes needed by method 1 and method 2 to achieve some desirable accuracy. The classical formula for the a.r.e. of method 2 w.r.t. method 1 is a.r.e. = σ_1^2/σ_2^2, motivated from approximate risk or from approximate length of confidence intervals; see Serfling [(1980), page 50–52]. This measure also plays a natural role in the Pitman approach to comparing test statistics; see Serfling's Section 10.2. Define last n variables $N_{\varepsilon,1}$ and $N_{\varepsilon,2}$ for the two estimator sequences. The ratio of sample sizes viewpoint invites the following as natural measures of asymptotic relative efficiency:

$$(2.5)\quad \lim_{\varepsilon\to 0}\frac{\operatorname{med}\{N_{\varepsilon,1}\}}{\operatorname{med}\{N_{\varepsilon,2}\}} = \frac{\sigma_1^2}{\sigma_2^2} = \text{a.r.e.},\qquad \lim_{\varepsilon\to 0}\frac{EN_{\varepsilon,1}}{EN_{\varepsilon,2}} = \frac{\sigma_1^2}{\sigma_2^2} = \text{a.r.e.}$$

(see Section 6 for convergence of moments). This provides fresh and independent motivation for the a.r.e. measure; see also 7A, 7B and 8E.

Let us now turn to the p-dimensional case. Let N_ε be defined as in (1.3) but with respect to some given distance function $\|\hat{\theta}_n - \theta_0\|$ in $\mathscr{R}^p$, for example, ordinary Euclidean distance. We have primarily distances of the type $\{(x - y)'A(x - y)\}^{1/2}$ in mind, where A is symmetric and positive definite, but require only that $\|x\|$ is a function on p-vectors with the properties $\|x + y\| \le \|x\| + \|y\|$, $\|x\| = 0$ if and only if $x = 0$, $\|x_n\| \to \|x\|$ when $x_n \to x$, $\|ax\| = |a|\|x\|$ for scalars a and $\|x\| = \|(x_1, \dots, x_p)'\| \le c\sum_{i=1}^p |x_i|$ for some constant c. See 8D for other distances.

THEOREM. *Suppose that*

$$(2.6)\qquad \hat{\theta}_n - \theta_0 = \Sigma_0^{1/2}\frac{1}{n}\sum_{i=1}^n Z_i + R_n,$$

where the Z_i's are i.i.d. with zero mean and the $p \times p$ identity matrix as covariance matrix. Suppose further that

$$(2.7)\qquad D_m = \sqrt{m}\sup_{n\ge m}\|R_n\| \to_p 0;$$

in particular $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N_p(0, \Sigma_0)$. Let $G_p(s) = \Sigma_0^{1/2}W(s)$, where $W(s) = (W_1(s), \dots, W_p(s))'$ is a vector of p independent Brownian motions, each evaluated at the same s. Then, as ε tends to 0,

$$(2.8)\quad \varepsilon^2 N_\varepsilon \to_d G_{p,\max}^2 = \Big\{\sup_{0\le s\le 1}\|G_p(s)\|\Big\}^2 = \sup_{0\le s\le 1}\big\|\Sigma_0^{1/2}W(s)\big\|^2.$$

PROOF. Somewhat more elaborate arguments are necessary now. We prove first that

$$(2.9)\qquad \sqrt{m}\sup_{n\geq m}\|\Sigma_0^{1/2}S_n/n\| \to_d G_{p,\max} = \sup_{0\leq s\leq 1}\|\Sigma_0^{1/2}W(s)\|,$$

where again the S_n's are partial sums of the Z_i's. Observe first that the stochastic process $(S_{1,[mt_1]}/\sqrt{m},\dots,S_{p,[mt_p]}/\sqrt{m})'$, where $S_{j,n}$ is the jth component of S_n, converges to $(W_1(t_1),\dots,W_p(t_p))'$ in each $D[b,c]^p$, equipped with the product Skorohod topology. [This p-variate version of Donsker's theorem follows from the 1-variate theorem by tightness and finite-dimensional convergence.] By the continuous mapping theorem, $\Sigma_0^{1/2}\sqrt{m}\,S_{[mt]}/[mt]$ converges to $\Sigma_0^{1/2}W(t)/t$ in $D_p[1,c]$, the space of all right-continuous functions $[1,c]\to\mathscr{R}^p$ with left-hand limits, equipped with the Skorohod topology. And since the supremum mapping is continuous also,

$$\sqrt{m}\sup_{m\leq n\leq cm}\|\Sigma_0^{1/2}S_n/n\| = \sqrt{m}\sup_{1\leq t\leq c}\|\Sigma_0^{1/2}S_{[mt]}/[mt]\|$$

$$\to_d \sup_{1\leq t\leq c}\|\Sigma_0^{1/2}W(t)/t\| =_d \sup_{c^{-1}\leq s\leq 1}\|\Sigma_0^{1/2}W(s)\|.$$

Claim (2.9) follows since $\gamma_c = \limsup_{m\to\infty}\Pr\{\sqrt{m}\sup_{n\geq cm}\|\Sigma_0^{1/2}S_n/n\|\geq\delta\}$ tends to 0 as c grows, by a simple inequality relating this quantity to a sum of p one-dimensional analogues; see the proof of the lemma.

The rest of the proof follows from (2.9) and regularity condition (2.7). For let again m and y_0 be as in (1.4). Then

$$\Pr\{\varepsilon^2N_\varepsilon\geq y\} = \Pr\Big\{\sqrt{m}\sup_{n\geq m}\|\Sigma_0^{1/2}S_n/n + R_n\|\geq\sqrt{y_0}\Big\}$$

is seen to converge to $\Pr\{\sup_{0\leq s\leq 1}\|\Sigma_0^{1/2}W(s)\|\geq\sqrt{y}\}$, which is the same as $\varepsilon N_\varepsilon^{1/2}\to_d G_{p,\max}$ or $\varepsilon^2N_\varepsilon\to_d G_{p,\max}^2$. □

In a parametric model the maximum likelihood estimator sequence achieves the smallest possible limit covariance matrix and therefore also achieves the stochastically smallest possible limit distribution for N_ε, regardless of distance measure $\|\hat\theta_n-\theta_0\|$, see Remark (iii) above for the one-dimensional case.

COROLLARY. *Let conditions be as in the theorem and let* $\|\hat\theta_n-\theta_0\| = \{(\hat\theta_n-\theta_0)'\Sigma_0^{-1}(\hat\theta_n-\theta_0)\}^{1/2}$ *be* Σ_0*-weighted Mahalanobis distance. Then*

$$(2.10)\qquad \sqrt{m}\sup_{n\geq m}\|\hat\theta_n-\theta_0\|\to_d \chi_{p,\max}\quad\text{and}\quad \varepsilon^2N_\varepsilon\to_d\chi_{p,\max}^2,$$

as, respectively, $m\to\infty$ *and* $\varepsilon\to 0$, *where* $\chi_{p,\max}^2 = \max_{0\leq s\leq 1}\sum_{i=1}^pW_i(s)^2$.

In particular the limit distribution is the very same one in each estimation problem with p parameters. One can prove that (2.10) continues to hold when Σ_0 is replaced by a strongly consistent estimate $\hat\Sigma_n$ ($\hat\Sigma_n\to_p\Sigma_0$ does not suffice). Details are in Hjort and Fenstad (1990).

The (in some sense) natural extension of (2.5) to the p-dimensional case would be a.r.e. = $EH(\Sigma_1)/EH(\Sigma_2)$, where $H(\Sigma) = \max_{0 \le s \le 1} W(s)'\Sigma W(s)$, since $\varepsilon^2 EN_\varepsilon \to EH(\Sigma)$ under Euclidean distance. There is no simple formula for $EH(\Sigma)$, however. A simple explicit a.r.e. measure emerges in 7B.

3. Applications to special cases. In this section we confirm that the necessary regularity conditions (2.6) and (2.7) indeed pertain in the usual situations, both under parametric and nonparametric circumstances. In the one-dimensional case, suppose that $\hat{\theta}_n$ admits the representation

$$(3.1) \qquad \hat{\theta}_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \sigma_0 Z_i + R_n, \qquad R_n = \delta_n \overline{U}_n = \delta_n \frac{1}{n}\sum_{i=1}^{n} U_i,$$

where the U_i's are i.i.d. with mean zero and finite variance and $\delta_n \to 0$ a.s. Then (2.3) holds, since

$$D_m = \sqrt{m} \sup_{n \ge m} |\delta_n \overline{U}_n| \le \sup_{n \ge m} |\delta_n| \sqrt{m} \sup_{n \ge m} |\overline{U}_n| \to_p 0,$$

in that the second term has a limit in distribution, by the lemma of Section 1, and the first term tends to 0 in probability, by the definition of $\delta_n \to 0$ a.s. There is a similar result for the p-dimensional case: If (3.1) holds, with Σ_0 replacing σ_0^2 and where the U_i's are i.i.d. vectors with mean 0 and finite covariance matrix and δ_n is a matrix with components that all tend to 0 a.s., then D_m of (2.7) tends to zero in probability. This follows essentially by the one-dimensional argument. To see this, let norm(δ_n) be the matrix norm of δ_n, defined as the maximum of $\|\delta_n x\|$ over $\|x\| \le 1$; for Euclidean distance-norm norm(δ_n) is equal to the largest eigenvalue, for example. Then $\|\delta_n \overline{U}_n\| \le \text{norm}(\delta_n)\|\overline{U}_n\|$, which goes a.s. to 0 by the continuity of the $\|\cdot\|$ norm.

3A. *Smooth functions of averages.* Suppose $\hat{\theta}_n = h(\overline{B}_n)$ and $\theta_0 = h(b)$, where $\overline{B}_n$ is the average of i.i.d. variables B_i with $EB_i = b$ and Var $B_i = \tau^2$. If h has a continuous derivative in a neighbourhood of b, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = h'(b)\sqrt{n}\left(\overline{B}_n - b\right) + \left\{h'\left(\tilde{b}_n\right) - h'(b)\right\}\sqrt{n}\left(\overline{B}_n - b\right)$$

for some random $\tilde{b}_n$ between b and $\overline{B}_n$. This is as in (3.1) with $\delta_n = h'(\tilde{b}_n) - h'(b)$. But it is easy to see that $\delta_n \to 0$ a.s. by the strong law of large numbers for $\overline{B}_n$. Hence (2.4) holds, with $\sigma_0^2 = h'(b)^2\tau^2$. More generally, suppose $\hat{\theta}_n$ is p-dimensional and that $\hat{\theta}_j = h_j(\overline{B}_{n,1}, \dots, \overline{B}_{n,r})$ for $j = 1, \dots, p$, for r averages of i.i.d. vectors $(B_{i,1}, \dots, B_{i,r})'$ with mean $b = (b_1, \dots, b_r)'$ and finite covariance matrix T, and let $h_j(b) = \theta_{0,j}$. If only $h_1, \dots, h_p$ have Jacobi matrix $J(x)$ with partial derivatives $\partial h_j(x)/\partial x_l$ that are continuous in a neighbourhood of $(b_1, \dots, b_r)$, then (2.7) holds. And this implies (2.8) with $\Sigma_0 = J(b)TJ(b)'$.

Example 1. Suppose $X_1, X_2, \dots$ are i.i.d. with finite sixth moment. Then $\hat{\theta}_n = (1/n)\Sigma_{i=1}^n (X_i - \overline{X})^3$, the natural and strongly consistent estimator of

$\theta_0 = E(X_i - EX_i)^3$, is a smooth function of the sample averages of X_i, X_i^2, X_i^3 and $\varepsilon^2 N_\varepsilon \to_d \sigma_0^2 W_{\max}^2$, where σ_0^2 is the limit variance of $\sqrt{n}\,(\hat{\theta}_n - \theta_0)$. [In fact, $\sigma_0^2 = (9 + \alpha_6 - 6\alpha_4 - \alpha_3^2)\tau^6$, where $\alpha_p = E(X_i - EX_i)^p/\tau^p$ and τ is the standard deviation for X_i.]

3B. *Maximum likelihood estimators.* The typical argument that leads to a limit distribution result for the maximum likelihood estimator uses Taylor expansion to get

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = J_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(X_i) \to_d N_p\{0, J_0^{-1}\}, \tag{3.2}$$

where $U(X_i) = \partial \log f(X_i, \theta_0)/\partial\theta$ is the score function and $J_n \to_p J_0$, the variance matrix for $U(X_i)$ computed under $f(x, \theta_0)$, that is, the familiar Fisher information matrix. It follows from previous arguments that (3.2) also secures convergence in distribution of $\varepsilon^2 N_\varepsilon$, as in (2.8), with $\Sigma_0 = J_0^{-1}$, provided there also is a.s. convergence $J_n \to J_0$. But this is true under weak conditions. It is, for example, not difficult to prove that the conditions used in Lehmann's (1983) Section 6.4 suffice.

One can also prove that if the model specifies $f(x, \theta)$, but the true density f does not belong, then (2.8) holds again under mild regularity conditions, but with a different interpretation of θ_0 and a different matrix. The θ_0 that now enters is not "true," but rather "least false" or "best fitting" and can be characterised as the parameter value that minimises the Kullback–Leibler distance $d[f\colon f(\cdot, \theta)] = \int f(x) \log\{f(x)/f(x,\theta)\}\,dx$. Furthermore, $\Sigma_0 = J_0^{-1} K_0 J_0^{-1}$, where K_0 is the variance matrix, under the true f, of the score function computed at θ_0; and J_0 is minus the expected value, under the true f, of the twice differentiated log-density also computed at θ_0. If the model happens to be perfect, then θ_0 deserves to be called "true" and $J_0 = K_0$. Proofs and discussion of these claims about maximum likelihood under the agnostic viewpoint can be found in Hjort [(1986), Chapter 3].

Example 2. Consider again maximum likelihood estimation in a given parametric family $f(x, \theta)$. Let distance be measured in the invariant Mahalanobis way, $\|\theta - \theta_0\|^2 = (\theta - \theta_0)' J_0 (\theta - \theta_0)$, and let N_ε be the last n for which $\|\hat{\theta}_n - \theta_0\| \geq \varepsilon$. Then $\varepsilon^2 N_\varepsilon$ tends to $\max_{0 \leq s \leq 1} W(s)' J_0^{-1/2} K_0 J_0^{-1/2} W(s)$ (which is $\chi^2_{p,\max}$ if the model is correct). For a specific example, suppose the model specifies the normal density $f(x, \theta) = f(x, \mu, \sigma)$, but assume only that the true f is symmetric with finite fourth moment. Then the least false parameters are $\mu_0 = E_f X_i$ and $\sigma_0 = \text{stdev}_f X_i$. One also finds $J_0^{-1/2} K_0 J_0^{-1/2} = \text{diag}(1, 1 + \beta_2/2)$, where $\beta_2 = E\{(X - \mu_0)/\sigma_0\}^4 - 3$ is the kurtosis. Hence $\varepsilon^2 N_\varepsilon$ tends to $\max_{0 \leq s \leq 1}\{W(s)^2 + (1 + \beta_2/2) W_2(s)^2\}$, where W_1 and W_2 are independent Brownian motions.

Example 3. Let $(Y_1, \dots, Y_p)$ be multinomial $(n, \theta_1, \dots, \theta_p)$, with $\sum_{i=1}^{p} \theta_i = 1$ and $\sum_{i=1}^{p} Y_i = n$. Let N_ε be the last n at which $\sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2/\theta_i \geq \varepsilon^2$, where

$\hat{\theta}_i = Y_i/n$ is the usual maximum likelihood estimator of θ_i. This corresponds in fact to measuring distance from $(\hat{\theta}_1, \dots, \hat{\theta}_{p-1})$ to $(\theta_1, \dots, \theta_{p-1})$ in the Mahalanobis way; see the Corollary ending Section 2. Hence $\varepsilon^2 N_\varepsilon$ tends to $\chi^2_{p-1,\max}$. The same is true if θ_i's are replaced with $\hat{\theta}_i$'s in the denominators; see Section 8F.

3C. *Differentiable functionals.* In many situations, the estimator $\hat{\theta}_n$ can be thought of as a functional T evaluated at the empirical distribution function F_n, while the true parameter θ_0 correspondingly is equal to $T(F)$ for the true F. Suppose T is so-called locally Lipschitz differentiable at F w.r.t. the supremum norm $\|G - F\| = \sup_x |G(x) - F(x)|$, which means that $T(G) - T(F) = \int I(F, x)\{dG(x) - dF(x)\} + O(\|G - F\|^2)$, featuring the influence function $I(F, x) = \lim_{\varepsilon \to 0}\{T((1 - \varepsilon)F + \varepsilon\delta_x) - T(F)\}/\varepsilon$. This might be interpreted as a reasonable minimum amount of smoothness on the part of $T(\cdot)$. Examples are given in Shao (1989), including general L- and M-estimators. In particular the somewhat nonsmooth median functional is still locally Lipschitz differentiable. Under this assumption it holds that

$$\hat{\theta}_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} I(F, X_i) + O\big(\|F_n - F\|^2\big).$$

But it is known that $\|F_n - F\|^2 \le K n^{-1} \log\log n$ a.s., for some large K; see, for example, Shao (1989). It follows that $|R_n| \le K' n^{-1} \log\log n$ in representation (2.2) and this implies (2.3). Consequently (2.4), or (2.8) in the p-dimensional case, are true for functionals that are locally Lipschitz differentiable.

4. The last n for Glivenko–Cantelli.

Let $X_1, X_2, \dots$ be independent from some continuous F and let $F_n(t)$ be the empirical distribution function $(1/n)\sum_{i=1}^{n} I\{X_i \le t\}$ based on the first n data points. Then

$$\sqrt{m}\sup_{n \ge m} |F_n(t) - F(t)| \to_d \{F(t)(1 - F(t))\}^{1/2} W_{\max},$$
$$(4.1) \qquad \varepsilon^2 N_\varepsilon(t) \to_d F(t)(1 - F(t)) W^2_{\max},$$

by previous efforts, where $N_\varepsilon(t)$ is the last n for which $|F_n(t) - F(t)| \ge \varepsilon$. Can we obtain similar results for the supremum distance $\|F_n - F\|$?

The answer to these somewhat more involved questions must involve asymptotic arguments in n and t simultaneously. Let $K_0(s, t)$ be a Kiefer process on $[0, \infty) \times [0, 1]$. This is a two-parameter zero mean Gaussian process with continuous sheets and

$$(4.2) \qquad \operatorname{cov}\{K_0(s_1, t_1), K_0(s_2, t_2)\} = (s_1 \wedge s_2)(t_1 \wedge t_2 - t_1 t_2).$$

It behaves like a Brownian bridge in t for fixed s and like Brownian motion in s for fixed t. Note that $K(s, t) = s K_0(s^{-1}, t)$ is another Kiefer process.

THEOREM. *Let N_ε be the last n at which $\|F_n - F\| \ge \varepsilon$ and let $K_{\max}$ be the maximum of $|K(s,t)|$ over the unit square* $[0,1] \times [0,1]$. *Then*

$$\sqrt{m} \sup_{n \ge m} \|F_n - F\| \to_d K_{\max} \quad \text{and} \quad \varepsilon^2 N_\varepsilon \to_d K_{\max}^2,$$

as, respectively, $m \to \infty$ *and* $\varepsilon \to 0$.

PROOF. Considerations involving the inverse transformation $X_i' = F^{-1}(\xi_i)$, where the ξ_i's are i.i.d. from the uniform distribution $F_0(t) = t$ on $[0,1]$, reveal that the distribution of the full sequence of $\|F_n - F\|$ is equal to that of $\|F_{n,0} - F_0\|$, where $F_{n,0}$ is the empirical distribution of the n first ξ_i's. Accordingly we might as well take F to be F_0 from the outset, and this simplifies matters below.

The Le Cam–Bickel–Wichura–Müller theorem states that the process

$$K_m(s,t) = \frac{1}{\sqrt{m}} \sum_{i=1}^{[ms]} [I\{X_i \le t\} - F_0(t)] = \frac{[ms]}{\sqrt{m}} \{F_{[ms]}(t) - t\}$$

converges in distribution to $K_0(s,t)$ in $D\{[b,c] \times [0,1]\}$ with the Skorohod metric for each $[b,c]$ interval; see, for example, Shorack and Wellner [(1986), Chapter 3.5]. For us it is more convenient to study

$$H_m(s,t) = \sqrt{m}\{F_{[ms]}(t) - t\} = \frac{m}{[ms]} K_n(s,t) \to_d s^{-1} K_0(s,t) = K(s^{-1},t).$$

[The (4.1) results follow anew from this.] By the continuous mapping theorem,

$$\sqrt{m} \sup_{n \ge m} \|F_n - F_0\| = \sup_{s \ge 1} \sup_{0 \le t \le 1} |H_m(s,t)| \to_d \sup_{s \ge 1} \sup_{0 \le t \le 1} |K(1/s,t)| = K_{\max}.$$

Reasoning once more as in (1.4), we also obtain $\varepsilon^2 N_\varepsilon \to_d K_{\max}^2$. This incidentally also gives a sequential fixed-width nonparametric simultaneous confidence band for F.

The argument presented here is heuristic at one point, since convergence in distribution of the H_m process is only guaranteed on each $[1,c] \times [0,1]$. Therefore only convergence of $\sqrt{m}\, \sup_{m \le n \le cm} \|F_n - F\|$ to the maximum of $|K(s,t)|$ over $[1/c, 1] \times [0,1]$ is rigorously proved, so far. What needs to be ascertained is that

$$(4.3) \qquad \gamma_c = \limsup_{m \to \infty} \Pr\left\{\sqrt{m} \sup_{n \ge cm} \|F_n - F_0\| \ge \delta\right\} \to 0 \quad \text{as } c \to \infty;$$

see Billingsley's (1968) Theorem 4.2 and the corresponding technical point in the proof of the lemma of Section 2. It will suffice to prove

$$(4.4) \qquad \Pr\left\{\sqrt{m} \sup_{n \ge m} \|F_n - F_0\| \ge b\right\} \le A/b^4 \quad \text{for all } b \text{ and } m$$

for some large enough constant A, since this implies $\gamma_c \le A/c^2\delta^4$. [An alternative route is to prove weak convergence in some appropriate function space on $[1,\infty)\times[0,1]$, with suitable metric. A result of Müller (1970) is of this type. Yet another way is via strong Hungarian approximations, as pointed out to us by David Pollard.]

To prove (4.4) we shall use general fluctuation inequalities provided by Bickel and Wichura (1971) for two-parameter processes. For neighbouring blocks B and C in the unit square, one can show $E\{K_m(B)^2K_m(C)^2\}\le 3\mu(B)\mu(C)$, where μ is Lebsegue measure, see Shorack and Wellner [(1986), Chapter 3.5]. This implies $\Pr\{\sup_{s,t\in[0,1]}|K_m(s,t)|\ge b\}\le A/b^4$ for some universal constant A, by Bickel and Wichura's Theorem 1 in conjunction with their inequality (1). But

$$\frac{1}{2}\sqrt{m}\max_{m/2\le n\le m}\|F_n-F_0\|\le\sqrt{m}\max_{m/2\le n\le m}\frac{n}{m}\|F_n-F_0\|$$
$$\le\sqrt{m}\max_{n\le m}\frac{n}{m}\|F_n-F_0\|=\sup_{s,t\in[0,1]}|K_m(s,t)|.$$

This is soon translated into $\Pr\{\sqrt{m}\max_{m\le n\le 2m}\|F_n-F_0\|\ge b\}\le A/(b/\sqrt{2})^4$ for all m and b. Let $2^k\le m<2^{k+1}$. Then the left-hand side of (4.4) is bounded by the sum of $\Pr\{\sqrt{2^i}\max_{2^i\le n<2^{i+1}}\|F_n-F_0\|\ge\sqrt{2^i}b/\sqrt{m}\}$ for $i\ge k$. Bounding each of these in the way just described gives at the end of the night (4.4), with constant $64A/3$, which concludes our proof. □

The two-parameter stochastic process approach is very powerful and allows us to reach other related results as well. As but one example, let $\mathrm{CM}_n^2=\int\{F_n(t)-F(t)\}^2\,dF(t)$ be the Cramér–von Mises statistic. Using the H_m process from the proof above we have

$$m\sup_{n\ge m}\mathrm{CM}_n^2=_d\sup_{s\ge 1}\int_0^1 H_m(s,t)^2\,dt\to_d\sup_{0\le s\le 1}\int_0^1 K(s,t)^2\,dt=\Lambda^2.$$

We also have $\varepsilon^2N_\varepsilon\to\Lambda^2$ if N_ε is the last n where $\mathrm{CM}_n\ge\varepsilon$. [Ways of simulating the distributions of Λ^2 and $K_{\max}$ are described in Hjort and Fenstad (1990).]

Asymptotic optimality of F_n. Are there estimators better than F_n, as measured by expected smallness of N_ε as ε tends to 0? The answer to this question is no, if one rules out the superefficiency phenomenon. This follows from the Hájek convolution type representation theorem for limit distributions for $\sqrt{n}(\tilde{F}_n-F)$ proved by Beran (1977), in conjunction with the arguments used above.

5. The last n for nonparametric density estimators. Consider a kernel type estimator $f_n(x)=(1/n)\sum_{i=1}^nK((x-X_i)/h_n)/h_n$ for the unknown density $f(x)$ based on the first n data points in an i.i.d. sequence. What is the size of N_ε, the last time $|f_n(x)-f(x)|\ge\varepsilon$? Techniques from Sections 2 and 3 can be employed to reach a limit distribution result though some extra care is needed since h_n varies with sample size (the minimum requirement for strong consistency is $h_n \to 0$ and $nh_n \to \infty$).

Suppose f has two continuous derivatives around the given x and let the kernel density K have mean zero and unit variance and finite $\beta_K = \int K(u)^2\,du$. Let $h_n = cn^{-1/5}$, well known to be the optimal rate. Study $Z_m(t) = m^{2/5}\{f_{[mt]}(x) - f(x)\}$ for the fixed x. It splits into a bias term $b_m(t)$ and a zero mean term $Z_m^0(t)$. The first can be seen to converge to the function $c^2f''(x)/(2t^{2/5})$, uniformly over finite t-intervals. The second can be proved to converge in distribution to a Gaussian zero mean process with covariance function of the form $c^{-1}g(s/t)f(x)/t^{4/5}$, where in fact $g(z) = z^{1/5}\int K(u)K(z^{1/5}u)\,du$. Hence

$$Z_m(t) \to_d Z(t) = \left[\tfrac{1}{2}c^2f''(x) + c^{-1/2}f(x)^{1/2}V(t)\right]\Big/t^{2/5}$$

for a certain normal zero mean stochastic process $V(\cdot)$ with constant variance β_K. From this result, using arguments parallel to those of Section 2, it is not difficult to derive

$$\varepsilon^{5/2}N_\varepsilon \to_d Z_{\max}^{5/2} = \Big\{\sup_{t\ge 1}|Z(t)|\Big\}^{5/2}.$$

How should c in $cn^{-1/5}$ be chosen? The approximate mean squared error is $h^4f''(x)^2/4 + \beta_K f(x)/nh$ and is minimised for $h_n = c_0(x)n^{-1/5}$, where $c_0(x) = \{\beta_K f(x)/f''(x)^2\}^{1/5}$. One version of the variable kernel approach to density estimation is to aim for this value, using a smooth pilot estimate to reach $\hat{c}_0(x)$, say. We could try to make EN_ε as small as possible by making $E|Z(t)|^{5/2}$ as small as possible. But this expectation can be written $a^{-5/4}E|a^{5/4}/2 + N(0,1)|^{5/2}$ times other terms not depending on c, where $a = c/c_0(x)$. Careful numerical integrations reveal that minimum occurs for $a_0 = 1.008$. Hence $1.008c_0(x)n^{-1/5}$ is best from the EN_ε point of view; see also 7D.

The derivative of f is even more difficult to estimate with good precision. This is reflected in high values for N'_ε, the last n for which $|f'_n(x) - f'(x)|$ exceeds ε. By techniques similar to those sketched above, one can prove that $\varepsilon^{7/2}N'_\varepsilon$ tends to some appropriate $(Z'_{\max})^{7/2}$ in distribution.

It would be interesting to reach results for N_ε's connected to global deviance measures like $\int(f_n - f)^2/f\,dx$ or the statistically natural but technically difficult $\int|f_n - f|\,dx$ as well. Techniques from Bickel and Rosenblatt (1973) would be appropriate but we have not pursued this.

6. Convergence of moments. We have proved that $\varepsilon^2N_\varepsilon \to_d \sigma_0^2W_{\max}^2$ (in the one-dimensional case) and it is clear that $\varepsilon^2EN_\varepsilon$ should tend to $\sigma_0^2EW_{\max}^2$ under conditions pertaining to uniform integrability. The present section derives this and a couple of related results under natural conditions.

We should like to prove

$$E\varepsilon^2N_\varepsilon = \int_0^\infty \Pr\{\varepsilon^2N_\varepsilon \ge y\}\,dy \to \int_0^\infty \Pr\{\sigma_0^2W_{\max}^2 \ge y\}\,dy = E\sigma_0^2W_{\max}^2,$$

and this holds by Lebesgue's theorem on dominated convergence provided we can bound

$$\Pr\{\varepsilon^2 N_\varepsilon \geq y\} \tag{6.1}$$
$$= \Pr\Big\{\sqrt{m} \sup_{n \geq m} \big|\hat{\theta}_n - \theta_0\big| \geq \sqrt{y_0}\Big\}, \qquad m = \langle y/\varepsilon^2 \rangle, \qquad y_0 = m\varepsilon^2,$$

with some integrable function uniformly in ε. A sufficient condition is therefore that for some positive ε_0,

$$\Pr\Big\{\sqrt{m} \sup_{n \geq m} \big|\hat{\theta}_n - \theta_0\big| \geq a\Big\} \leq K/a^{2+\lambda} \quad \text{when } 0 < a/\sqrt{m} \leq \varepsilon_0 \tag{6.2}$$

for some positive λ and some companion constant K.

We start with the simplest case $\hat{\theta}_n - \theta_0 = \sigma_0 S_n/n$, with partial sums of Z_i's that are i.i.d. with mean zero and variance 1, as in the lemma of Section 2.

LEMMA. *Suppose* $E|Z_i|^{2+\lambda} < \infty$ *for some* $\lambda \geq 0$. *Then there is a constant* $c_{2+\lambda}$ *such that*

$$E|S_n|^{2+\lambda} \leq c_{2+\lambda} n^{1+\lambda/2} E|N(0,1)|^{2+\lambda} \quad \textit{for all } n \tag{6.3}$$

(*and* $c_{2+\lambda}$ *can be replaced by* 1.001 *if we change "for all* n*" to "for all large* n*"*). *Furthermore,*

$$\Pr\left\{\sqrt{m} \sup_{n \geq m} \left|\frac{S_n}{n}\right| \geq a\right\} \leq \frac{6.75 c_{2+\lambda} E|N(0,1)|^{2+\lambda}}{a^{2+\lambda}} \quad \textit{for all } m \textit{ and } a. \tag{6.4}$$

PROOF. Of course $S_n/\sqrt{n} \to_d N(0,1)$. Results from von Bahr (1965) can be used to show $E|S_n/\sqrt{n}|^{2+\lambda} = E|N(0,1)|^{2+\lambda} + r_n$, where $|r_n| \leq M/\sqrt{n}$ for some M. In particular there is convergence and (6.3) (with accompanying parenthetical remark) follows from this. As a step in the rest of the proof we utilise a generalisation of Kolmogorov's inequality, namely

$$\Pr\Big\{\max_{i \leq n} |S_i| \geq a\Big\} \leq E|S_n|^{2+\lambda}/a^{2+\lambda},$$

which can be found in Loéve [(1960), page 263], for example. Let $q > 1$, suppose $q^k \leq m < q^{k+1}$, and let us abbreviate $c_{2+\lambda} E|N(0,1)|^{2+\lambda}$ with K. Then

$$\begin{aligned}
\Pr\left\{\sqrt{m} \sup_{n \geq m} \left|\frac{S_n}{n}\right| \geq a\right\} &\leq \sum_{i=k}^{\infty} \Pr\left\{\max_{q^i \leq n < q^{i+1}} |S_n| \geq \frac{aq^i}{\sqrt{m}}\right\} \\
&\leq \sum_{i=k}^{\infty} \frac{K(q^{i+1})^{1+\lambda/s}}{\left(aq^i/\sqrt{m}\right)^{2+\lambda}} \\
&= \frac{K}{a^{2+\lambda}} m^{1+\lambda/2} q^{1+\lambda/2} \sum_{i=k}^{\infty} \left(\frac{1}{q}\right)^{i(1+\lambda/2)} \\
&\leq \frac{K}{a^{2+\lambda}} \frac{q^{3+3\lambda/2}}{q^{1+\lambda/2} - 1}.
\end{aligned}$$

The best value of q corresponds to $q^{1+\lambda/2} = 3/2$ and the result follows. □

Note that the right-hand side of (6.4) becomes $6.75/a^2$ for $\lambda = 0$; this was needed in the proof of the lemma of Section 2. Robbins, Siegmund and Wendel (1968) have inequality (6.4) for this simplest $\lambda = 0$ case (but with constant 8 instead of 6.75).

This basic lemma can now be used to prove $E\varepsilon^2 N_\varepsilon \to 1.832\sigma_0^2$ in various situations. Consider smooth functions of averages. Suppose $\hat{\theta}_n = h(\overline{B}_n)$ estimates $\theta_0 = h(b)$, where $\overline{B}_n$ is the average of i.i.d. variables B_i with mean b and variance τ^2 as in Section 3A. In particular, $\varepsilon^2 N_\varepsilon \to_d \sigma_0^2 W_{\max}^2$, where $\sigma_0^2 = h'(b)^2\tau^2$, if only h has a continuous derivative around b.

THEOREM. *Suppose in addition that* $E|B_i|^{2+\lambda}$ *is finite for some positive* λ. *Then* $\varepsilon^2 EN_\varepsilon \to 2G\sigma_0^2$, *where* $G = 0.915966\ldots$ *is the Catalanian constant* (*see Section* 8A).

PROOF. We are to prove (6.2). This is very immediate if h' is bounded, but some care is needed to cover all the interesting cases where h' is unbounded; see Example 1 of Section 3A. With notation as in Section 3A we have

$$\begin{aligned}
&\Pr\Big\{\sqrt{m}\sup_{n\ge m}\big|\hat{\theta}_n - \theta_0\big| \ge 2a\Big\}\\
&\quad\le \Pr\Big\{\sqrt{m}\sup_{n\ge m}\big|h'(b)\big(\overline{B}_n - b\big)\big| \ge a\Big\}\\
&\qquad + \Pr\Big\{\sqrt{m}\sup_{n\ge m}\big|\big(h'(\tilde{b}_n) - h'(b)\big)\big(\overline{B}_n - b\big)\big| \ge a\Big\}\\
&\quad\le \frac{K'|h'(b)|^{2+\lambda}\tau^{2+\lambda}}{a^{2+\lambda}} + \Pr\Big\{\sqrt{m}\sup_{n\ge m}\rho\big(\big|\overline{B}_n - b\big|\big)\big|\overline{B}_n - b\big| \ge a\Big\},
\end{aligned}$$

where K' is a new constant and writing $\rho(r)$ for the maximum of $|h'(x) - h'(b)|$ as $|x - b| \le r$. Let ε_0 be such that $\rho(r) \le 1$ when $r \le \varepsilon_0$ [we even have $\rho(r) \to 0$ as $r \to 0$] and let $g(r) = \rho(r)r$, a continuously increasing function. The second term above is bounded by

$$\Pr\Big\{\sup_{n\ge m}\big|\overline{B}_n - b\big| \ge g^{-1}\Big(\frac{a}{\sqrt{m}}\Big)\Big\} \le \frac{K'\tau^{2+\lambda}}{\big\{\sqrt{m}\,g^{-1}\big(a/\sqrt{m}\big)\big\}^{2+\lambda}},$$

which again is bounded by $K'\tau^{2+\lambda}/a^{2+\lambda}$, provided $\sqrt{m}\,g^{-1}(a/\sqrt{m}) \ge a$, or $a/\sqrt{m} \ge g(a/\sqrt{m})$, or $1 \ge \rho(a/\sqrt{m})$. But this holds when $a/\sqrt{m} \le \varepsilon_0$, which proves (6.2). □

This result extends without serious difficulties to p-dimensional $\hat{\theta}_n$ being a smooth function of r averages. With notation as in Section 3A the proviso for correct convergence of $E\varepsilon^2 N_\varepsilon$ is finiteness of $E|B_{i,j}|^{2+\lambda}$ for some positive λ for $j = 1,\ldots,r$. One may also look for conditions in the maximum likelihood estimator case. The essential requirement is $E|\partial \log f(X_i, \theta_0)/\partial\theta_j|^{2+\lambda} < \infty$ for $j = 1,\ldots,p$.

7. The number of ε-misses. We have been able to reach rather general and elegant results for N_ε by the stochastic process approach, working with $\sqrt{m}\,(\hat{\theta}_{[mt]} - \theta_0)$ and its limiting process $\sigma_0 W(t)/t$. This approach can also successfully be applied to other random nonobservable quantities of interest, thereby broadening the perspective.

7A. *The one-dimensional case.* To illustrate this point, consider $Q_\varepsilon(a)$, the number of times among $n \geq a/\varepsilon^2$ where $|\hat{\theta}_n - \theta_0| \geq \varepsilon$. Then

$$(7.1) \qquad \varepsilon^2 Q_\varepsilon(a) \to_d \sigma_0^2 Q\big(a/\sigma_0^2\big) = \sigma_0^2 \mu\big\{t \geq a/\sigma_0^2\colon |W(t)/t| \geq 1\big\},$$

in which μ is Lebesgue measure on the halfline. This can be proved as follows, under conditions (2.2)–(2.3). Write $Q_\varepsilon(a)$ cleverly as $\int_{\langle a/\varepsilon^2\rangle}^{\infty} I\{|\hat{\theta}_{[s]} - \theta_0| \geq \varepsilon\}\,ds$, then let $m = 1/\varepsilon^2$ and change to $t = s/m$. After tending to details similar to those of Section 2, the result is

$$\varepsilon^2 Q_\varepsilon(a) = \int_{\langle ma\rangle/m}^{\infty} I\Big\{\sqrt{m}\,\big|\hat{\theta}_{[mt]} - \theta_0\big| \geq 1\Big\}\,dt \to_d \int_a^{\infty} I\{\sigma_0 |W(t)/t| \geq 1\}\,dt,$$

and the limit can be rewritten as $\sigma_0^2 Q(a/\sigma_0^2)$ above. There is also simultaneous convergence in distribution of $(\varepsilon^2 N_\varepsilon, \varepsilon^2 Q_\varepsilon(a))$ to $\sigma_0^2(\sup_{t\geq 1}|W(t)/t|^2, Q(a/\sigma_0^2))$. This follows by measurability and a.s. continuity of the appropriate functionals on $D[a,b]$-spaces and an extra argument to take care of the tail. It can also be proved via the continuous mapping theorem on the function space on $[0,\infty)$ described in Remark (i) of Section 2. These results can also be proved for $a = 0$. We leave the details out, but regularity conditions (2.2)–(2.3) suffice once more. In particular, ε^2 times the total number of ε-misses goes to $\sigma_0^2 Q(0) = \sigma_0^2 \mu\{t \geq 0\colon\ |W(t)/t| \geq 1\}$. Note that $EQ(b) = E(\chi_1^2 - b)I\{\chi_1^2 \geq b\}$, using Fubini's theorem. In particular, $EQ(0) = 1$ and $EQ(0.95) = 1/2$, which means that the estimator sequence has about $1/(2\varepsilon^2)$ misses of size $\sigma_0\varepsilon$ for $n \leq 0.95/\varepsilon^2$ and about $1/(2\varepsilon^2)$ misses of size $\sigma_0\varepsilon$ for $n \geq 0.95/\varepsilon^2$. We mention finally that Kao (1978) has the $\varepsilon^2 Q_\varepsilon(a)$ result, but only for the special case of simple i.i.d. averages and $a = 0$.

7B. *The multidimensional case.* One result is the following, under conditions (2.6)–(2.7): Let $Q_\varepsilon(a)$ be the number of times, among $n \geq a/\varepsilon^2$, where $(\hat{\theta}_n - \theta_0)'\Sigma_0^{-1}(\hat{\theta}_n - \theta_0) \geq \varepsilon^2$, with notation and conditions as in the theorem of Section 2. Then $\varepsilon^2 Q_\varepsilon(a)$ tends to $Q(a) = \mu\{t \geq a\colon \sum_{i=1}^{p} W_i(t)^2 > t^2\}$. Note that $Q(a)$ has mean value $E(\chi_p^2 - a)I\{\chi_p^2 \geq a\}$, which is easy to compute. In particular, the total number of ε-misses for the estimator sequence (with the Mahalanobis distance) is about p/ε^2.

Another result with two interesting consequences is as follows: Let distance function and conditions be as in the theorem of Section 2 and let Q_ε be the total number of $\|\hat{\theta}_n - \theta_0\| \geq \varepsilon$ cases. Then $\varepsilon^2 Q_\varepsilon$ tends to $Q = \mu\{t \geq 0\colon \|\Sigma_0^{1/2} W(t)/t\| \geq 1\}$. Our first point is yet another asymptotic optimality property for the maximum likelihood sequence: In the limit, as $\varepsilon \to 0$, provided the underlying parametric model is correct, *no other estimator sequence has stochastically fewer ε-misses*. Our second point is that EQ can be computed and leads to an a.r.e. measure in the multidimensional case; see the discussion that led to (2.5) and the end remark of Section 2. Taking $\|x - y\| = \{(x - y)'A(x - y)\}^{1/2}$, the mean of Q is $\int_0^\infty \Pr\{W(t)'\Sigma_0^{1/2}A\Sigma_0^{1/2}W(t)/t \geq t\}\,dt$, which becomes $EZ'\Sigma_0^{1/2}A\Sigma_0^{1/2}Z$, where $Z \sim N_p(0, I_p)$. Hence $EQ = \mathrm{Tr}(A\Sigma_0)$. One can also prove convergence of $\varepsilon^2 EQ_\varepsilon$ to EQ under conditions that are in fact simpler than those of Section 6. Suppose $\sqrt{n}\,(\hat{\theta}_{n,j} - \theta_0) \to_d N(0, \Sigma_j)$ for $j = 1, 2$, and let $Q_{\varepsilon,j}$ be the number of ε-misses for method j. Then the arguments presented before (2.5) suggest

$$\text{(7.2)} \qquad \text{a.r.e.} = \lim_{\varepsilon \to 0} \frac{EQ_{\varepsilon,1}}{EQ_{\varepsilon,2}} = \frac{\mathrm{Tr}(A\Sigma_1)}{\mathrm{Tr}(A\Sigma_2)}.$$

Under ordinary Euclidean distance, a.r.e. becomes $\mathrm{Tr}(\Sigma_1)/\mathrm{Tr}(\Sigma_2)$; see also 8E.

7C. *The number of ε-misses for Glivenko–Cantelli.* Consider the more complicated situation of Section 4. Let Q_ε be the number of times $\|F_n - F\| \geq \varepsilon$. Combining arguments above with those of Section 4, one can show that $\varepsilon^2 Q_\varepsilon \to_d Q = \mu\{s\colon A(s) \geq 1\}$, in which $A(s) = \max_{0 \leq t \leq 1} |K_0(s, t)/s|$. But, for fixed s, $K_0(s, \cdot\,)/s$ is distributed like $W^0(\cdot)/\sqrt{s}$, where $W^0(\cdot)$ is a Brownian bridge, so that $A(s) =_d \|W^0\|/\sqrt{s}$, where $\|W^0\|$ is the maximum of $|W^0(t)|$. This leads to $EQ = \int_0^\infty \Pr\{\|W^0\| \geq \sqrt{s}\}\,ds = E\|W^0\|^2 = \pi^2/12$. Accordingly, the full estimator sequence will have about $0.822/\varepsilon^2$ cases of $\|F_n - F\| \geq \varepsilon$. Similarly, ε^2 times the total number of cases of $\int (F_n - F)^2\,dF \geq \varepsilon^2$ will converge in distribution to a variable with expected value $1/6$. And for a final example of a nontrivial result reached using these methods, let Q_ε^* denote the number of $\int |F_n - F|\,dF \geq \varepsilon$ cases. Then $\varepsilon^2 Q_\varepsilon^*$ tends to an appropriate Q^* and $\varepsilon^2 EQ_\varepsilon^*$ tends to EQ^*, which can be proved to be equal to $E\{\int_0^1 |W^0(t)|\,dt\}^2$, and which is found to be $7/60$ by quite strenuous calculations.

7D. *The number of ε-misses for a density estimator.* Let finally Q_ε be the total number of times $|f_n(x) - f(x)| \geq \varepsilon$ in the density estimation problem considered in Section 5. Analysis similar to that above leads to $\varepsilon^{5/2} Q_\varepsilon \to_d Q = \mu\{s\colon |Z(s)| \geq 1\}$, where $Z(\cdot)$ is the process defined in Section 5. One can then show that

$$\text{(7.3)} \qquad EQ = E\left|c^2 f''(x)/2 + c^{-1/2} f(x)^{1/2} \beta_K^{1/2} N(0,1)\right|^{5/2}.$$

The value of c that gives the smallest expected number of ε-misses in the limit as $\varepsilon \to 0$ can be shown to be $1.008c_0(x)$, as in Section 5. Similar but more cumbersome calculations can be carried out for Q_ε', the number of times $|f_n'(x) - f'(x)| \geq \varepsilon$, under the optimal scheme $h_n = cn^{-1/7}$. One finds that $\varepsilon^{7/2} Q_\varepsilon'$ tends to a certain Q'. The best value of c from the point of view of approximate mean squared error is $c_0(x) = \{3\gamma_K f(x)/f'''(x)^2\}^{1/7}$, where $\gamma_K = \int K'(u)^2\,du$. But the value of c that minimises EQ' can by determined efforts be shown to be $1.049c_0(x)$.

8. Complementary remarks and results.

8A. *Numerical information.* Central in our limit distribution results is the variable $W_{\max} = \max_{0 \le s \le 1} |W(s)|$. Its distribution can be found in Shorack and Wellner [(1986), page 35], for example. One can prove that $EW_{\max} = \sqrt{\pi/2} = 1.2533$; $EW^2_{\max} = 2G = 1.8319$, featuring Catalan's constant; $\operatorname{Var} W_{\max} = 2G - \pi/2 = 0.5110^2$; $\operatorname{stdev}(W^2_{\max}) = (EW^4_{\max} - 4G^2)^{1/2} = 1.6055$. In the case of a single parameter, therefore, the following holds, in the notation of Section 6: $\varepsilon E N_\varepsilon^{1/2} \to \sqrt{\pi/2}\,\sigma_0$, if only $E|Z_i|^2 < \infty$; $\varepsilon^2 E N_\varepsilon \to 2G\sigma_0^2$, if $E|Z_i|^{2+\lambda}$ is finite for some positive λ; $\varepsilon^2 \operatorname{stdev}(N_\varepsilon) \to 1.6055\sigma_0^2$, if $E|Z_i|^{4+\lambda}$ is finite. The distribution of N_ε is skewed to the right, as $\operatorname{skew}(N_\varepsilon) = E\{(N_\varepsilon - EN_\varepsilon)/\operatorname{stdev}(N_\varepsilon)\}^3 \to 2.3308$ if $E|Z_i|^{6+\lambda}$ is finite. In the case of several parameters and the Mahalanobis distance, we have proved $\varepsilon^2 N_\varepsilon \to_d \chi^2_{p,\max}$, the maximum of $\chi^2_p(s) = \sum_{i=1}^{p} W_i(s)^2$ over $[0, 1]$. A way of computing its distribution is provided by DeLong (1980), along with a few quantiles. [More details and a fuller table of quantiles, arrived at by simulation of Brownian motions, are given in Hjort and Fenstad (1990).]

8B. *Extension to non-i.i.d. situations.* Our basic results read $\varepsilon^2 N_\varepsilon \to_d \sigma_0^2 W^2_{\max}$ and $\varepsilon^2 Q_\varepsilon(0) \to \sigma_0^2 Q(0)$ (in the one-parameter case), where σ_0^2 is the variance of the limit distribution for $\sqrt{m}\,(\hat{\theta}_m - \theta_0)$. These continue to hold for large classes non-i.i.d. situations. The key ingredients are process convergence $\sqrt{m}\,(\hat{\theta}_{[mt]} - \theta_0) \to_d \sigma_0 W(t)/t$ in $D[b, c]$-spaces (tightness and convergence of finite-dimensional distributions) and a tail inequality for $[c, \infty)$. Proving this for a particular case requires attention to technical details depending upon that case, however. In the technical report version of this article, such attention is given to linear regression and to a situation with auto-correlation.

8C. *A slow minimax estimator.* Let $X_1, X_2, \dots$ be independent Bernoulli trials with success probability p. The maximum likelihood estimator for p after n trials is $\hat{p}_n = Y_n/n$, where Y_n is the number of successes in the first n trials. From earlier results, we known that $\varepsilon^2 N_\varepsilon \to_d p(1-p)W^2_{\max}$, where N_ε is the last time $|\hat{p}_n - p| \ge \varepsilon$. Now consider the minimax estimator $p_n^* = (\sqrt{n}\,\hat{p}_n + 1/2)/(\sqrt{n} + 1)$ and the accompanying N'_ε, the last time $|p_n^* - p| \ge \varepsilon$. Some analysis reveals that

$$\sqrt{m}\,(p^*_{[mt]} - p) \to_d \sqrt{p(1-p)}\left[\frac{W(t)}{t} + \frac{(1/2) - p}{\sqrt{p(1-p)}}\,\frac{1}{\sqrt{t}}\right] \quad \text{in } D[1, c].$$

This can be used to prove $\varepsilon^2 N_\varepsilon^* \to_d p(1-p)\max_{0 \le s \le 1}|W(s) + b(p)\sqrt{s}\,|^2$, where $b(p) = (1/2 - p)/\{p(1-p)\}^{1/2}$. Accordingly, N_ε^* for p_n^* is stochastically larger in the limit that N_ε for $\hat{p}_n$ (unless $p = 1/2$). There is a similar story for Q_ε and Q_ε^*, the number of times $\hat{p}_n$ and p_n^* miss with more than ε. One can prove that $\varepsilon^2 Q_\varepsilon \to_d Q$ and $\varepsilon^2 Q_\varepsilon^* \to_d Q^*$, where $EQ = p(1-p)$ and $EQ^* = p(1-p) + (1/2 - p)^2 = 1/4$.

There are analogous results for the cdf estimator $F_n^* = (\sqrt{n}F_n + 1/2)/(\sqrt{n} + 1)$, which can be shown to be minimax when the loss function is $\int(\hat{F} - F)^2\,dw$, w any given weight function with mass 1. Then F_n^* can expect $1/(4\varepsilon^2)$ instances with loss greater than or equal to ε^2, regardless of the underlying F, whereas the nonminimax estimator F_n can expect $\int F(1 - F)\,dw/\varepsilon^2$ such instances.

8D. *Other distances.* Our basic result (2.8) was phrased in terms of a distance function $\|\hat{\theta}_n - \theta_0\|$. The arguments carry through also for other measures of distance that are not of the norm type. As a particular example of some interest, let $d[\theta_0 : \theta]$ be the Kullback–Leibler distance $\int f(x, \theta_0) \log\{f(x, \theta_0)/f(x, \theta)\}\,dx$ in some model with a p-dimensional parameter. Let $\hat{\theta}_n$ be the maximum likelihood estimator and let M_ε be the last n at which $d[\theta_0 : \hat{\theta}_n] \geq \varepsilon$. Then $2\varepsilon M_\varepsilon \to_d \chi^2_{p,\max}$ of (2.10) can be proved under mild conditions. Note that the limit is the same regardless of the actual parametric family. This holds when $f(x, \theta_0)$ represents the true model. A more general result, valid under the agnostic viewpoint presented in Section 3B, is given in Hjort and Fenstad (1990), along with further examples with other distance functions.

8E. *Second order results.* Our a.r.e. measures in (2.5) and (7.2) do not distinguish between estimators with the same limiting distribution. To do so requires second order asymptotics for $\varepsilon^2 N_\varepsilon$ and $\varepsilon^2 Q_\varepsilon$. In Hjort and Fenstad (1991) and Hjort and Khasminskii (1991), the limiting behaviour of differences between Q_ε's has been sorted out in cases where their ratio tends to 1, thereby making it possible to find second order optimal estimator sequences in many cases of interest. Thus, in the binomial situation, the $(Y_n + 2/3)/(n + 4/3)$ sequence can be expected to make 2.667 fewer ε-errors than the traditional Y_n/n sequence, for example, regardless of the underlying p parameter. And among all estimators of the form $\Sigma_{i=1}^n(X_i - \bar{X}_n)^2/(n + c)$ for a normal variance the one with denominator $n - 1/3$ can be expected to make the fewest ε-errors.

8F. *Sequential fixed-volume confidence regions.* Suppose (2.5) and (2.6) hold, and write N_ε^* for the last time $(\hat{\theta}_n - \theta_0)'\Sigma_n^{-1}(\hat{\theta}_n - \theta_0) \geq \varepsilon^2$. Then $\varepsilon^2 N_\varepsilon^*$ tends to $\chi^2_{p,\max}$ of (2.10), provided merely that $\hat{\Sigma}_n \to \Sigma_0$ a.s. (convergence in probability does not suffice). Let ε be small and given, find c such that $\Pr\{\chi^2_{p,\max} \leq c\} = 0.95$, put $m = [c/\varepsilon^2]$, and consider $I_n^* = \{\theta\colon (\theta - \hat{\theta})'\hat{\Sigma}_n^{-1}(\theta - \hat{\theta}_n) \leq \varepsilon^2\}$. Then $\Pr\{\theta_0 \in I_n^* \text{ for all } n \geq m\} \doteq 0.95$. The details of this construction are in Hjort and Fenstad (1990).

8G. *Shrinking boundaries and tests with power* 1. Methods of this paper can be used to construct sequential confidence regions with shrinking volume as well as sequential tests with power 1; see Hjort and Fenstad (1990).

8H. *The probabilities of $N_{\varepsilon,1} < N_{\varepsilon,2}$ and $Q_{\varepsilon,1} < Q_{\varepsilon,2}$.* Consider two competing estimator sequences with accompanying last ε-miss variables $N_{\varepsilon,j}$ and number of ε-misses variables $Q_{\varepsilon,j}$ as in (2.5) and (7.2). The probabilities $\Pr\{N_{\epsilon,1} < N_{\varepsilon,2}\}$ and $\Pr\{Q_{\epsilon,1} < Q_{\varepsilon,2}\}$ will usually converge as ε goes to 0; in fact $\varepsilon^2(N_{\varepsilon,1}, N_{\varepsilon,2}, Q_{\varepsilon,1}, Q_{\varepsilon,2})$ has a joint limiting distribution in terms of two correlated Brownian motions under natural conditions. These limits are found in Hjort and Fenstad (1990). As an example, consider the average estimator $\hat{\theta}_{n,1} = \bar{X}_n$ and the median estimator $\hat{\theta}_{n,2} = M_n$ for the mean parameter in the normal model. Then $N_{\varepsilon,1} < N_{\varepsilon,2}$ with probability about 0.72 and $Q_{\varepsilon,1} < Q_{\varepsilon,2}$ with probability about 0.69. To give $\bar{X}_n$ a harder match, replace the second estimator with the solution of $\sum_{i=1}^{n} \arctan(X_i - \theta) = 0$, an M-estimator with a smooth and bounded influence function. Then the figures become, respectively, 0.56 and 0.55. These are simulation-based figures computed using the exact limit distributions.

Acknowledgment. We are grateful to a reviewer and to David Siegmund for comments that inspired improvements and for pointing out some of the probabilistic literature which we had missed in an earlier version.

REFERENCES

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

von Bahr, B. (1965). On convergence of moments in the central limit theorem. *Ann. Math. Statist.* **36** 808–818.

Beran, R. (1977). Estimating a distribution function. *Ann. Statist.* **5** 400–404.

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. [Corrigenda (1975) **3** 1370.]

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

DeLong, D. M. (1980). Some asymptotic properties of a progressively censored nonparametric test for multiple regression. *J. Multivariate Anal.* **10** 363–370.

Hjort, N. L. (1986). *Statistical Symbol Recognition*. Research monograph, Norwegian Computing Cent., Oslo.

Hjort, N. L. and Fenstad, G. (1990). On the last time a strongly consistent estimator is more than ε from its target value. Statistical research report, Dept. Mathematics, Univ. Oslo.

Hjort, N. L. and Fenstad, G. (1991). Some second order asymptotics for the number of times an estimator is more than ε from its target point. Statistical research report, Dept. Mathematics, Univ. Oslo.

Hjort, N. L. and Khasminskii, R. Z. (1991). On the time a diffusion process spends along a line. Research report, Mathematical Sciences Research Institute, Berkeley.

Kao, C-S. (1978). On the time and the excess of linear boundary crossings of sample sums. *Ann. Statist.* **6** 191–199.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Loève, M. (1960). *Probability Theory*, 2nd ed. Van Nostrand, Toronto.

Müller, D. W. (1968). Verteilungs-Invarianzprinzipien für das starke Gesetz der großen Zahl. *Z. Wahrsch. Verw. Gebiete* **10** 173–192.

Müller, D. W. (1970). On Glivenko–Cantelli convergence. *Z. Wahrsch. Verw. Gebiete* **16** 195–210.

MÜLLER, D. W. (1972). Randomness and extrapolation. *Proc. Sixth Berkeley Symp. Math. Statist Probab.* **2** 1–31. Univ. California Press, Berkeley.

ROBBINS, H., SIEGMUND, D. and WENDEL, J. (1968). The limiting distribution of the last time $s_n \geq n\varepsilon$. *Proc. Nat. Acad. Sci. U.S.A.* **61** 1228–1230.

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

SHAO, J. (1989). Functional calculus and asymptotic theory for statistical analysis. *Statist. Probab. Lett.* **8** 397–405.

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

STUTE, W. (1983). Last passage time of *M*-estimators. *Scand. J. Statist.* **10** 301–305.

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF OSLO
P.B. 1053 BLINDERN
N-0316 OSLO 3
NORWAY