The Annals of Statistics
1987, Vol. 15, No. 3, 1131–1154

STOCHASTIC ESTIMATION AND TESTING[1]

BY R. BERAN AND P. W. MILLAR

University of California, Berkeley

Stochastic procedures are randomized tests, estimates and confidence sets with two properties:

(i) They are functions of an original sample and one or more artificially constructed auxiliary samples.
(ii) They become nearly nonrandomized when the auxiliary samples increase in size.

The stochastic procedures of this paper, which arise as approximations to numerically intractable procedures, involve iterated bootstrap techniques and random sampling schemes over abstract populations. A general methodology is applied to the asymptotic study of stochastic minimum distance tests, stochastic maximum likelihood estimates, stochastic confidence bands and several other stochastic procedures.

1. Introduction. Stochastic procedures are randomized tests, estimates and confidence sets with two properties:

(i) They are functions of an original sample and one or more artificially constructed auxiliary samples.
(ii) They become nearly nonrandomized when the auxiliary samples are increased in size.

A simple example is a bootstrap confidence set for a parameter, where the boundary of the confidence set is determined by Monte Carlo simulations of an appropriate bootstrap distribution. The stochastic procedures discussed in this paper are more complex, involving, for example, iterated bootstrap techniques and randomized sampling schemes over abstract populations. However, like the bootstrap Monte Carlo just mentioned, they are randomized procedures, in the sense of decision theory, which arise as useful approximations to numerically intractable procedures. This paper introduces several new stochastic procedures (described later on) and develops some asymptotic theory for these which recognizes the effects of the auxiliary randomization.

1.1. *Stochastic minimum distance estimates.* Minimum distance estimates are attractive because of their robustness and good rate-of-convergence in a relatively wide range of models. Let $\{Q_\theta,\ \theta \in \Theta\}$ be a family of cdf's on R^q, where Θ is an open subset of R^d. Let $\mathbf{x}_n = (x_1, \dots, x_n)$ be a sample of size n

Received June 1985; revised October 1986.
[1]Research supported by National Science Foundation Grant MCS-84-03239.
AMS 1980 *subject classification*. Primary 62E20.
Key words and phrases. Confidence set, critical value, bootstrap, Monte Carlo, minimum distance estimate, goodness-of-fit test, likelihood ratio test, maximum likelihood estimate, stochastic search, confidence band for a measure, empirical measure.

from an unknown Q_θ, let $\hat{Q}_n$ be the empirical cdf of $\mathbf{x}_n$ and let $|\cdot|$ denote some norm on bounded real functions of R^q, such as the supremum norm. A minimum distance estimate is any Θ-valued random variable that comes within n^{-1} of minimizing the discrepancy function $D_n(\theta, \mathbf{x}_n) = |\hat{Q}_n - Q_\theta|$. If the discrepancy function has several relative minima, or if the dimension d of Θ is large, calculation of a minimum distance estimate can be difficult.

A result from real variables provides another view of the problem. Let μ be a probability measure on Θ and suppose $s_1, s_2, \ldots$ are i.i.d. random vectors, each distributed according to μ. Then

$$\inf\{D_n(s_j, \mathbf{x}_n)\colon 1 \le j < \infty\} = \operatorname*{ess\,inf}_{\mu} D_n(\theta, \mathbf{x}_n) \quad \text{w.p.1}, \tag{1.1}$$

the ess inf notation meaning essential infimum with respect to μ. Moreover, if $D_n(\theta, \mathbf{x}_n)$ is continuous in θ and if μ gives positive probability to every open subset of Θ, then $\operatorname{ess\,inf}_\mu D_n(\theta, \mathbf{x}_n)$ coincides with $\inf_\theta D_n(\theta, \mathbf{x}_n)$.

This viewpoint motivates a stochastic approximation to the minimum distance estimate. Let $\mathbf{s}_n = (s_1, \ldots s_{j_n})$ be a random sample of j_n elements in Θ, drawn as described previously or in some other manner (see the following). Define the stochastic minimum distance estimate $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ by the requirement that

$$D_n(\hat{\theta}_n, \mathbf{x}_n) = \min\{D_n(s_j, \mathbf{x}_n)\colon 1 \le j \le j_n\}, \tag{1.2}$$

that is, $\hat{\theta}_n$ minimizes $D_n(\theta, \mathbf{x}_n)$ as θ ranges over values in $\mathbf{s}_n$. The estimate $\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ is a stochastic procedure, in the sense described earlier, since it is a function of the original sample $\mathbf{x}_n$ and of the auxiliary random search sample $\mathbf{s}_n$.

The global search strategy takes $s_1, \ldots s_{j_n}$ to be i.i.d. μ, where μ is a fixed probability on Θ with full support. Under the regularity assumptions of Section 4, the resulting *global* stochastic minimum distance estimate $\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ has the same asymptotic distribution as the actual minimum distance estimate, provided $\lim_{n\to\infty} j_n n^{-d/2} = \infty$, where d is the dimension of Θ. The condition on j_n has an intuitive explanation: It ensures that the expected number of search points $\{s_j\colon 1 \le j \le j_n\}$ falling within a ball of radius $O(n^{-1/2})$ about the true parameter value tends to infinity as n increases. This is needed because the discrepancy function $D_n(\theta, \mathbf{x}_n)$ typically approaches its infimum within such a ball. When dimension d is large, global search wastes too many observations searching unimportant parts of the parameter space.

A more sophisticated local search strategy is often available. Suppose $\tilde{\theta}_n = \tilde{\theta}_n(\mathbf{x}_n)$ is a preliminary estimate which is $n^{1/2}$-consistent for the unknown parameter and takes its values in the parameter space Θ. Conditionally on $\mathbf{x}_n$, draw j_n independent bootstrap samples $\mathbf{x}_1^*, \ldots, \mathbf{x}_{j_n}^*$, each of size n, from the fitted model $Q_{\tilde{\theta}_n(\mathbf{x}_n)}$. Let $s_j = \tilde{\theta}_n(\mathbf{x}_j^*)$, the value of the preliminary estimate recalculated from the jth parametric bootstrap sample. Set $\mathbf{s}_n = (s_1, \ldots, s_{j_n})$ and determine the stochastic minimum distance estimate as in (1.1). Under the regularity conditions of Section 4, this *local* stochastic minimum distance estimate has the same asymptotic distribution as the actual minimum distance estimate, provided only $\lim_{n\to\infty} j_n = \infty$. This last condition does not involve the dimension d of Θ because the local search described previously automatically concentrates on balls of radius $O(n^{-1/2})$ about the true parameter value. Other ways of constructing local search samples with this key property are, of course, possible and lead to the same asymptotics. Note, however, that the bootstrap local search automatically generates points s_j in Θ, a useful property when the parameter space is constrained (e.g., when Θ is a set of positive definite covariance matrices).

1.2. *Stochastic minimum distance tests.* Consider the null hypothesis that the actual cdf of each observation in the i.i.d. sample $\mathbf{x}_n$ is Q_{θ_0}, where $\theta_0 \in \Theta$ is unknown. The minimum distance test for this hypothesis rejects when the test statistic $\inf_\theta D_n(\theta, \mathbf{x}_n)$ is sufficiently large. A stochastic approximation to this test is based upon the test statistic $D_n[\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n), \mathbf{x}_n] \doteq \min\{D_n(s_j, \mathbf{x}_n): 1 \le j \le j_n\}$, where $\mathbf{s}_n$ is the local search sample described in Section 1.1 and $\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ is the local stochastic minimum distance estimate. An asymptotically valid critical value for the stochastic test can be obtained as follows, by parametric bootstrapping. Given $(\mathbf{x}_n, \mathbf{s}_n)$, let $\mathbf{y}_1^*, \dots, \mathbf{y}_{k_n}^*$ be k_n bootstrap samples, each of size n, drawn from the fitted model $Q_{\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)}$. Let $\hat{r}_n$ be a $(1-\alpha)$th quantile of the empirical distribution of the values $\{D_n[\hat{\theta}_n(\mathbf{y}_k^*, \mathbf{s}_n), \mathbf{y}_k^*]: 1 \le k \le k_n\}$. The stochastic goodness-of-fit test which rejects whenever $D_n[\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n), \mathbf{x}_n] > \hat{r}_n$ has asymptotic level α (Section 4.2), provided $\lim_{n\to\infty} j_n = \lim_{n\to\infty} k_n = \infty$. Note that this test is a function of the original sample $\mathbf{x}_n$ and of the auxiliary samples $\mathbf{s}_n$ and $\mathbf{y}_1^*, \dots, \mathbf{y}_{k_n}^*$. The approach just described overcomes well-known difficulties in constructing minimum distance goodness-of-fit tests [cf. Durbin (1973) and Pollard (1980)].

1.3. *Stochastic maximum likelihood estimates.* Maximum likelihood estimation has a close formal similarity to minimum distance estimation: Instead of minimizing a discrepancy function $D_n(\theta, \mathbf{x}_n)$, we seek to maximize the likelihood function $L_n(\theta; \mathbf{x}_n) = \prod_{i=1}^n f(\theta; x_i)$, where $f(\theta; x)$ is the density of Q_θ. The stochastic maximum likelihood estimate $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ is defined by the requirement

$$(1.3) \qquad L_n(\hat{\theta}_n; \mathbf{x}_n) = \max\{L_n(s_j; \mathbf{x}_n): 1 \le j \le j_n\},$$

where $\mathbf{s}_n = (s_1, \dots, s_{j_n})$ is a random search sample. Under the regularity conditions of Section 3, the stochastic maximum likelihood estimate, like the true maximum likelihood estimate, is asymptotically efficient and asymptotically normal, provided $\lim_{n\to\infty} j_n = \infty$ (when $\mathbf{s}_n$ is a local search sample) or $\lim_{n\to\infty} j_n n^{-d/2} = \infty$ (when $\mathbf{s}_n$ is a global search sample).

Finite sample behavior of the local stochastic maximum likelihood estimate and of a familiar competitor, the method of scoring, were studied numerically in the Cauchy location model on the real line. The preliminary estimate $\tilde{\theta}_n$ was taken to be the sample median. The local stochastic maximum likelihood estimate $\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ was improved by appending $\tilde{\theta}_n$ to the bootstrap local search sample $\mathbf{s}_n$ described in Section 1.1; this change does not affect the asymptotics of

TABLE 1

Estimated efficiencies relative to the sample median of the j_n-step method of scoring estimate and of the local stochastic maximum likelihood estimate (SMLE) based on a search sample of size j_n. The data are a sample of size n from the Cauchy location model. The efficiency estimates are based on 1000 Cauchy samples. Asymptotic relative efficiency is $\pi^2/8 = 1.23$.

n \ j_n	5		11		21		41	
	j_n-step	SMLE	j_n-step	SMLE	j_n-step	SMLE	j_n-step	SMLE
1	1.05	1.21	1.18	1.20	1.23	1.05	1.20	1.00
2	1.06	1.27	1.23	1.15	1.27	1.20	1.22	1.03
5	1.03	1.28	1.24	1.21	1.29	1.25	1.23	1.12
10	1.03	1.29	1.23	1.27	1.28	1.31	1.22	1.14
25	1.01	1.29	1.19	1.30	1.29	1.30	1.24	1.21
50	1.02	1.20	1.19	1.28	1.27	1.31	1.25	1.22

this paper. The j_n-step method of scoring estimate in the Cauchy location model is $\bar{\theta}_{n,j_n}$, where

$$\begin{aligned} \bar{\theta}_{n,0} &= \bar{\theta}_n, \\ \bar{\theta}_{n,j} &= \bar{\theta}_{n,j-1} + n^{-1} \sum_{i=1}^{n} v\big(x_i - \bar{\theta}_{n,j-1}\big), \qquad 1 \le j \le j_n, \end{aligned} \tag{1.4}$$

and $v(x) = 4x/(1+x^2)$.

For various choices of sample size n and of j_n, Table 1 reports the observed efficiency relative to sample median, in 1000 Monte Carlo trials, of the j_n-step method of scoring estimate $\bar{\theta}_{n,j_n}$ and of the stochastic MLE based upon a local search sample of size j_n (with sample median appended). The asymptotic relative efficiency of the local stochastic MLE is $\pi^2/8 = 1.23$, provided $j_n \to \infty$ as $n \to \infty$. The asymptotic relative efficiency of the j_n-step method of scoring estimate is also $\pi^2/8$, if j_n is fixed as $n \to \infty$ [Le Cam (1974)]. The entries in Table 1 reflect the differing asymptotics clearly: For $n = 41$, the one-step method of scoring estimate compares in performance to the stochastic MLE having a search sample of size 25.

On the other hand, at $n = 5$, the local stochastic MLE dominates the method of scoring estimate, no matter how many iterations of the latter are computed. The ripples in the Cauchy likelihood when n is small confuse the method of scoring [Barnett (1966)] but do not affect stochastic search techniques.

1.4. *Confidence sets for an unknown probability.* Let $\hat{Q}_n$ be the empirical measure of n i.i.d. random variables with values in a space S and with common unknown distribution Q. Let $\mathbb{V}$ be a collection of subsets of S. Consider a confidence set $\hat{C}_n$ for the unknown Q of the form

$$\hat{C}_n = \{P: \sup|\hat{Q}_n(V) - P(V)| \le \hat{r}_n\},$$

where the supremum is over all sets V in $\mathbb{V}$. The classical Kolmogorov–Smirnov confidence sets are of this form. When S is Euclidean, other choices of $\mathbb{V}$ include the set of all halfspaces or the set of all ellipsoids. In general, construction of $\hat{C}_n$ founders on two difficulties: computation of the supremum and suitable choice of critical value $\hat{r}_n$. The asymptotic distribution theory is usually too intractable to give a value for $\hat{r}_n$.

To construct a stochastic approximation to $\hat{C}_n$, pick sets $V_1, \dots, V_{j_n}$ at random from $\mathbb{V}$ and replace the supremum over $\mathbb{V}$ by the supremum over the $\{V_j: 1 \le j \le j_n\}$. Estimating the conditional distribution of $n^{1/2}\max\{|\hat{Q}_n(V_j) - P(V_j)|: 1 \le j \le j_n\}$, given the $\{V_j: 1 \le j \le j_n\}$, by nonparametric bootstrapping yields a critical value $\tilde{r}_n$ such that the stochastic confidence set

$$\left\{P: n^{1/2} \max_{1 \le j \le j_n} |\hat{Q}_n(V_j) - P(V_j)| \le \tilde{r}_n\right\}$$

has the desired asymptotic coverage probability. Section 8 gives details. The special case when $\mathbb{V}$ consists of all halfspaces in an Euclidean space was discussed by Beran and Millar (1986). The treatment here differs from the earlier one in three respects: Here the random search extends over all sets in $\mathbb{V}$, $\mathbb{V}$ is not restricted to halfspaces and the effect of approximating the relevant bootstrap distribution by Monte Carlo methods is analyzed.

1.5. *Other stochastic procedures*. Several other stochastic procedures are analyzed in this paper. Section 6 treats stochastic likelihood ratio tests for composite hypothesis in parametric models. These tests replace the double maximization of the classical likelihood ratio tests by two local stochastic search maximizations and replace the chi-squared approximation to the null distribution by a bootstrap estimate. There are heuristic grounds for believing that the bootstrap approach reduces the level error of the test.

Section 3.3 discusses confidence sets for a parameter θ based on $n^{1/2}|\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n) - \theta|$, where $\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n)$ is the local stochastic maximum likelihood estimate. Critical values for the confidence set are obtained from a parametric bootstrap estimate of the conditional distribution of $n^{1/2}|\hat{\theta}_n(\mathbf{x}_n, \mathbf{s}_n) - \theta|$, given the local search sample $\mathbf{s}_n$. There is reason to expect that studentizing before bootstrapping may yield a stochastic confidence set with smaller coverage probability error. [See the discussion in Abramovitch and Singh (1985) for the case with $j_n = \infty$.] The second-order asymptotics for this problem need further development.

Section 7 studies a stochastic Bayes estimate in which integration with respect to a prior distribution is replaced by averaging over a sample drawn from that prior. The asymptotics for this stochastic procedure are sensitive to the dimension of the parameter space. This example illustrates that stochastic procedures are not necessarily linked to resampling methods or to a random search for an extremum.

We observe that stochastic procedures have a history in statistical practice, which begins with the use of Monte Carlo methods to approximate critical values for tests of a simple hypothesis [e.g., Dwass (1957)]. More recent stochastic procedures include tests and confidence sets based on Monte Carlo approximations to bootstrap distribution [e.g., Efron (1979) and Beran (1986)], approximations to maximum likelihood estimates based on Monte Carlo estimates of the log-likelihood function [Diggle and Gratton (1984)], the stochastic confidence sets for an unknown probability cited in Section 1.4 [Beran and Millar (1986)] and stochastic tests for a simple hypothesis which foreshadow the confidence sets of Section 1.4 [Pyke (1984)]. The present paper goes farther in two directions: It introduces and studies the method of local stochastic search and it develops an asymptotic analysis for stochastic procedures which recognizes the effects of the auxiliary randomization. It is ironic that stochastic procedures— the most practical of randomized procedures—should come to the fore so late in the story of decision theory.

2. Basic tools. The distribution of a stochastic procedure depends upon the joint distribution of the augmented sample, which consists of the original sample and the artificial auxiliary samples. Section 1 described how augmented samples are constructed in practice. A different mathematical viewpoint is needed for the asymptotic theory. From now on, we regard a stochastic procedure as a random variable (that is, a measurable function) defined on a space of elementary events which consists of all possible values of the augmented sample. The space is endowed with the probability measure which corresponds to the constructions of augmented samples in Section 1. This section introduces some notation and several weak convergence results which will be used repeatedly in the remainder of the paper.

2.1. *Empirical estimates of random probability measures.* For each $n \geq 1$, let $(\mathbb{X}_n, \mathbb{F}_n, P_n)$ be a probability space. Let $\mathbb{A}$ be a complete metric space with Borel σ-algebra Σ and, for each $j_n \geq 1$, let $\mathbb{A}^{j_n}$ denote the j_n-fold product of $\mathbb{A}$ with product σ-algebra Σ^{j_n}. Let $\mathbb{M}$ be the set of all probability measures on $(\mathbb{A}, \Sigma)$, metrized by the Prohorov metric. Let μ_n be a random probability measure on $\mathbb{X}_n/\mathbb{A}$, defined as follows: μ_n is a real-valued function on $\mathbb{X}_n \times \Sigma$ such that $\mu_n(\mathbf{x}_n, \cdot)$ is an element of $\mathbb{M}$ for each $\mathbf{x}_n \in \mathbb{X}_n$ and the mapping $\mathbf{x}_n \to \mu_n(\mathbf{x}_n, \cdot)$ is measurable. The notation $\mu_n(\mathbf{x}_n, D)$ then denotes the probability assigned to the set $D \in \Sigma$ by the probability measure $\mu_n(\mathbf{x}_n, \cdot)$.

Define the random product measure $\mu_n^{j_n}$ on $\mathbb{X}_n/\mathbb{A}^{j_n}$ by

$$(2.1) \qquad \mu_n^{j_n}(\mathbf{x}_n, B) = \int_B \prod_{j=1}^{j_n} \mu_n(\mathbf{x}_n, da_j),$$

where $\mathbf{x}_n \in \mathbb{X}_n$, $\mathbf{a}_n = (a_1, \ldots, a_{j_n}) \in \mathbb{A}^{j_n}$ and $B \in \Sigma^{j_n}$. Let $P_n \otimes \mu_n^{j_n}$ be the probability measure on $\mathbb{X}_n \times \mathbb{A}^{j_n}$ given by

$$(2.2) \qquad (P_n \otimes \mu_n^{j_n})(C) = \int_C \mu_n^{j_n}(\mathbf{x}_n, d\mathbf{a}_n) P_n(d\mathbf{x}_n),$$

where $\mathbf{x}_n \in \mathbb{X}_n$, $\mathbf{a}_n \in \mathbb{A}^{j_n}$ and C is any set in the product σ-algebra $\mathbb{F}_n \otimes \Sigma^{j_n}$. The probability models $(\mathbb{X}_n, \mathbb{F}_n, P_n)$ and $(\mathbb{X}_n \times \mathbb{A}^{j_n}, \mathbb{F}_n \otimes \Sigma^{j_n}, P_n \otimes \mu_n^{j_n})$ admit the following interpretation. The space $\mathbb{X}_n$ is the set of all possible values for the original sample and P_n is the distribution of the original sample. For each

$\mathbf{x}_n \in \mathbb{X}_n$, $\mu_n^{j_n}(\mathbf{x}_n, \cdot)$ is the distribution of an auxiliary sample of size j_n drawn from the probability measure $\mu_n(\mathbf{x}_n, \cdot)$. The space $\mathbb{A}^{j_n}$ is the set of all possible values for such an auxiliary sample. The space $\mathbb{X}_n \times \mathbb{A}^{j_n}$ is the set of all possible values for the augmented sample and $P_n \otimes \mu_n^{j_n}$ is the distribution of the augmented sample.

The empirical distribution of $\hat{\mu}_n$ of an auxiliary sample is defined formally as the random probability measure on $\mathbb{X}_n \times \mathbb{A}^{j_n}/\mathbb{A}$ given by

$$\hat{\mu}_n((\mathbf{x}_n, \mathbf{a}_n), D) = j_n^{-1} \sum_{j=1}^{j_n} I_D(a_j), \tag{2.3}$$

where $\mathbf{x}_n \in \mathbb{X}_n$, $\mathbf{a}_n = (a_1, \dots, a_{j_n}) \in \mathbb{A}^{j_n}$ and I_D is the indicator of the set $D \in \Sigma$. Let ρ denote Prohorov distance on $\mathbb{M}$, the set of all probability measures on $(\mathbb{A}, \Sigma)$.

THEOREM 2.1. *Let μ_0 be a nonrandom probability measure in $\mathbb{M}$. Let $\{\mu_n: n \geq 1\}$ be a sequence of random probability measures on $\mathbb{X}_n/\mathbb{A}$ such that $\rho(\mu_n, \mu_0) \to 0$ in P_n probability as $n \to \infty$. If $\lim_{n\to\infty} j_n = \infty$, then $\rho(\hat{\mu}_n, \mu_0) \to 0$ in $P_n \otimes \mu_n^{j_n}$ probability as $n \to \infty$.*

Note that j_n can approach infinity at any rate and that the sequence $\{P_n\}$ need not converge. In the special nonrandom measure case $\mu_n = \mu_0$, $n \geq 1$, the theorem is due to Varadarajan (1958). When $\mathbb{A}$ is Euclidean, which is the case needed in this paper, Theorem 2.1 follows from the exponential bound of Kiefer and Wolfowitz (1958) on the Kolmogorov distance between $\hat{\mu}_n$ and μ_n. The theorem as stated is proved in Beran, Le Cam and Millar (1987).

2.2. *Substitution arguments.* Whenever convenient, weak convergence results will be derived from ordinary pointwise convergence properties, in the manner of Skorohod. The two lemmas stated below—both consequences of a theorem due to Wichura (1970)—will be applied repeatedly in the remainder of the paper.

Let $\mathbb{Z}, \mathbb{Y}, \mathbb{W}$ be metric spaces endowed with their Borel σ-algebras. Let $\{\Lambda_n: n \geq 1\}$ be a sequence of $\mathbb{W}$-valued functionals defined on $\mathbb{Z} \times \mathbb{Y}$ and let Λ be a fixed $\mathbb{W}$-valued functional defined on $\mathbb{Z} \times \mathbb{Y}$. A first substitution lemma is obtained under the following hypotheses:

(2.4a) $\{(Z_n, Y_n): n \geq 1\}$ and (Z, Y) are $\mathbb{Z} \times \mathbb{Y}$-valued random variables; (Z_n, Y_n) converges weakly in $\mathbb{Z} \times \mathbb{Y}$ to (Z, Y), whose support is separable; the support of Y lies in $\mathbb{Y}_0 \subset \mathbb{Y}$.

(2.4b) $\Lambda_n(z_n, y_n) \to \Lambda(z, y)$ in $\mathbb{W}$ wherever $(z_n, y_n) \to (z, y) \in \mathbb{Z} \times \mathbb{Y}_0$.

LEMMA 2.1. *If* (2.4a) *and* (2.4b) *hold, then $\Lambda_n(Z_n, Y_n)$ converges weakly in $\mathbb{W}$ to $\Lambda(Z, Y)$.*

REMARKS. A special case of assumption (2.4a), to be used in the sequel, is

(2.4c) Z_n converges in probability to z_0, a fixed element of $\mathbb{Z}$; Y_n converges weakly in $\mathbb{Y}$ to Y, whose support lies in $\mathbb{Y}_0$ and is separable.

There are many variants of Lemma 2.1. Some of these arise because stronger conditions on the $\{z_n, y_n\}$ in hypothesis (2.4b) may be necessary in order to ensure that $\Lambda_n(z_n, y_n)$ converge. For example, let $\{\delta_n\}$ be a sequence of real numbers, $\delta_n \downarrow 0$. Suppose (2.4b) holds but only for sequences $\{z_n\}$ such that for some c, $\delta_n^{-1}d(z_n, z_0) \le c$; here d is the metric of $\mathbb{Z}$ and z_0 is a fixed element of $\mathbb{Z}$. If (2.4c) holds and if $\{\delta_n^{-1}d(Z_n, z_0)\}$ is tight, then the conclusion of Lemma 2.1 continues to hold. Another variant arises if Λ_n is real-valued and the first half of (2.4b) is replaced by (say) $\limsup_n \Lambda_n(z_n, y_n) \le \Lambda(z, y)$, then the conclusion of Lemma 2.1 can be replaced by $\limsup_n Ef(\Lambda_n(z_n, Y_n)) \le Ef(\Lambda(Z, Y))$, where f is any increasing function for which the expectations exist.

A second substitution lemma is obtained under the following hypotheses:

(2.5a) $\{(Z_n, Y_n): n \ge 1\}$ and (Z, Y) are $\mathbb{Z} \times \mathbb{Y}$-valued random variables; the $\{Z_n\}$ and the $\{Y_n\}$ are independent; Y_n converges weakly in $\mathbb{Y}$ to Y, whose support is separable and lies in $\mathbb{Y}_0 \subset \mathbb{Y}$.

(2.5b) $\Lambda_n(Z_n, y_n)$ converges in probability in $\mathbb{W}$ to $\Lambda(Z, y)$ whenever $y_n \to y \in \mathbb{Y}_0$.

Let $\mathbb{L}_n(z)$ denote the distribution of $\Lambda_n(z, Y_n)$ and $\mathbb{L}(z)$ denote the distribution of $\Lambda(z, Y)$.

LEMMA 2.2. *If* (2.5a) *and* (2.5b) *hold, then*

(i) $\Lambda_n(Z_n, Y_n)$ *converges weakly in* $\mathbb{W}$ *to* $\Lambda(Z, Y)$;

(ii) $\mathbb{L}_n(Z_n)$ *converges weakly to* $\mathbb{L}(Z)$, *as random elements of the space of probability measures on* $\mathbb{W}$ *metrized by Prohorov metric.*

PROOF. By applying Wichura's (1970) theorem to the $\{Y_n\}$ and Y, construct versions of $\{(Y_n, Z_n)\}$ and (Y, Z) such that $Y_n \to Y$ a.s. in addition to (2.5a) and (2.5b). Assume without loss of generality that the metric d on $\mathbb{W}$ is bounded above by 1. Then, by Fubini's theorem and dominated convergence,

$$\lim_{n\to\infty} Ed[\Lambda_n(Z_n, Y_n), \Lambda(Z, Y)] = 0, \tag{2.6}$$

which implies part (i) of the lemma.

Let β denote the bounded Lipschitz metric on the space of all probability measures on $\mathbb{W}$. For every choice of constants $\{z_n\}$, z in $\mathbb{Z}$,

$$\beta[\mathbb{L}_n(z_n), \mathbb{L}(z)] \le Ed[\Lambda_n(z_n, Y_n), \Lambda(z, Y)]. \tag{2.7}$$

Since $\{Y_n\}$, Y are independent of $\{Z_n\}$, Z,

$$E\beta[\mathbb{L}_n(Z_n), \mathbb{L}(Z)] \le Ed[\Lambda_n(Z_n, Y_n), \Lambda(Z, Y)]. \tag{2.8}$$

Part (ii) of the lemma follows from (2.8) and (2.6). □

REMARK. Several variants of Lemma 2.2 exist, in which hypothesis (2.5b) is weakened by imposing more restrictions upon the $\{y_n\}$ (cf. the comments following Lemma 2.1).

3. Stochastic MLE: Local search.

3.1. *Construction.* Let $\{Q_\theta, \theta \in \Theta\}$ be a family of mutually absolutely continuous probabilities on a space $\mathbb{X}$; for simplicity let Θ be R^d. Let $\mathbb{X}_n$ be the n-fold product of $\mathbb{X}$, and Q_θ^n the product measure of Q_θ. Let $f^n(\theta; \cdot)$, $\theta \in \Theta$ be the density of Q_θ^n. The maximum likelihood estimator of θ is a variable that, for each $\mathbf{x}_n$, maximizes $f^n(\theta; \mathbf{x}_n)$ as a function of θ. Let $\tilde{\theta}_n = \tilde{\theta}_n(\mathbf{x}_n)$ be a preliminary estimate of θ, i.e., $\tilde{\theta}_n$ is a map from $\mathbb{X}_n$ to Θ. The local stochastic maximum likelihood estimate is defined to be any measurable function $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n)$ on $\mathbb{X}_n \times (R^d)^{j_n}$ satisfying

$$f^n(\hat{\theta}_n; \mathbf{x}_n) = \max_{i \le j_n} f(\tilde{\theta}_n(\mathbf{x}_n) + n^{-1/2} t_i; \mathbf{x}_n), \tag{3.1}$$

where $\mathbf{x}_n \in \mathbb{X}_n$ and $\mathbf{t}_n = (t_1, \dots, t_{j_n}) \in (R^d)^{j_n}$.

Let $\mathbb{D}_n(\theta)$ be the distribution of $n^{1/2}(\tilde{\theta}_n - \theta)$ under Q_θ^n. Define the random probability measure μ_n by

$$\mu_n(\mathbf{x}_n, \cdot) = \mathbb{D}_n(\tilde{\theta}_n(\mathbf{x}_n)). \tag{3.2}$$

If θ_n is the true parameter at time n, put the measure $Q_{\theta_n}^n \otimes \mu_n^{j_n}$ on $\mathbb{X}_n \times (R^d)^{j_n}$. The local stochastic maximum likelihood estimate $\hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n)$ defined in (3.1) has the same distribution as the one defined in Section 1.3. Only the underlying probability space has been changed for the theoretical development.

3.2. *Asymptotic normality.* For $\theta, \theta_1 \in \Theta$ and $\mathbf{x}_n \in \mathbb{X}_n$ define the likelihood ratio

$$L_n(\theta_1, \theta; \mathbf{x}_n) = f^n(\theta_1, \mathbf{x}_n)/f^n(\theta, \mathbf{x}_n). \tag{3.3}$$

Often, $L_n(\theta_1, \theta)$ will be written in place of $L_n(\theta_1, \theta; \mathbf{x}_n)$. Asymptotic normality will be proved under the following hypotheses. Fix θ_0 in R^d. Let $C_0(R^d)$ denote the set of all continuous functions on R^d which vanish at infinity, metrized by the supremum norm.

(3.4a) For every $c > 0$ and every sequence $\{\theta_n\}$ with $n^{1/2}|\theta_n - \theta_0| \le c$, $\mathbb{D}_n(\theta_n)$ converges to a measure μ_0 that depends only on θ_0.

(3.4b) For every sequence $\{\theta_n\}$ as in (3.4a) the processes $\{L_n(\theta_n + n^{-1/2}u, \theta_n; \mathbf{x}_n): u \in \Theta\}$ converge in distribution under $Q_{\theta_n}^n$, as a random element of $C_0(R^d)$, to a process $W = \{W(u): u \in R^d\}$, where the distribution of $\log W(u)$ is that of $\langle u, N\rangle - 2^{-1}\langle u, I(\theta_0)u\rangle$. Here N is normal with mean 0 and covariance $I(\theta_0)$, $I(\theta_0)$ is the Fisher information matrix (assumed nonsingular) and the brackets denote the inner product in R^d.

Note that these two hypotheses can hold outside the i.i.d. setting we have chosen.

THEOREM 3.1. *Assume* (3.4*a*) *and* (3.4*b*) *and* $\{\theta_n\}$ *is a sequence satisfying* $|\theta_n - \theta_0| \le cn^{-1/2}$. *If* $j_n \to \infty$, *then*

$$n^{1/2}(\hat{\theta}_n - \theta_n) \Rightarrow N(0, I^{-1}(\theta_0)), \quad \text{under } Q_{\theta_n}^n \otimes \mu_n^{j_n}. \tag{3.5}$$

REMARKS. Hypothesis (3.4b) is satisfied for certain canonical exponential family models, such as the normal location-scale model, but is too strong otherwise. It is used here to keep the proof of Theorem 3.1 straightforward. A more general account of stochastic maximum likelihood estimates would draw on the more refined hypothesis of Le Cam (1970) and of Ibragimov and Has'minskii (1981), Chapter 2.

Hypothesis (3.4a) asserts that the preliminary estimate $\tilde{\theta}_n$ is regular in Hájek's sense [Hájek (1970)]. The theorem implies that the local stochastic maximum likelihood estimate is regular, efficient and locally asymptotically minimax in the usual framework [Ibragimov and Has'minskii (1981), Chapter 2, and Millar (1983), Chapter 7]. Indeed $\hat{\theta}_n$ is very close to the actual maximum likelihood estimate, as shown by the corollary to the proof given later. An important problem is to compare the performance of stochastic MLE's with other approximate MLE's, such as those obtained by Le Cam's method of fitting parabolas [Le Cam (1974)]. Theoretically, such developments involve second-order asymptotics and are beyond the scope of this paper. Section 1.3 gives a numerical comparison for the Cauchy location model. The proof of Theorem 3.1, given later, also yields

COROLLARY. *If* $\bar{\theta}_n$ *is the actual MLE and* $\hat{\theta}_n$ *is the SMLE, then* $n^{1/2}(\bar{\theta}_n - \hat{\theta}_n) \to 0$ *in* $Q_{\theta_n}^n \otimes \mu_n^{j_n}$ *probability.*

The proof of Theorem 3.1 depends on a convergence lemma for the ess sup operation. Let $\mathbb{Z}$ be the metric space of all probabilities on Θ with, e.g., the Prohorov metric. Let $\mathbb{Y} = C_0(R^d)$ and let $\mathbb{Y}_0$ be those elements of $\mathbb{Y}$ with a unique maximum. For measures $m \in \mathbb{Z}$ and functions $g \in \mathbb{Y}$ define the functional $\Lambda_n(m, g) = t^*$, where t^* is any point satisfying $g(t^*) > \operatorname{ess\,sup}_m g - n^{-1}$. Define $\Lambda(g)$ to be the unique point at which $g \in \mathbb{Y}_0$ achieves its maximum. Let m be a fixed measure putting positive mass on every open subset of R^d.

CONVERGENCE LEMMA 3.1. *If* $m_n \to m$ *in* $\mathbb{Z}$ *and* $g_n \to g \in \mathbb{Y}_0$, *then*

$$\operatorname*{ess\,sup}_{m_n} g_n \to \sup_t g(t), \tag{3.6a}$$

$$\Lambda_n(m_n, g_n) \to \Lambda(g). \tag{3.6b}$$

PROOF OF THEOREM 3.1. Since the Q_θ's are mutually absolutely continuous, $\hat{\theta}_n$ satisfies $L_n(\hat{\theta}_n, \theta_n) = \max_{i \le j_n} L_n(\tilde{\theta}_n + n^{-1/2}t_i, \theta_n)$. Let W_n be the stochastic process with paths in $C_0(R^d)$ given by $W_n(t) = L_n(\theta_n + n^{-1/2}t, \theta_n)$. Let $\hat{\mu}_n$ be the empirical distribution of $(t_1, \dots, t_{j_n})$ viewed as a function of $\mathbb{X}_n \times (R^d)^{j_n}$. Let $\hat{\mu}_{nc}$ be the random centering of $\hat{\mu}_n$ given by $\hat{\mu}_{nc}(A) = \hat{\mu}_n(A - Y_n)$, where

$Y_n = (\tilde{\theta}_n - \theta_n)n^{1/2}$. Then $\hat{\theta}_n$ satisfies

$$L_n(\hat{\theta}_n, \theta_n) = \operatorname*{ess\,sup}_{\hat{\mu}_{nc}} W_n(\cdot) \equiv W_n(T_n), \tag{3.7}$$

where $T_n = \Lambda_n(\hat{\mu}_{nc}, W_n) = n^{1/2}(\hat{\theta}_n^* - \theta_n)$.

Note that W_n converges in $C_0(R^d)$ to W [hypothesis (3.4b)] and that W has a unique maximum at $I^{-1}(\theta_0)N$. Next, let $\{\theta_n'\}$ be any sequence satisfying $n^{1/2}|\theta_n' - \theta_0| \le c$. By hypothesis (3.4a), $\rho(\mathbb{D}_n(\theta_n'), \mu_0) \to 0$. The tightness of $n^{1/2}(\tilde{\theta}_n - \theta_0)$ under $Q^n_{\theta_n}$ and Lemma 2.1 then imply $\rho(\mu_n, \mu_0) \equiv \rho(\mathbb{D}_n(\tilde{\theta}_n), \mu_0) \to 0$ in $Q^n_{\theta_n}$-probability. By Theorem 2.1, $\rho(\hat{\mu}_n, \mu_0) \to 0$ in $Q^n_{\theta_n} \otimes \mu_n^{j_n}$ probability. Hájek's (1970) convolution theorem implies that μ_0, the limit of $n^{1/2}(\tilde{\theta}_n - \theta_n)$, is the convolution of some probability with $N(0, I^{-1}(\theta_0))$; hence, μ_0 gives positive mass to every open set. Lemma 2.1 implies that $\hat{\mu}_{nc}$, as a random probability measure, converges in distribution to the random measure $\mu_0(\cdot - Y)$, where Y has distribution μ_0.

Each realization of this limit measure gives positive mass to every open set. The sequence $\{(\hat{\mu}_{nc}, W_n)\}$ is tight; application of Lemma 2.1 and Convergence Lemma 3.1 to a weakly convergent subsequence of $\{(\hat{\mu}_{nc}, W_n)\}$, indexed by n', shows that $\Lambda_{n'}(\hat{\mu}_{n'c}, W_{n'})$ converges to $I^{-1}(\theta_0)N$. Since this limit does not depend on the subsequence selected, the entire sequence converges. □

PROOF OF CONVERGENCE LEMMA 3.1. Let $[f]_n$ denote $\operatorname{ess\,sup}_{m_n} f$. Then by the triangle inequality

$$|[g_n]_n - [g]_n| \le [|g_n - g|]_n \le \sup_t |g(t) - g_n(t)| \to 0. \tag{3.8}$$

If $A_\varepsilon = \{x\colon\ g(x) > \sup g - \varepsilon\}$, then A_ε is open, $m(A_\varepsilon) > 0$ and so $\liminf m_n(A_\varepsilon) \ge m(A_\varepsilon) > 0$. Hence, $[g]_n \ge \sup g - \varepsilon$ for all large n, implying part (a). To prove (b), let $B_n = \{v\colon\ g_n(v) > [g_n]_n - n^{-1}\}$. Then for sufficiently large n,

$$\begin{aligned} B_n &\subset \{v\colon g(v) > [g_n]_n - 2n^{-1}\}, \quad \text{when } \|g - g_n\| < n^{-1} \\ &\subset \{v\colon g(v) > \sup g - 3n^{-1}\}, \quad \text{using part (a)}. \end{aligned} \tag{3.9}$$

This last set shrinks down, as $n \to \infty$, to the unique maximizing point of g. □

3.3. *Estimated distribution of the stochastic MLE.* This subsection analyzes a bootstrap estimate of the distribution of $n^{1/2}(\hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n) - \theta)$ under $Q^n_{\theta_n} \otimes \mu_n^{j_n}$ when θ is unknown and where $\hat{\theta}_n$ is the local stochastic MLE. This estimate $\hat{\nu}_n$, described later, has the same distribution as the estimate of Section 1.5; however, in order to make apparent the application of Theorem 2.1 and the substitution theorems, it is represented on a more convenient space and its dependence on various auxiliary empiricals is made explicit.

For $\theta \in \Theta = R^d$, $\mathbf{x}_n \in \mathbb{X}_n$ and m a probability on R^d, define $\theta_n(m, \mathbf{x}_n)$ to be any point satisfying $L_n(\theta_n(m, \mathbf{x}_n), \theta; \mathbf{x}_n) \ge \operatorname{ess\,sup}_m L_n(\theta + n^{-1/2}(\cdot), \theta; \mathbf{x}_n) - n^{-1}$. Let $\xi_n(m, \theta)$ be the distribution of $n^{1/2}(\theta_n(m, \mathbf{x}_n) - \theta)$ when $\mathbf{x}_n$ has distribution

Q_θ^n. For $(\mathbf{x}_n, \mathbf{t}_n) \in \mathbb{X}_n \times (R^d)^{j_n}$, let $\nu_n((\mathbf{x}_n, \mathbf{t}_n), \cdot) = \xi_n(\hat{\mu}_n(\mathbf{x}_n, \mathbf{t}_n), \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n))$, where $\hat{\theta}_n$ is the stochastic MLE and $\hat{\mu}_n$ is the empirical distribution of $\mathbf{t}_n = (t_i, \dots, t_{j_n})$ viewed as a function on $\mathbb{X}_n \times (R^d)^{j_n}$. Let $\hat{\nu}_n$ be the empirical distribution of $\mathbf{u}_n = (u_1, \dots, u_{k_n})$: If $\mathbf{x}_n \in \mathbb{X}_n$, $\mathbf{t}_n \in (R^d)^{j_n}$ and $\mathbf{u}_n \in (R^d)^{k_n}$, then $\hat{\nu}_n(\mathbf{x}_n, \mathbf{t}_n, \mathbf{u}_n; \cdot)$ puts mass k_n^{-1} at each u_i. If $(Q_{\theta_n}^n \otimes \mu_n^{j_n}) \otimes \nu_n^{k_n}$ is the measure put on the foregoing product space, then given $\mathbf{x}_n, \mathbf{t}_n$, $\hat{\nu}_n$ is the empirical distribution of a sample of size k_n from $\nu_n((\mathbf{x}_n, \mathbf{t}_n), \cdot)$.

THEOREM 3.2. *Let* $\nu_0 = N(0, I(\theta_0)^{-1})$. *Assume hypotheses* (3.4a) *and* (3.4b) *and let* $\{\theta_n\}$ *satisfy* $n^{1/2}|\theta_n - \theta_0| \le c$. *Then, if* $k_n, j_n \to \infty$,

$$\rho(\hat{\nu}_n, \nu_0) \to 0, \quad \text{in } \left(Q_{\theta_n}^n \otimes \mu_n^{j_n}\right) \otimes \nu_n^{k_n} \text{ probability.} \tag{3.10}$$

PROOF. By Theorem 2.1 it suffices to show that $\rho(\nu_n, \nu_0) \to 0$ in $Q_{\theta_n}^n \otimes \mu_n^{j_n}$ probability. The argument given in the proof of Theorem 3.1 shows that whenever $\{\eta_n\}$ is a sequence of probability measures converging weakly to μ_0 and whenever $\{\theta_n'\}$ satisfies $n^{1/2}|\theta_n' - \theta_0| \le c$, then $\xi_n(\eta_n, \theta_n')$ converges to ν_0 in Prohorov distance. Theorem 3.1 also showed that the stochastic maximum likelihood estimate $\hat{\theta}_n$ has the property that $n^{1/2}|\hat{\theta}_n - \theta_0|$ is tight, and that $\hat{\mu}_n \to \mu_0$ in probability. By Lemma 2.1, $\xi_n(\hat{\mu}_n, \hat{\theta}_n) \equiv \nu_n$ converges to ν_0. □

Confidence sets. Let $|\cdot|$ denote any norm on R^d, and let $d_{n,\alpha} = d_\alpha(\mathbf{x}_n, \mathbf{t}_n, \mathbf{u}_n)$ be $(1-\alpha)$th quantile of the cdf $\hat{\nu}_n\{y \in R^d: |y| \le r\}$. Set $\hat{C}_{n,\alpha} = \{\theta: |\theta - \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n)| \le d_{n,\alpha}(\mathbf{x}_n, \mathbf{t}_n, \mathbf{u}_n)\}$. Theorem 3.2 implies that $\hat{C}_{n,\alpha}$ is a confidence set for θ having asymptotic level $1-\alpha$. Moreover, the local asymptotic minimax property described in Beran and Millar (1985) holds for $\hat{C}_{n,\alpha}$. This confidence set has the same distribution as the one described in the introduction. Variants of the method of this section will yield stochastic analogues of confidence ellipsoids of the classical theory as well. The possibility of bootstrapping a centered statistic in order to produce confidence sets was first pointed out by Efron (1979). More recent theory for bootstrap procedures can be found in Beran (1984), Bickel and Freedman (1981) and Singh (1981), for example.

4. Stochastic minimum distance procedures.

4.1. *Construction and asymptotics*. Let $\{Q_\theta, \theta \in \Theta\}$ be a family of cdf's on R^q, where Θ is an open subset of R^d. Let $\mathbb{X}_n$ be the n-fold product of R^q; if $\mathbf{x}_n = (x_1, \dots, x_n) \in \mathbb{X}_n$, let $\hat{Q}_n = \hat{Q}_n(\mathbf{x}_n, \cdot)$ be the empirical cdf of $\mathbf{x}_n$. Let $|\cdot|$ be a norm on real functions of R^q, such as L^p-norm or a supremum norm (required properties of $|\cdot|$ are given later). The minimum distance test statistic for the hypothesis $\{Q_\theta, \theta \in \Theta\}$ is

$$n^{1/2} \inf_\theta |\hat{Q}_n - Q_\theta| \tag{4.1}$$

and a minimum distance estimate $\bar{\theta}_n$ is any point in Θ satisfying

$$\inf_\theta n^{1/2}|\hat{Q}_n - Q_\theta| \ge n^{1/2}|\hat{Q}_n - Q_{\bar{\theta}_n}| - n^{-1}. \tag{4.2}$$

To define stochastic versions of these quantities, let $\tilde{\theta} = \tilde{\theta}_n(\mathbf{x}_n)$ be a preliminary estimate of θ. The stochastic minimum distance estimate is any random variable $\hat{\theta}_n$ on $\mathbb{X}_n \times (R^d)^{j_n}$ satisfying

$$\min_{i \le j_n} \left|\hat{Q}_n - Q(\tilde{\theta}_n + n^{-1/2} t_i)\right| = \left|\hat{Q}_n - Q(\hat{\theta}_n)\right|, \tag{4.3}$$

where $\mathbf{t}_n = (t_1, \dots, t_{j_n}) \in (R^d)^{j_n}$ and we have written $Q_\theta = Q(\theta)$. Then $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n)$. The stochastic minimum distance test statistic is

$$n^{1/2}\left|\hat{Q}_n - Q(\hat{\theta}_n)\right|. \tag{4.4}$$

Let $\mathbb{D}_n(\theta)$ be the distribution of $n^{1/2}(\tilde{\theta}_n - \theta)$ when $\mathbf{x}_n$ has distribution Q_θ^n and define the Markov kernel μ_n by

$$\mu_n(\mathbf{x}_n, \cdot) = \mathbb{D}_n(\tilde{\theta}_n). \tag{4.5}$$

If θ_n is the true parameter at time n, put the measure $Q_{\theta_n}^n \otimes \mu_n^{j_n}$ on $\mathbb{X}_n \times (R^d)^{j_n}$; notice that this is exactly the search structure used in the MLE case.

Asymptotics for the stochastic minimum distance estimate and test statistic will be developed under the following hypothesis. Fix $\theta_0 \in \Theta$ and let B be a Banach space of real functions with norm $|\cdot|$ such that $\hat{Q}_n - Q_\theta \in B$ for all θ. Assume:

(4.6a) If $Q_{\theta_n} - Q_\theta$ converges to 0 in B, then $\theta_n \to \theta$.

(4.6b) There is a bounded linear map l from R^d to B such that $Q_\theta - Q_{\theta_0} = l(\theta - \theta_0) + o(|\theta - \theta_0|)$ and such that $|l(\theta)| \ge c_0|\theta|$ for all θ, for some positive c_0.

(4.6c) If $\{\theta_n\}$ is a sequence such that $n^{1/2}|\theta_n - \theta_0| \le c$, then there is a B-valued r.v. W such that $n^{1/2}(\hat{Q}_n - Q_{\theta_n})$ converges to W, weakly in B, under $Q_{\theta_n}^n$.

(4.6d) The preliminary estimator $\tilde{\theta}_n$ satisfies (3.4a).

Hypotheses (4.6a–4.6c) are standard for the minimum distance problem: See, e.g., Pollard (1980) or Millar (1984). Hypothesis (4.6c) is known to be satisfied if $|\cdot|$ is supremum norm or an L^p-norm relative to a finite measure.

THEOREM 4.1. *Assume* (4.6a)–(4.6d), *that* $\{\theta_n\}$ *satisfies the condition in* (4.6c) *and that* $j_n \to \infty$.

(a) *Under* $Q_{\theta_n}^n \otimes \mu_n^{j_n}$,

$$n^{1/2}\left|\hat{Q}_n - Q(\hat{\theta}_n)\right| \Rightarrow \min_\theta |W - l(\theta)|. \tag{4.7}$$

(b) *Let M be the random set of θ's that achieve* $\min|W - l(\theta)|$. *If A is a closed set in R^d and C is open, then*

$$\limsup_{n\to\infty} Q_{\theta_n}^n \otimes \mu_n^{j_n}\left\{(\hat{\theta}_n - \theta_n)n^{1/2} \in A\right\} \le P\{M \cap A \ne \varnothing\} \tag{4.8}$$

and

$$\liminf_{n\to\infty} Q_{\theta_n}^n \otimes \mu_n^{j_n}\left\{(\hat{\theta}_n - \theta_n)n^{1/2} \in C\right\} \ge P\{M \subset C\}. \tag{4.9}$$

In case M consists of a single random variable η, the conclusions of (b) simplify; unfortunately, the existence of asymptotically unique minima appear easily guaranteed only in the case that $|\cdot|$ is Hilbertian.

COROLLARY. *If the random set M consists of a single random variable η, then*

$$n^{1/2}(\hat{\theta}_n - \theta_n) \Rightarrow \eta, \quad \text{under } Q^n_{\theta_n} \otimes \mu^{j_n}_n. \tag{4.10}$$

REMARKS. If $\bar{\theta}_n$ is the actual minimum distance estimate, the conclusions (a) and (b) hold with $\bar{\theta}_n$ replacing $\hat{\theta}_n$. See Pollard (1980) for a formulation related to (b) but in terms of convergence of random sets. Theorem 4.1 presents only a relatively simple illustration of the use of stochastic methods in minimum distance problems. For example, more complicated functionals [such as those in Millar (1984)] are amenable to similar analysis. Characterization of η in the Hilbert case can be found in Millar (1984); in this case, the stochastic minimum distance estimate is locally asymptotically minimax and asymptotically normal. In general M will contain more than one point; it is easy to see, however, that it is compact and convex. The proof of Theorem 4.1 requires a convergence lemma, whose proof is similar to that in Section 3.

CONVERGENCE LEMMA 4.1. *Let g_n, g be nonnegative continuous functions on R^d, g_n converging uniformly to g on balls. Assume that for $n \geq n_0$, there is a ball containing all the minima of g_n. Let m_n, m be probabilities on R^d, $m_n \Rightarrow m$, m having full support. Define $\Lambda_n(m_n, g_n) = t^*_n$, where t^*_n is any point satisfying $g_n(t^*_n) \leq \operatorname{ess\,inf}_{m_n} g_n + n^{-1}$. Let M be the set of minima of g, and let $\Lambda^0_n(m_n, g_n) = \operatorname{ess\,inf}_{m_n} g_n$. Then*

$$\Lambda^0_n(m_n, g_n) \to \min_t g(t), \tag{4.11a}$$

$$d(t^*_n, M) \to 0. \tag{4.11b}$$

PROOF OF THEOREM 4.1. Let $W_n = (\hat{Q}_n - Q_{\theta_n})n^{1/2}$. Let $\hat{\mu}_n$ and $\hat{\mu}_{nc}$ be defined as in the proof of Theorem 3.1. Let M_n be the set of θ such that $n^{1/2}|\hat{Q}_n - Q_\theta| \leq \min_\theta |\hat{Q}_n - Q_\theta| n^{1/2} + n^{-1}$. The theory of minimum distance estimators [see Pollard (1980) or Millar (1984)] shows that M_n is contained in a ball about θ_0 of radius $\hat{C}_n n^{-1/2}$, where $\{\hat{C}_n\}$ is a tight sequence. The differentiability hypothesis (4.6b) implies that $n^{1/2}[\hat{Q}_n - Q_{\theta_0}(\theta_0 + vn^{-1/2})]$ is approximatable by $W_n + l(v)$, uniformly in v, for v constrained to balls of fixed radius. Moreover, the minima of $|W_n + l(v)|$ are all contained in a v-ball of random radius η_n, where $\{\eta_n\}$ is tight. Regard $W_n + l(v)$, $v \in R^d$, as a random element of $C(R^d)$, where the metric is that of uniform convergence on balls. Then the B-valued random elements $W_n + l(v)$ converge weakly to $W + l(v)$, by (4.6c).

For all large n, therefore, $n^{1/2}|\hat{Q}_n - Q(\hat{\theta}_n)| = \Lambda^0_n(\hat{\mu}_{nc}, \hat{g}_n)$, where $\hat{g}_n = |W_n + l(v)|$. The auxiliary empiricals $\hat{\mu}_{nc}$ converge, as in the proof of Theorem 3.1. Part (a) of the theorem now follows from Lemma 2.1, applied to part (a) of Convergence Lemma 4.1.

To prove part (b), note that for all large n, $n^{1/2}(\hat{\theta}_n - \theta_n) = \Lambda_n(\hat{\mu}_{nc}, \hat{g}_n)$ (this assertion depends on the lack of uniqueness in the definition of Λ_n). A variant of Lemma 2.1, mentioned in Section 2, applied to part (b) of Convergence Lemma 4.1, shows that if $\varepsilon > 0$, then for any Borel set D

$$\begin{aligned} P\{M^{\varepsilon} \subset D\} &\leq \liminf_{n\to\infty} Q^n_{\theta_n} \otimes \mu_n^{j_n}\{n^{1/2}(\hat{\theta}_n - \theta_n) \in D\} \\ &\leq \limsup_{n\to\infty} Q^n_{\theta_n} \otimes \mu_n^{j_n}\{n^{1/2}(\hat{\theta}_n - \theta_n) \in D\} \qquad (4.12) \\ &\leq P\{M^{\varepsilon} \cap D \neq \varnothing\}, \end{aligned}$$

where $M^{\varepsilon} = \{y\colon d(y, M) < \varepsilon\}$. Let $\varepsilon \downarrow 0$ in (4.12) for D open, closed to complete the proof. □

4.2. *Estimated distribution of the stochastic minimum distance test statistic.* This subsection analyzes a bootstrap estimate of the distribution of the stochastic discrepancy $n^{1/2}|\hat{Q}_n - Q(\hat{\theta}_n)|$, where $\hat{\theta}_n$ is the stochastic MDE. The development closely parallels that of Section 3.3.

For $\theta \in \Theta$ and m a probability on R^d, define $\xi_n(m, \theta)$ to be the distribution of $\operatorname{ess\,inf}_m n^{1/2}|\hat{Q}_n(\mathbf{x}_n, \cdot) - Q(\theta)|$ when $\mathbf{x}_n$ has distribution Q^n_{θ}. For $(\mathbf{x}_n, \mathbf{t}_n) \in \mathbb{X}_n \times (R^d)^{j_n}$, let $\nu_n(\mathbf{x}_n, \mathbf{t}_n; \cdot) = \xi_n(\hat{\mu}_n(\mathbf{x}_n, \mathbf{t}_n), \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n))$. Let $\hat{\nu}_n$ be the empirical distribution of a sample of size k_n from ν_n (cf. Section 3.3), viewed as a function on $[\mathbb{X}_n \times \Theta^{j_n}] \times (R^d)^{k_n}$, and let ν_0 be the distribution of $\min_v|W + l(v)|$. Then $\hat{\nu}_n$ is a bootstrap estimate of the distribution of the stochastic minimum distance test statistic.

THEOREM 4.2. *Assume hypothesis* (4.6); *let* $\{\theta_n\}$ *satisfy* $n^{1/2}|\theta_n - \theta_0| \leq c$. *Suppose* $k_n, j_n \to \infty$. *Then*

$$\rho(\hat{\nu}_n, \nu_0) \to 0, \quad \text{in } \left(Q^n_{\theta_n} \otimes \mu_n^{j_n}\right) \otimes \nu_n^{k_n} \text{ probability.}$$

The proof is similar to that of Theorem 3.1, and will be omitted.

Testing. Let $\hat{r}_n$ be a $(1-\alpha)$th quantile of the cdf $\hat{\nu}_n$. The test which rejects when $|\hat{Q}_n - Q(\hat{\theta}_n)|n^{1/2} > \hat{r}_n$ is (by Theorem 4.1) an asymptotically level α test of the null hypothesis $\{Q_\theta, \theta \in \Theta\}$. This test has the same distribution as the one described in the introduction; it is just represented on a more convenient space. Computational feasibility is one of the attractive features of this goodness-of-fit test. Comparisons of performance with other goodness-of-fit tests would be of interest; such an undertaking could involve simulated power functions as in Beran (1986).

5. Stochastic MLE: Global search.

5.1. *Asymptotic normality.* Let $\{Q_\theta, \theta \in \Theta\}$, $\mathbb{X}_n$, L_n and $f^n(\theta; \cdot)$ be as in Section 3 and take $\Theta = R^d$ for convenience. For $\mathbf{x}_n \in \mathbb{X}_n$, let $\hat{\Gamma}_n(\mathbf{x}_n)$ be a $d \times d$ random matrix and $\hat{C}_n(\mathbf{x}_n)$ an R^d-valued random variable. The global stochastic

MLE, $\hat{\theta}_n$, is defined by

(5.1) $$\max_{i \le j_n} f^n\big(\hat{\Gamma}_n(\mathbf{x})\big(z_i - \hat{C}_n(\mathbf{x}_n)\big); \mathbf{x}_n\big) = f^n\big(\hat{\theta}_n; \mathbf{x}_n\big),$$

for $\mathbf{z}_n = (z_1, \dots, z_{j_n}) \in \Theta^{j_n}$. Then $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}_n, \mathbf{z}_n)$ is an R^d-valued random variable on $\Theta^{j_n} \times \mathbb{X}_n$.

Let μ_0 be a probability on Θ and put the measure $Q^n_{\theta_n} \times \mu_0^{j_n}$ on $\mathbb{X}_n \otimes \Theta^{j_n}$. Then $\mathbf{z}_n = (z_1, \dots, z_{j_n}) \in \Theta^{j_n}$ is a realized vector of i.i.d. (μ_0) random variables, independent of $\mathbf{x}_n$; the search in (5.1) then consists of scanning randomly centered and scaled z_i's, where the centering and scaling variables can depend on the original data $\mathbf{x}_n$.

Make the following assumptions. Fix θ_0.

(5.2a) $|\hat{\Gamma}_n|$, $|\hat{\Gamma}_n^{-1}|$ and $\hat{C}_n$ are tight under $Q^n_{\theta_n}$ whenever $\{\theta_n\}$ satisfies $|\theta_n - \theta_0| \le cn^{-1/2}$. (Here $|\hat{\Gamma}_n|$ is the usual operator norm.)

(5.2b) μ_0 has a density with respect to Lebesgue measure that is bounded away from 0 on compacts.

(5.2c) Hypothesis (3.4b) holds.

THEOREM 5.1. *Assume* (5.2) *and suppose* $\lim j_n n^{-d/2} = \infty$. *For every* $\{\theta_n\}$ *such that* $n^{1/2}|\theta_n - \theta_0| \le c$,

$$n^{1/2}\big(\hat{\theta}_n - \theta_n\big) \Rightarrow N\big(0, I^{-1}(\theta_0)\big), \quad \textit{under } Q^n_{\theta_n} \otimes \mu_0^{j_n}.$$

REMARKS. As in the local case, this result implies that $\hat{\theta}_n$ is efficient and locally asymptotic minimax. Despite the rate requirement on j_n, the global approach may be of use when one does not have a preliminary estimate with which to begin the local search. Hypotheses can be weakened but at the price of a much longer proof. The proof depends upon a real variable lemma.

CONVERGENCE LEMMA 5.1. *Let g_n, g be continuous, real, bounded functions on R^d, with g_n converging to g in $C_0(R^d)$. Assume g has a unique maximum at b_0. Put the probability measure μ_0^∞ on $(Z_1, Z_2, \dots)$, where μ_0 satisfies* (5.2b). *Let $\{M_n\}$ be a sequence of matrices with $|M_n|$, $|M_n^{-1}|$ bounded. Let $\{a_n\}$ be a bounded sequence in R^d. Let b_n be defined by*

(5.3) $$\max_{i \le j_n} g_n\big(n^{1/2} M_n(z_i - a_n)\big) \le g\big(b_n\big) + n^{-1}.$$

If $j_n n^{-d/2} \to \infty$, then

(5.4a) $$\max_{i \le j_n} g_n\big(n^{1/2}(M_n(z_i - a_n))\big) \to g\big(b_0\big) \quad a.e.\ \big(\mu_0^\infty\big),$$

(5.4b) $$b_n \to b_0 \quad a.e.\ \big(\mu_0^\infty\big).$$

PROOF OF THEOREM 5.1. Using the notation of Convergence Lemma 5.1, let $Z = (z_1, z_2, \dots)$ and let $\Lambda_n(Z_n; g_n, M_n, a_n) = b_n$, where Z_n is the r.v. defined by $Z_n(z) = (z_1, \dots, z_{j_n})$. Define the stochastic process W_n by $W_n(t) = L_n(\theta_n + n^{-1/2}t, \theta_n)$. Replace g_n in Λ_n by W_n, M_n by $\hat{\Gamma}_n$ and a_n by $\hat{C}_{nc} \equiv \hat{C}_n + \theta_n$.

Take $Y_n = (W_n, \hat{\Gamma}_n, \hat{C}_{nc})$ and note that Z_n, Y_n are independent; Y_n can be assumed to converge weakly by passing to a subsequence since $\{\hat{\Gamma}_n\}, \{\hat{C}_n\}$ are tight. By (5.2c), W_n converges weakly in $C_0(R^d)$ to W, which has a unique maximum at $I^{-1}(\theta_0)N$ [see (3.4b) for the definition of N]. Application of Lemma 2.2(i) implies that $\hat{b}_n \equiv \Lambda_n(Z_n; W_n, \hat{\Gamma}_n, \hat{C}_n)$ converges in distribution to $I^{-1}(\theta_0)N$. On the other hand, since the Q_θ's are mutually absolutely continuous, the global stochastic MLE $\hat{\theta}_n$ satisfies

$$(5.5)\quad L_n(\hat{\theta}_n, \theta_n) = \max_{i \le j_n}\left(\hat{\Gamma}_n(z_i - \hat{C}_n), \theta_n\right) = \max_{i \le j_n} W_n\left(n^{1/2}\hat{\Gamma}_n(z_i - \hat{C}_{nc})\right),$$

so $\hat{\theta}_n = \theta_n + n^{-1/2}\hat{b}_n$, proving the result. □

5.2. *The estimated distribution.* Let $\mathbb{D}_n(\mathbf{z}_n, \theta)$ be the distribution of $n^{1/2}(\hat{\theta}_n(\mathbf{x}_n, \mathbf{z}_n) - \theta)$ when $\mathbf{z}_n = (z_1, \dots, z_{j_n})$ is fixed and $\mathbf{x}_n$ has distribution Q_θ^n. Let $\nu_n(\mathbf{x}_n, \mathbf{z}_n; \cdot)$ be the random probability measure on $\mathbb{X}_n \times \Theta^{j_n}/\Theta$ given by $\nu_n(\mathbf{x}_n, \mathbf{z}_n; \cdot) = \mathbb{D}_n(\mathbf{z}_n, \hat{\theta}_n(\mathbf{x}_n, \mathbf{z}_n))$. Let $\hat{\nu}_n$ be the distribution of $\mathbf{u}_n = (u_1, \dots, u_{k_n}) \in \Theta^{k_n}$, viewed as a function on $(\mathbb{X}_n \times \Theta^{j_n}) \times \Theta^{k_n}$. If $Q_{\theta_n}^n$ is the true distribution of $\mathbf{x}_n$ and if the measure $(Q_{\theta_n}^n \otimes \mu_0^{j_n}) \otimes \nu_n^{k_n}$ is placed on $(\mathbb{X}_n \times \Theta^{j_n}) \times \Theta^{k_n}$, then $\hat{\nu}_n$ is the *estimated distribution* of the normalized stochastic MLE. This particular $\hat{\nu}_n$ assumes that the search variables are held fixed throughout the generation of the k_n auxiliary variables (cf. Section 1.3).

THEOREM 5.2. *Assume the hypothesis of Theorem* 5.1, *and let* $\nu_0 = N(0, I^{-1}(\theta_0))$. *Then if* $k_n \to \infty$,

$$(5.6)\qquad \rho(\hat{\nu}_n, \nu_0) \to 0, \quad \textit{in } \left(Q_{\theta_n}^n \otimes \mu_0^{j_n}\right) \otimes \nu_n^{k_n} \textit{ probability.}$$

PROOF. By Theorem 2.1 it suffices to show that $\rho(\nu_n, \nu_0) \to 0$ in $Q_{\theta_n}^n \otimes \mu^{j_n}$ probability. Let $\{\theta_n'\}$ be a sequence such that $|\theta_n' - \theta_0| \le c_n^{-1/2}$. In the proof of Theorem 5.1 apply Lemma 2.2(ii) instead of Lemma 2.2(i). This shows that $\rho(\mathbb{D}_n(\mathbf{z}_n, \theta_n'), \nu_0) \to 0$ in $\mu_0^{j_n}$ probability. By Theorem 5.1, $n^{1/2}(\hat{\theta}_n - \theta_0)$ is tight under $Q_{\theta_n}^n \otimes \mu_0^{j_n}$. Therefore, by a variant of Lemma 2.2(i), $\rho(\mathbb{D}_n(\mathbf{z}_n, \hat{\theta}_n), \nu_0) \to 0$, which is the desired result. □

5.3. *Proof of Convergence Lemma* 5.1. Let $u(n, i) = n^{1/2}M_n(z_i - a_n)$. By the triangle inequality and the definition of supremum norm $\|\cdot\|$,

$$(5.7)\qquad \left|\max_{i \le j_n} g_n(u(n,i)) - \max_{i \le j_n} g(u(n,i))\right| \le \|g_n - g\| \to 0.$$

Let $g^+ = g(b_0)$. Then

$$(5.8)\qquad \mu_0^\infty\left\{\max_{i \le j_n} g(u(n,i)) > g^+ - \varepsilon\right\} = 1 - \delta_n^{j_n},$$

where $\delta_n = \mu_0^\infty\{g(u(n,1)) \le g^+ - \varepsilon\}$, so it suffices to show $\delta_n^{j_n} \to 0$ to prove (a). Since g is continuous, $A_\varepsilon \equiv \{x: g(x) > g^+ - \varepsilon\}$ is open and so contains a ball

$B(z_0, r)$ of center z_0, radius r. Hence,

$$(5.9)\quad \begin{aligned} 1 - \delta_n &\geq \mu_0\{u(n,1) \in B(z_0, r)\} \\ &= \mu_0\{x: x \in a_n + n^{-1/2}M_n^{-1}(B(z_0, r)\}. \end{aligned}$$

The x-set on the right side of (5.9) contains a ball with center $b_n = a_n + n^{-1/2}M_n^{-1}z_0$, which is bounded as $n \to \infty$, and radius $r_n \geq qrn^{-1/2}$, where $q \geq \min_n |M_n|^{-1} > 0$, since $|M_n|, |M_n^{-1}|$ are bounded. But $\mu_0(B(b_n, qrn^{-1/2})) \geq c_0$Lebesgue meas$\{B(b_n, qrn^{-1/2}\} \geq c_1 n^{-k/2}$ for some constants c_1, c_0, since b_n is bounded and the density of μ_0 is bounded away from 0 compacts. Hence, $1 - \delta_n \geq c_1 n^{-k/2}$ and so $\delta_n^{j_n} \leq (1 - c_1 n^{-k/2})^{j_n} \to 0$, proving (a). Part (b) is proved much as Lemma 3.1. □

6. Stochastic likelihood ratio statistic.

6.1. *Asymptotic distribution.* Let $\{Q_\theta, \theta \in \Theta\}$, $\mathbb{X}_n$, $f^n(\theta; \mathbf{x}_n)$, $L_n(\theta, \theta_1; \mathbf{x}_n)$, $\tilde{\theta}_n$, Θ, $\mathbb{D}_n(\theta)$ and μ_n be as in Section 3; let $\hat{\theta}_n$ be the local stochastic MLE defined in (3.2). Let d_0, d_1 be integers, $d = d_0 + d_1$, where d is the dimension of Θ. Any $s \in R^d$ can be written $s = (s_0; s_1)$, where $s_0 \in R^{d_0}$, $s_1 \in R^{d_1}$. Fix $\theta_{00} \in R^{d_0}$ and let $\Theta_0 = \{\theta \in \Theta: \theta = (\theta_{00}; \theta_1)$ for some $\theta_1 \in R^{d_1}\}$.

Decompose the preliminary estimator $\tilde{\theta}_n$ into $\tilde{\theta}_n = (\tilde{\theta}_{n0}; \tilde{\theta}_{n1})$, $\tilde{\theta}_{ni} \in R^{d_i}$. Define $\tilde{\theta}_n^0 = (\theta_{00}, \tilde{\theta}_{n1})$ and define $\hat{\theta}_n^0$, the stochastic MLE when θ is restricted to Θ_0, via the auxiliary sample $(t_1, \dots, t_{j_n^0})$, $t_i \in R^d$, by

$$f^n(\hat{\theta}_n^0; \mathbf{x}_n) = \max_{i \leq j_n^0} f^n(\tilde{\theta}_n^0 + n^{-1/2}t_i; \mathbf{x}_n),$$

where $\mathbf{x}_n \in \mathbb{X}_n$.

Let $\mu_n^0(\mathbf{x}_n, \cdot) = \mathbb{D}_n(\tilde{\theta}_n^0)$. Note that, for each $\mathbf{x}_n$, $\mu_n^0(\mathbf{x}_n, \cdot)$ is a probability on $\{0\}^{d_0} \times R^{d_1}$. If $\mu_n^0(\mathbf{x}_n, \cdot)$ is the conditional distribution put upon the t_i, then $\hat{\theta}_n^0$ is of the form $\hat{\theta}_n^0 = (\theta_{00}; \hat{\theta}_{n1}^0)$ for some $\hat{\theta}_{n1}^0 \in R^{d_1}$.

If θ_n is true at time n, put the measure $Q_{\theta_n}^n \otimes [(\mu_n^0)^{j_n^0} \times (\mu_n)^{j_n}]$ on $\mathbb{X}_n \times (R^d)^{j_n^0} \times (R^d)^{j_n}$. Then conditional on $\mathbf{x}_n$, the search samples for constructing $\hat{\theta}_n$ and $\hat{\theta}_n^0$ are independent. The stochastic likelihood ratio statistic for the null hypothesis Θ_0 is

$$(6.1)\quad T_n = f^n(\hat{\theta}_n; \mathbf{x}_n)/f^n(\hat{\theta}_n^0; \mathbf{x}_n).$$

Fix $\theta_0 = (\theta_{00}; \theta_{01})$ and let $\theta_n = \theta_n(h) \equiv \theta_0 + n^{-1/2}h = (\theta_{00} + n^{-1/2}h_0; \theta_{01} + n^{-1/2}h_1)$. Let $J(\theta_0) = I^{-1}(\theta_0)$, where $I(\theta_0)$ is the information matrix defined in (3.4b), and let $J_{00}(\theta_0)$ be the upper left $d_0 \times d_0$ submatrix of $J(\theta_0)$. Let $\nu_0(h)$ be the chi-squared distribution with d_0 degrees of freedom and noncentrality parameter $\langle h_0, J_{00}^{-1}(\theta_0)h_0\rangle$, where the brackets denote the inner product for R^{d_0}.

THEOREM 6.1. *Assume hypothesis* (3.4a) *and* (3.4b) *and that* n, j_n, j_n^0 *all tend to* ∞. *Then, under* $Q_{\theta_n}^n \otimes [(\mu_n^0)^{j_n^0} \times (\mu_n)^{j_n}]$ *and with* $\theta_n = \theta_n(h)$ *as above,* $2\log T_n \Rightarrow \nu_0(h)$.

PROOF. Since the Q_θ's are mutually absolutely continuous, $T_n = L_n(\hat{\theta}_n, \theta_n; x) / L_n(\hat{\theta}_n^0, \theta_n; x)$. Let $W_n(h) = L_n(\theta_n + n^{-1/2}h, \theta_n)$. Let $\hat{\mu}_n$ be the empirical distribution of the j_n auxiliary variables used to construct $\hat{\theta}_n$ (same $\hat{\mu}_n$ as in Section 3) and let μ_n^0 be the empirical of the j_n^0 auxiliary variables used to construct $\hat{\theta}_n^0$. For real functions f and a measure m, let $[f]_m$ denote $\operatorname{ess\,sup}_m f$. Then

$$2\log T_n = 2\log L_n(\hat{\theta}_n, \theta_n) - 2\log L_n(\hat{\theta}_n^0, \theta_n)$$
$$= 2\log[W_n(\cdot + Y_n)]_{\hat{\mu}_n} - 2\log[W_n((-h_0; Y_n^0) + \cdot)]_{\hat{\mu}_n^0},$$

where $Y_n = n^{1/2}(\hat{\theta}_n - \theta_n)$, $Y_n^0 = n^{1/2}(\hat{\theta}_{n1}^0 - \theta_{n1})$. By Section 3, $\hat{\mu}_n$ converges to a measure that puts positive mass m on every open set in R^d; the same argument shows that $\hat{\mu}_n^0$ has the same property on R^{d_1}. Also, W_n converges weakly in $C_0(R^d)$ to W and $\{Y_n\}$, $\{Y_n^0\}$ are tight. A simple variant of part (a) of Convergence Lemma 3.1 together with Lemma 2.1, now shows that $2\log T_n$ converges weakly to

$$\max_{t\in R^d} \log W(t) - \max_{t\in R^{d_1}} \log W((-h_0; t)).$$

Since $\log W(t) = t'N - 2^{-1}t'I(\theta_0)t$, one may compute these maxima explicitly; simplification of their difference by standard matrix identities eventually yields the desired result. □

6.2. *Estimated distribution.* For $\theta \in \Theta$ and m a probability on R^d, define $\theta_n(m, \mathbf{x}_n)$ by $f^n(\theta_n(m, \mathbf{x}_n); \mathbf{x}_n) = \operatorname{ess\,sup}_m f^n(\theta + n^{-1/2}(\cdot); \mathbf{x}_n)$. Decompose the θ just given into $\theta = (\theta_0; \theta_1)$ (cf. Section 6.1) and set $\theta^0 = (\theta_{00}; \theta_1)$. For m^0 a probability on R^{d_1} define $\theta_n^0(m^0; \mathbf{x}_n)$ by

$$f^n(\theta_n^0(m, \mathbf{x}_n); \mathbf{x}_n) = \operatorname*{ess\,sup}_{m^0} f^n(\theta^0 + n^{-1/2}(0; \cdot); \mathbf{x}_n).$$

Let $\xi_n(m, m^0; \theta)$ be the distribution of

$$2\log[f^n(\theta_n(m, \mathbf{x}_n); \mathbf{x}_n)/f^n(\theta_n^0(m^0, \mathbf{x}_n); \mathbf{x}_n)]$$

if $\mathbf{x}_n$ has distribution Q_θ^n. For $u \in \mathbb{U}_n \equiv \mathbb{X} \times (R^d)^{j_n^0} \times (R^d)^{j_n}$ and h_0 a fixed vector in R^{d_0}, define

$$\nu_n(u, \cdot) = \xi(\hat{\mu}_n, \hat{\mu}_n^0; \hat{\theta}_n^0 + n^{-1/2}(h_0; 0)). \tag{6.2}$$

Let $\hat{\nu}_n$ be the empirical distribution of k_n observations from ν_n, viewed as a function on $\mathbb{U}_n \times R^{k_n}$. Then $\hat{\nu}_n$ is the *estimated distribution* of the stochastic likelihood ratio statistic under the hypothesis $(\theta_{00} + n^{-1/2}h_0; \theta_1)$, the second component θ_1 being fixed but unknown.

Let $P_n = Q_{\theta_n}^n \otimes [(\mu_n)^{j_n^0} \times (\mu_n)^{j_n}]$, and let $\nu_0(h)$ be the chi-squared distribution given in Section 6.1.

THEOREM 6.2. *Assume the hypothesis of Theorem* 6.1. *If* $k_n \to \infty$, *then* $\rho(\hat{\nu}_n, \nu_0(h)) \to 0$ *in* $P_n \otimes \nu_n^{k_n}$ *probability.*

PROOF. By Theorem 2.1, it suffices to show that $\rho(\nu_n, \nu_0) \to 0$ in P_n probability. The argument for Theorem 6.1 shows that $\xi(m_n, m_n^0, \theta_n') \Rightarrow \nu_0$ whenever m_n, m_n^0 converge to measures with full support and when $\{\theta_n'\}$ satisfies $\theta_n' = (\theta_{00} + n^{-1/2}h_0; \theta_{01} + n^{-1/2}h_{1n})$, where $|h_{1n}| \le c$. Since analogous stochastic properties hold for $\hat{\mu}_n$, $\hat{\mu}_n^0$ and $\hat{\theta}_n^0 + n^{-1/2}(h_0; 0)$ by the proof of Theorem 6.1, we may apply Lemma 2.2(i) to complete the proof. □

REMARKS. Specialized to $h_0 = 0$, Theorem 6.2 implies that the $(1 - \alpha)$th quantile of $\hat{\nu}_n$ (i.e., the bootstrap critical value of the stochastic likelihood ratio test) converges in $P_n \otimes \nu_n^{k_n}$ probability to the $(1 - \alpha)$th quantile of the chi-squared distribution with d_0 degrees of freedom. It also implies that the asymptotic power of the stochastic likelihood ratio test against any sequence of alternatives of the form $(\theta_{00} + n^{-1/2}h_0, \theta_{01} + n^{-1/2}h_1)$ is the same as that of the classical likelihood ratio test.

7. Stochastic Bayes estimates. Let $\{Q_\theta, \theta \in \Theta\}$ be a family of probabilities on a space $\mathbb{X}$, with $\Theta = R^d$. Assume the Q_θ's are mutually absolutely continuous and each Q_θ is absolutely continuous with respect to Lebesgue measure. For $\theta_1, \theta \in \Theta$ and $\mathbf{x}_n \in \mathbb{X}_n \equiv \mathbb{X}^n$, let $L_n(\theta_1, \theta; \mathbf{x}_n) = dQ_{\theta_1}^n(\mathbf{x}_n)/dQ_\theta^n(\mathbf{x}_n)$, let $f^n(\theta, \mathbf{x}_n) = dQ_\theta^n(\mathbf{x}_n)/d\lambda(\mathbf{x}_n)$, where λ is Lebesgue measure. For $\mathbf{x}_n \in \mathbb{X}_n$ and $\mathbf{t}_n = (t_1, \dots, t_{j_n}) \in \Theta^{j_n}$ define $\hat{\theta}_n$, the stochastic Bayes estimate, by

$$\hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n) = \sum_{i=1}^{j_n} t_i f^n(t_i; \mathbf{x}_n) \Big/ \sum_{i=1}^{j_n} f^n(t_i; \mathbf{x}_n). \tag{7.1}$$

See Section 1.5 for the motivation.

Fix $\theta_0 \in \Theta$ and let m be a probability on Θ. Assume

(7.2a) m has a density g with respect to Lebesgue measure which is bounded, continuous and positive;

(7.2b) for every sequence $\{\theta_n\}$, where $n^{1/2}|\theta_n - \theta_0| \le c$ for some c, $(L_n(\theta_n + n^{-1/2}t, \theta_n; \mathbf{x}_n), tL_n(\theta_n + n^{-1/2}t, \theta_n; \mathbf{x}_n))$ converges weakly under $Q_{\theta_n}^n$, as random elements of

$$\big(L_1(R^k) \cap L_2(R^k)\big) \times \big(L_1(R^k) \cap L_2(R^k)\big),$$

to $(W(t), tW(t))$, where $W(t)$ was defined in (3.4).

THEOREM 7.1. *Suppose* $\lim_{n\to\infty} j_n n^{-d/2} = \infty$ *and* (7.2a) *and* (7.2b) *hold. Let* $\{\theta_n\}$ *be any sequence such that* $n^{1/2}|\theta_n - \theta_0| \le c$. *Then, as* $n \to \infty$, $n^{1/2}(\hat{\theta}_n - \theta_n) \Rightarrow N(0, I^{-1}(\theta_0))$ *under* $Q_{\theta_n}^n \times m^{j_n}$.

The proof requires a lemma.

CONVERGENCE LEMMA 7.1. *Let* $\{h_n\}$ *be a sequence of real functions such that* $h_n \to h$ *in* $L_1(R^k) \cap L_2(R^k)$. *Let* $Z_1, \dots, Z_{j_n}$ *be i.i.d. m. Suppose* (7.2a) *holds and* $\lim_{n\to\infty} j_n n^{-d/2} = \infty$. *Then*

$$(7.3)\quad n^{k/2}\left[\frac{\Sigma_1^{j_n} h_n\big((Z_i-\theta_n)\sqrt{n}\big)}{n}\right] \to g(\theta_0)\int h(t)\,dt, \quad \textit{in } m^{j_n} \textit{ probability}.$$

PROOF. The expectation of the quantity in question converges to the desired limit, while the variance of this quantity goes to 0; this last convergence uses $j_n n^{-d/2} \to \infty$. □

PROOF OF THEOREM 7.1. Let $W_n(t) = L_n(\theta_n + n^{-1/2}t, \theta_n)$ and $\tilde{W}_n(t) = tW_n(t)$. Because the Q_θ's are mutually absolutely continuous,

$$(7.4)\qquad \hat{\theta}_n(\mathbf{x}_n, \mathbf{t}_n) = \frac{\Sigma_1^{j_n} t_i L_n(t_i, \theta_n)}{\Sigma_1^{j_n} L_n(t_i, \theta_n)} = \frac{\Sigma t_i L_n(\theta_n + (t_i - \theta_n), \theta_n)}{\Sigma L_n(\theta_n + (t_i - \theta_n), \theta_n)},$$

so that

$$(7.5)\qquad n^{1/2}\big(\hat{\theta}_n - \theta_n\big) = \frac{\Sigma \tilde{W}_n\big(n^{1/2}(t_i - \theta_n)\big)}{\Sigma W_n\big(n^{1/2}(t_i - \theta_n)\big)}.$$

Assumption (7.2), Lemma 7.1 and Lemma 2.2(i) immediately give

$$(7.6)\qquad n^{1/2}\big(\hat{\theta}_n - \theta_n\big) \Rightarrow \frac{\int tW(t)\,dt}{\int W(t)\,dt} = I^{-1}(\theta_0)N. \qquad \square$$

To estimate the distribution of $n^{1/2}(\hat{\theta}_n - \theta_n)$, define functionals F_n by

$$(7.7)\quad F_n(\mathbf{t}_n, \theta, \mathbf{x}_n) = \frac{\Sigma_1^{j_n} t_i L_n(t_i, \theta; \mathbf{x}_n)}{\Sigma_1^{j_n} L_n(t_i, \theta; \mathbf{x}_n)}, \qquad \mathbf{x}_n \in \mathbb{X}_n, \quad \mathbf{t}_n \in (R^d)^{j_n}.$$

Let $\mathbb{D}_n(\mathbf{t}_n, \theta)$ be the distribution of $n^{1/2}(F_n(\mathbf{t}_n, \theta, \mathbf{x}_n) - \theta)$ when $\mathbf{x}_n$ has distribution Q_θ^n and $\mathbf{t}_n, \theta$ are fixed. Define a random probability measure ν_n on $\mathbb{X}_n \times (R^d)^{j_n}/R^{k_n}$ by $\nu_n((\mathbf{x}_n, \mathbf{t}_n); \cdot) = \mathbb{D}_n(\mathbf{t}_n, \hat{\theta}_n(\mathbf{x}_n))$. Let $\nu_0 = N(0, I^{-1}(\theta_0))$. Let $\hat{\nu}_n$ be the empirical on $[\mathbb{X}_n \times (R^d)^{j_n}] \times R^{k_n}$ of a sample of size k_n from ν_n. □

THEOREM 7.2. *If* $k_n \to \infty$, $j_n n^{-d/2} \to \infty$ *and* (7.2a), (7.2b) *hold, then*

$$\rho(\hat{\nu}_n, \nu_0) \to 0, \quad \textit{in } \big[Q_{\theta_n}^n \times m^{j_n}\big] \otimes \nu_n^{k_n} \textit{ probability}.$$

PROOF. By Theorem 2.1, it suffices to show that $\rho(\nu_n, \nu_0)$ converges to 0 in $Q_{\theta_n}^n \times m^{j_n}$ probability. For this, apply Lemma 2.2(ii) instead of Lemma 2.2(i) to the arguments in the proof of Theorem 7.1; this shows that whenever $\{\theta_n'\}$ satisfies $n^{1/2}|\theta_n' - \theta_0| \le c$, then $\rho(\mathbb{D}_n(\mathbf{t}_n, \theta_n'), \nu_0) \to 0$ in m^{j_n}-probability. By Theorem 7.1, $n^{1/2}(\hat{\theta}_n - \theta_0)$ is tight under $Q_{\theta_n}^n \times m^{j_n}$. Hence, $\rho(\mathbb{D}_n(\mathbf{t}_n, \hat{\theta}_n), \nu_0) \to 0$ by Lemma 2.2(i). □

REMARK. Theorems 7.1 and 7.2 with their required rate on the search sample size, are reminiscent of the global stochastic MLE. To deal with this drawback, one can let the measure m depend on the original sample. One possibility is to use the posterior distribution of a diffuse prior for m; another is to let $\mathbf{t}_n$ be a local search sample, constructed as in Section 3 by bootstrapping a preliminary estimate of θ.

8. Stochastically normed empirical process. Let $(\Omega, \mathbb{F})$ be a measure space, and $\mathbb{V}$ a Vapnik–Červonenkis class of subsets of Ω satisfying the measurability conditions of Dudley (1978). Let $(T, \mathbb{T})$ be a measure space and ϕ a mapping of $\mathbb{V}$ to T which is 1–1 and onto. Any probability P on $\mathbb{F}$ is an element of $L_\infty(\mathbb{V})$ by identifying P with the map $V \to P(V)$. Such a probability is also an element of $L_\infty(T)$ by identifying P with the map $t \to P\{\phi^{-1}(t)\}$. If P, Q are two probabilities, then evidently the supremum norm of $P - Q$ in $L_\infty(\mathbb{V})$ is the same as the supremum norm of $P - Q$ in $L_\infty(T)$.

For convenience, call (T, ϕ) a parametrization of $\mathbb{V}$. A simple example of a parametrization is the representation of half spaces of R^k by $A(s, u) = \{x \in R^k\colon x's \le u\}$, where $s \in R^d$, $|s| = 1$ and $u \in R^1$, then $T = \{s\colon |s| = 1\} \times R^1$. Assume from now on that, for each probability P on $\mathbb{F}$, the mapping $t \to P\{\phi^{-1}(t)\}$ is $\mathbb{T}$-measurable.

Let m be a σ-finite measure on $\mathbb{T}$, and let L_m denote the Banach space of all real, essentially bounded (m) functions on T, with the ess sup norm denoted by $|\cdot|_m$. Then each probability P on $\mathbb{F}$ can be identified with an element of L_m.

Let $\mathbb{X}_n$ be the n-fold product of Ω and let $\hat{Q}_n = \hat{Q}_n(\mathbf{x}_n, \cdot)$ be the empirical measure of $\mathbf{x}_n \in \mathbb{X}_n$. If Q is a probability on $\mathbb{F}$, define the T-parametrized empirical process by

$$W_n(Q, \mathbf{x}_n, t) = n^{1/2}\left[\hat{Q}_n(\mathbf{x}_n, \phi^{-1}(t)) - Q(\phi^{-1}(t)\right], \qquad t \in T. \tag{8.1}$$

If $\mathbf{t}_n = (t_1, \dots, t_{j_n}) \in T^{j_n}$, the stochastically normed empirical process is

$$\max_{i \le j_n} |W_n(Q, \mathbf{x}_n, t_i)|. \tag{8.2}$$

If $Q^n \times m^{j_n}$ is the measure on $\mathbb{X}_n \times T^{j_n}$, then the stochastic norm is computed as the maximum over j_n sets $\phi^{-1}(t_i)$, chosen at random by taking the t_i i.i.d. m.

A bootstrap estimate of the distribution of the stochastically normal empirical process can be obtained as follows. If $\mathbf{t}_n \in T^{j_n}$, let $\xi_n(\mathbf{t}_n, Q)$ be the distribution of $\max_{i \le j_n}|W_n(Q, \mathbf{x}_n, t_i)|$ when $\mathbf{x}_n$ has distribution Q^n. Let $\nu_n(\mathbf{x}, \mathbf{t}_n; \cdot) = \xi_n(\mathbf{t}_n, \hat{Q}_n(\mathbf{x}_n, \cdot))$. Let $\hat{\nu}_n$ be the empirical distribution of a sample of size k_n from ν_n, viewed as a function on $[\mathbb{X}_n \times T^{j_n}] \times R^{k_n}$. Then $\hat{\nu}_n$ is the estimated distribution of the stochastically normal empirical process.

Fix Q_0, a probability on $(\Omega, \mathbb{F})$. Let $\{W(t), t \in T\}$ be the mean 0 Gaussian process on L_m having covariance

$$EW(t)W(s) = Q_0\{\phi^{-1}(t) \cap \phi^{-1}(s)\} - Q_0\{\phi^{-1}(t)\}Q_0\{\phi^{-1}(s)\}.$$

Let ν_0 be the distribution of $|W(\cdot)|_m$. Define a pseudometric d on probabilities by specifying $d(P, Q)$ to be the larger of $\sup\{|P(V) - Q(V)|\colon V \in \mathbb{V}\}$ and $\sup\{|P(V) - Q(V)|\colon V \in \mathbb{V} \cap \mathbb{V}\}$.

THEOREM 8.1. *Suppose* $j_n, k_n \to \infty$ *and that* $d(Q_n, Q_0) \to 0$. *Then*

(a) $\max_{i \le j_n} |W(Q_n, \mathbf{x}_n, t_i)| \Rightarrow |W|_m$ *under* $m^{j_n} \times Q_n^n$,
(b) $\rho(\hat{\nu}_n, \nu_0) \to 0$ *in* $[Q_n^n \times m^{j_n}] \otimes \nu_n^{k_n}$ *probability.*

REMARK. This theorem is related to, but different from, Theorem 3 in Beran and Millar (1986), which deals with empirical processes indexed by halfspaces. Because of the special structure of $\mathbb{V}$, there, the stochastic search was taken over only part of the parametrizing set T and it was possible to dispense with "ess sup" in the final statement. It is an open question, in general, whether the search measure m can be made dependent on the original sample so as to reduce the computation needed. The proof of Theorem 8 depends on a simple lemma whose proof is omitted.

CONVERGENCE LEMMA 8.1. *Let* g_n, g *be elements of* L_m, *such that* g_n *converges to* g *in the* $\operatorname{ess\,sup}_m$ *norm. Let* $Z_1, \dots, Z_{j_n}$ *be i.i.d. m. Then, if* $j_n \to \infty$,

$$\max_{i \le j_n} |g_n(Z_i)| \to |g|_m \quad a.e.\ m^{\infty}. \tag{8.3}$$

PROOF OF THEOREM 8.1. Building on the triangular array central limit theorem for empirical processes in Le Cam (1983), one may show that $W_n(Q_n, \mathbf{x}_n, \cdot)$ converges weakly in L_m to W; this requires the hypothesis that $d(Q_n, Q_0) \to 0$. Taking g_n in the lemma to be $W_n(Q_n, \mathbf{x}_n, \cdot)$ one may now employ Lemma 2.2(i) to deduce part (a) of Theorem 8.

To prove part (b), it suffices to show $\rho(\nu_n, \nu_0) \to 0$ in $m^{j_n} \times Q_n^n$ probability, by Theorem 2.1. By a triangular array version of the general Glivenko–Cantelli theorem, $d(\hat{Q}_n, Q_0) \to 0$ in Q_n^n probability. On the other hand, if $\{Q_n'\}$ is any sequence satisfying $d(Q_n', Q_0) \to 0$, then $\rho(\xi_n(\mathbf{t}_n, Q_n'), \nu_0) \to 0$ in m^{j_n} probability by the argument for (a) and Lemma 2.2(ii). It follows that $\rho(\xi_n(\mathbf{t}_n, \hat{Q}_n(\mathbf{x}_n)), \nu_0) \to 0$ in $Q_n^n \times m^{j_n}$ probability by Lemma 2.2(i). □

Confidence bands. Let $\hat{r}_n$ be a $(1-\alpha)$th quantile of $\hat{\nu}_n$. Theorem 8.1(b) and the fact that ν_0 has a continuous distribution [cf. Beran and Millar (1986)] imply that $\{Q: \max_{i \le j_n} |\hat{Q}_n - Q| n^{1/2} \le \hat{r}_n\}$ is a confidence band with asymptotic coverage probability equal to $1 - \alpha$. Bootstrapping a (nonstochastic) sup normed empirical process with the goal of obtaining confidence bands, has been considered by Bickel and Freedman (1981), Shorack (1982) and Beran (1984), all in the special case where $\mathbb{V}$ consisted of the quadrants in R^q. Using a stochastic version of sup norm resolves computational difficulties for other choices of $\mathbb{V}$. Pyke (1984) makes a similar point in discussing tests for a simple hypothesis.

REFERENCES

ABRAMOVITCH, L. and SINGH, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Statist.* **13** 116–132.

BARNETT, V. D. (1966). Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **53** 151–165.

BERAN, R. J. (1984). Bootstrap methods in statistics. *Jahresber. Deutsch. Math.-Verein.* **86** 14–30.

BERAN, R. J. (1986). Simulated power functions. *Ann. Statist.* **14** 151–173.

BERAN, R., LE CAM, L. and MILLAR, P. W. (1987). Convergence of stochastic empirical measures. *J. Multivariate Anal.* To appear.

BERAN, R. J. and MILLAR, P. W. (1985). Asymptotic theory of confidence sets. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 865–886. Wadsworth, Monterey, Calif.

BERAN, R. J. and MILLAR, P. W. (1986). Confidence sets for a multivariate distribution. *Ann. Statist.* **14** 431–443.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

DIGGLE, P. J. and GRATTON, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. Ser. B* **46** 193–227.

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929.

DURBIN, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function.* SIAM, Philadelphia.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Statist.* **28** 181–187.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. verw. Gebiete* **14** 323–330.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

KIEFER, J. and WOLFOWITZ, J. (1958). On the deviation of the empirical distribution function of a vector chance variable. *Trans. Amer. Math. Soc.* **87** 173–186.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

LE CAM, L. (1974). Notes on asymptotic methods in statistical decision theory. Centre de Recherches Mathématiques, Univ. de Montréal.

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

MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. *Lecture Notes in Math.* **976** 75–265. Springer, New York.

MILLAR, P. W. (1984). A general approach to the optimality of minimum distance estimation. *Trans. Amer. Math. Soc.* **286** 377–418.

POLLARD, D. (1980). The minimum distance method of testing. *Metrika* **27** 43–70.

PYKE, R. (1984). Discussion of "Some limit theorems for empirical processes", by E. Giné and J. Zinn. *Ann. Probab.* **12** 996–997.

SHORACK, G. R. (1982). Bootstrapping robust regression. *Comm. Statist. A—Theory Methods* **11** 961–972.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

VARADARAJAN, V. S. (1958). Weak convergence of measures on separable metric spaces. *Sankhyā* **19** 15–22.

WICHURA, M. J. (1970). On the construction of uniformly convergent random variables with given weakly convergent image laws. *Ann. Math. Statist.* **41** 284–291.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720