THE SEQUENTIAL DESIGN OF EXPERIMENTS FOR INFINITELY MANY STATES OF NATURE[1]

By Arthur E. Albert

Columbia University[2]

0. Summary. In [2] and [3], H. Chernoff discussed the Sequential Design of Experiments. In [2], a procedure was exhibited and was proved to be asymptotically optimal for the hypothesis testing problem when there are finitely many states of nature. This paper extends Chernoff's results to infinitely many states of nature.

1. Introduction. As a rule, when a scientist performs an experiment in order to obtain information about a certain phenomenon, the outcome of the experiment not only serves to cast light on the problem at hand, but also aids the experimenter in designing a more informative experiment. As more and more data is accumulated, his experiments can be made more and more informative until he reaches a point where he feels that further experiments are unnecessary. He then announces his results.

In [2] and [3], Chernoff dealt with this procedure (which he called the "Sequential Design of Experiments") and in [2], he proposed a sequential procedure which applies to the two action (i.e., hypothesis testing) problem when there are finitely many states of nature. It was shown that the risk under this procedure is approximately $-c \log c/I(\theta)$ when the cost c per experiment is very small (where $I(\theta)$ is an appropriately defined information number). It was also shown that in order for another procedure to do appreciably better for some value of the parameter (state of nature) θ, it must do worse by an *order* of *magnitude* for some other value of the parameter (as c tends to zero).

Chernoff's procedure can be partially described by saying that at each stage, the experimenter continues experimenting so long as the likelihood ratio is less than $1/c$. If another experiment is to be performed, the experimenter chooses the experiment as though he believed that the current value of the maximum likelihood estimate (m.l.e.) were the true value of the parameter. If the likelihood ratio is so large that no more observations are required, the experimenter accepts the hypothesis corresponding to the current value of the m.l.e. (It has been pointed out by one of the referees, that the idea of estimating the true situation by the m.l.e. and then using this estimate to decide what future course of action to take, seems to date back to A. Wald's work on sequential estimation in [9].)

Received June 9, 1960; revised January 11, 1961.

[1] This work was carried out under Contract N6onr-25140 while the author was at Stanford University.

[2] Now at Communications Biophysics Laboratory, Massachusetts Institute of Technology.

We shall deal with the extension of Chernoff's procedure and results to the case where the possible states of nature are infinite in number. A *class* of procedures will be exhibited which possesses the property that for any positive number ϵ, there is a member of this class for which the risk is no larger than $-(1 + \epsilon + o(1))c \log c/I(\theta)$ as c tends to zero.

The following example will serve as a prototype for the sequential design of experiments problem as applied to hypothesis testing:

Two random variables are independent and normally distributed with means m_1 and m_2 respectively, and unit variance. It is desired to test $H_1: m_1 \geqq m_2$ vs. $H_2: m_1 < m_2$. The cost of making the wrong decision (hereafter called the "regret") is a function of the distance from the true parameter $\theta = (m_1, m_2)$ to the boundary line $\{\theta': \theta' = (m_1', m_2'), m_1' = m_2'\}$. Two experiments are available. These are e_1 : Observe the first random variable, and e_2 : Observe the second random variable. After each experiment, the statistician must decide whether to perform another (independent) experiment or to stop. If he continues, he must decide which experiment to perform next. If he stops, he must decide whether to accept H_1 or H_2.

2. The Relevance of Kullback-Leibler (K.L.) Information Numbers. In [2] and [3], extensive heuristic arguments were set forth to motivate the use of K.L. information numbers in the sequential design problem. (See [7] for a wider realm of application.) Chernoff's arguments can be briefly summarized as follows:

Suppose an experiment is repeated many times, yielding independent observations $Y_1, Y_2, \cdots, Y_n, \cdots$. Let H_1 be the hypothesis that the observations have a density $f_1(x)$ and let H_2 be the hypothesis that the observations have a density $f_2(x)$. The Bayes strategies are the Wald sequential likelihood ratio tests.

A sequential likelihood-ratio test is characterized by two numbers A and B, $(A > B)$: After the nth observation, continue sampling if

$$B < \sum_{j=1}^{n} \log [f_1(Y_j)/f_2(Y_j)] < A.$$

Stop sampling and accept H_1 if

$$\sum_{j=1}^{n} \log [f_1(Y_j)/f_2(Y_j)] \geqq A.$$

Stop sampling and accept H_2 if

$$\sum_{j=1}^{n} \log [f_1(Y_j)/f_2(Y_j)] \leqq B.$$

The appropriate numbers A and B are determined by the *a priori* probabilities and the costs. However, when c is very small, compared to the regret, it turns out that A is approximately equal to $-\log c$ and B is approximately equal to $\log c$.

Denote the probability of error (when H_i is true) by $\alpha_i (i = 1, 2)$ and the expected sample size (when H_i is true) by $N_i (i = 1, 2)$. In [10], Wald showed

that for small c, $N_1 \approx -B/I_1$, $N_2 \approx A/I_2$, $\alpha_1 \approx e^{-A}$ and $\alpha_2 \approx e^{B}$, where

$$I_1 = \int \log [f_1(y)/f_2(y)] f_1(y)\, dy$$

and

$$I_2 = \int \log [f_2(y)/f_1(y)] f_2(y)\, dy.$$

(The quantities $I_i (i = 1, 2)$ have subsequently come to be known as Kullback-Leibler information numbers.)

If the regret for making an incorrect decision (when H_i is true) is $r_i (i = 1, 2)$ then the average regret (or risk) under H_i can be approximated by

$$R_i = cN_i + r_i\alpha_i \approx [(-c \log c)/I_i]$$

when c is small compared to $r_i (i = 1, 2)$. (See [2]).

Suppose that a design element is introduced: Assume that two equally costly experiments e_1 and e_2 are available for testing H_1 against H_2. If the experimenter chooses e_j, performs it exclusively, and proceeds in an optimal fashion, his risk under H_i will be approximately inversely proportional to $I_i(e_j)$ when c is small. Hence, if $I_1(e_1) > I_1(e_2)$ and $I_2(e_1) > I_2(e_2)$, it obviously behooves the statistician to select e_1.

However, if $I_1(e_1) > I_1(e_2)$ and $I_2(e_1) < I_2(e_2)$, e_1 is better than e_2 if H_1 is true, but e_2 is better than e_1 if H_2 is true.

If the cost per experiment is small compared to the cost of making an incorrect decision, the experimenter may find it expedient to perform an additional experiment, even though he is virtually convinced that H_1 (for instance) is the true hypothesis. In this case ($I_1(e_1) > I_1(e_2)$) it would seem that he would be wisest to choose e_1.

Owing to the uncertainty about the true state of nature, the statistician is bound to make mistakes at the early stages of experimentation, but if the probability laws are such that the true hypothesis becomes more and more evident as data accumulates, the small cost of experimentation will make initial mistakes in choosing experiments relatively unimportant, and eventually the statistician will begin performing the most advantageous experiment and stick to it until he decides to make his terminal decision.

If the hypotheses are composite and if a finite number of experiments are available to the experimenter, considerations of the sort mentioned above suggest that if the experimenter is almost positive that θ is the true state of nature (say, $\theta \in H_1$), he should choose his next experiment so as to maximize $\inf_{\varphi \in H_2} I(\theta, \varphi, e)$, where

$$I(\theta, \varphi, e) = \int \log [f(y, \theta, e)/f(y, \varphi, e)] f(y, \theta, e)\, dy,$$

and $f(y, \theta, e)$ is the density of the random variable observed under the experiment e.

The appearance of an expression of the form $\max_e \min_\varphi I(\theta_0, \varphi, e)$ immediately calls to mind a resemblance to similar-looking expressions which occur in the theory of games. By interpreting $I(\theta_0, \cdot, \cdot)$ as a payoff matrix, we recall that it is sometimes possible to do a better job of maximizing $\min_\varphi I(\theta_0, \varphi, e)$ with respect to e if we utilize *randomized* strategies.

A randomized experiment can easily be interpreted when the collection of available experiments is finite (or countable). If the statistician consults a table of random numbers chooses experiment e with probability $\lambda\{e\}$ ($\sum_e \lambda\{e\} = 1$), and then performs experiment e, this process constitutes a randomized experiment which can be denoted by λ. It will be shown that a Kullback-Leibler information number for the randomized experiment λ can be consistently defined by

$$I(\theta, \varphi, \lambda) = \sum_e I(\theta, \varphi, e)\lambda\{e\}.$$

3. General formulation. We now extend the notions of the previous section to the case where the parameter space is not finite.

Suppose a statistician is contemplating two courses of action in connection with a problem of inference. The true state of nature is unknown to the statistician, but corresponds to a point in an abstract space Θ. Denoting the two (terminal) actions by a_1 and a_2 we assume that Θ can be partitioned into three sets:

$$\Theta = \Theta_0 \cup \Theta_1 \cup \Theta_2$$

If the true state of nature is in Θ_0, either action is acceptable, but if the true s.o.n. lies in Θ_i, then a_i is preferred ($i = 1, 2$). If $\theta \in \Theta_1 \cup \Theta_2$ is the true s.o.n. and the non-preferred action is taken, the regret is given by $r(\theta) > 0$. We can extend the domain of definition of r to Θ by setting $r(\theta) = 0$ for $\theta \in \Theta_0$.

The statistician has at his disposal a finite set of (pure) experiments

$$\mathcal{E} = \{e_1, e_2, \cdots, e_M\}.$$

(From now on, e with or without subscripts, will denote a generic element of $\mathcal{E}$.)

By performing a sequence of experiments, the statistician hopes to amass enough data to make an intelligent guess (or terminal decision) as to whether the true s.o.n. θ lies in Θ_1 or in Θ_2, and then will take action a_1 or a_2 accordingly. (He is not concerned if $\theta \in \Theta_0$, for then, either action is acceptable to him.)

If experiment e is performed, the random variable Y_e, which takes its values in a measure space $(\mathcal{Y}_e, \mu_e)$, is observed. It is assumed that Y_e has a density with respect to (w.r.t.)μ_e for each $\theta \in \Theta$. Hereafter, we shall denote this density $f(y, \theta, e)$.

If the $n + 1$st experiment $e^{(n+1)}$ is chosen according to any measurable rule (i.e., $e^{(n+1)} = e^{(n+1)}(Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)})$ is a measurable function of the previous n observations $Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)}$), the outcome of this experiment (once $e^{(n+1)}$ is specified) is assumed independent of the n previous outcomes. No matter which experiment is chosen, we assume a sampling cost of c units per observation.

When a *randomized* experiment λ is performed, two random variables (r.v.'s) are, in effect, being observed. First, the statistician observes the value of the r.v. E, whose probabalistic behavior is governed by the relation

$$P[E = e] = \lambda\{e\}.$$

After observing E, he observes Y_E. The probabalistic behavior of Y_E (given that $E = e$) is, of course dependent upon the true state of nature θ. But aside from that, it is known that for every (μ_e) measureable subset B of $\mathcal{Y}_e$,

$$P_\theta[Y_e \varepsilon B] = \int_B f(y, \theta, e)\, d\mu_e(y).$$

It will be convenient for us to have a notation for dealing with r.v.'s which are associated with a random experiment λ: When the (randomized) experiment λ is performed, the statistician is actually observing the values of the random variable

$$X_\lambda = (E, Y_E)$$

which takes its values in the space

$$\mathcal{X} = \{x: x = (e, y),\ e \varepsilon \mathcal{E},\ y \varepsilon \mathcal{Y}_e\}.$$

If S is a subset of $\mathcal{X}$, we define the projection of S on $\mathcal{Y}_e$ by

$$S_e = \{y: (e, y) \varepsilon S\}.$$

We define a set function μ on all sets $S \subseteq \mathcal{X}$ having the property that S_e is (μ_e) measureable for all e, by

$$\mu(S) = \sum_e \mu_e(S_e).$$

It is easy to see that the domain of μ is a σ-algebra of sets and that μ is a measure on $\mathcal{X}$. Since it has been assumed that the (pure) experiments are such that $Y^{(n+1)}$ is independent of the past given $e^{(n+1)}$, it follows that the outcome, $X^{(n+1)} = X_{\lambda^{(n+1)}}$, of the $(n + 1)$st randomized experiment is also conditionally independent of the previous n observations, once $\lambda^{(n+1)}$ is specified.
It is clear that X_λ has a density (w.r.t. μ) over $\mathcal{X}$:

$$f(x, \theta, \lambda) = f(y, \theta, e)\lambda\{e\} \quad \text{if} \quad x = (e, y) \quad \text{and} \quad y \varepsilon \mathcal{Y}_e.$$

We define the K.L. information number for the experiment λ by

$$I(\theta, \varphi, \lambda) = \int \log \frac{f(x, \theta, \lambda)}{f(x, \varphi, \lambda)}\, f(x, \theta, \lambda)\, d\mu(x).$$

It is plain then, that

$$I(\theta, \varphi, \lambda) = \sum_e I(\theta, \varphi, e)\lambda\{e\}.$$

4. An "optimal" class of procedures. We now define the class of sequential

procedures with which we shall deal. As is the custom, we shall define our procedure(s) by giving the stopping rule, the terminal decision rule and in addition, we shall prescribe a rule for choosing the next (possibly randomized) experiment if the data and stopping rule allow an additional sample.

Given the data from the first n (possibly randomized) experiments $X^{(1)}$, $X^{(2)}, \cdots, X^{(n)}$, we define $L_n(\theta)$ to be the log of the likelihood:

$$L_n(\theta) = \sum_{j=1}^{n} \log f(x^{(j)}, \theta, \lambda^{(j)}), \tag{4.1}$$

where

$$X^{(j)} = X_{\lambda^{(j)}} .$$

Let

$$L_{i,n} = \sup_{\theta' \varepsilon \Theta_i} L_n(\theta') \qquad (i = 1, 2). \tag{4.2}$$

The generalized log of the likelihood ratio can be defined by

$$L_n = \max \{L_{1n} - L_{2n}, L_{2n} - L_{1n}\}. \tag{4.3}$$

If, for each $\theta \varepsilon \Theta_1 \cup \Theta_2$, we define $a(\theta)$ to be the hypothesis alternative to the one containing θ:

$$a(\theta) = (\Theta_1 \cup \Theta_2) - \Theta_i \quad \text{if} \quad \theta \varepsilon \Theta_i, \qquad (i = 1, 2), \tag{4.4}$$

then it is easily verified that

$$L_n = \sup_{\theta' \varepsilon \Theta_1 \cup \Theta_2} \left[\inf_{\theta'' \varepsilon a(\theta')} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \theta', \lambda^{(j)})}{f(X^{(j)}, \theta'', \lambda^{(j)})} \right]. \tag{4.5}$$

For any $\rho (0 < \rho \leqq 1)$, we say that $\tilde{\theta}_n$ is a ρ-pseudo maximum likelihood estimate (ρ-p.m.l.e.) over $\Theta = \Theta_0 \cup \Theta_1 \cup \Theta_2$ if:

(1) $\tilde{\theta}_n = \tilde{\theta}_n(X^{(1)}, X^{(2)}, \cdots, X^{(n)})$ is a function of the first n observation $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$, and

$$(2) \qquad \prod_{j=1}^{n} f(X^{(j)}, \tilde{\theta}_n, \lambda^{(j)}) \geqq \rho \sup_{\theta \varepsilon \Theta} \prod_{j=1}^{n} f(X^{(j)}, \theta', \lambda^{(j)}).$$

For any $\rho (0 < \rho < 1)$, $\tilde{\theta}_n$ will always exist, and we tacitly assume that a measureable version of $\tilde{\theta}_n$ is available. If $\tilde{\theta}_n$ exists for $\rho = 1$, it corresponds to the usual m.l.e. Throughout the remainder of this paper, we assume ρ to be fixed and less than unity.

We define Λ to be the set of probability distributions over $\mathcal{E}$ and Λ_γ to be the set of probability distributions over $\mathcal{E}$ which assign at least probability $\gamma (0 \leqq \gamma \leqq 1/M)$ to each element of $\mathcal{E}$.

Given $\theta \varepsilon \Theta_1 \cup \Theta_2$ and $\gamma (0 \leqq \gamma \leqq 1/M)$, define $\lambda^{\gamma}{}_{\theta}$ to be that element of Λ_γ which maximizes $\inf_{\varphi \varepsilon a(\theta)} [I(\theta, \varphi, \lambda)]$. (Such an element exists by Theorem 2.4.2. of [1], since Λ_γ is convex and has a finite set of extreme points.)

Given $\gamma_1 (0 \leqq \gamma_1 \leqq 1/M)$, $\gamma_2 (\gamma_2 \geqq 0)$ and $c (0 < c < 1)$, we define procedure $A(\gamma_1, \gamma_2)$ as follows:

(1) On the first trial, perform any randomized experiment from Λ_{γ_1}.

(2) After the nth trial, if $L_n \geqq -(1 + \gamma_2) \log c$, stop sampling and accept the hypothesis corresponding to

$$\Theta_1 \quad \text{if} \quad L_n = L_{1n} - L_{2n}$$

$$\Theta_2 \quad \text{if} \quad L_n = L_{2n} - L_{1n}$$

(3) If, after the *nth* observation, $L_n < -(1 + \gamma_2) \log c$, compute the ρ-p.m.l.e. $\tilde{\theta}_n$, and perform the $(n + 1)$st randomized experiment with

$$\lambda^{(n+1)} = \begin{cases} \lambda_{\tilde{\theta}_n}^{\gamma_1} & \text{if} \quad \tilde{\theta}_n \varepsilon \Theta_1 \cup \Theta_2 \\ \lambda^{(n)} & \text{if} \quad \tilde{\theta}_n \varepsilon \Theta_0 . \end{cases}$$

Each member of the class depends upon four parameters: $\gamma_1 (0 \leqq \gamma_1 \leqq 1/M)$, where M is the number of pure experiments available to the statistician),

$$\gamma_2 (\gamma_2 \geqq 0), \ \rho (0 < \rho \leqq 1), \quad \text{and} \quad c (0 < c < 1).$$

Throughout the course of our discussion, ρ will remain fixed and we shall investigate the risk associated with the procedure as c approaches zero. For these reasons, we suppress c and ρ when we talk about a typical procedure "$A(\gamma_1, \gamma_2)$" from this class.

It should be pointed out that the procedure proposed by Chernoff in [2] and [3] for the case when Θ is finite, corresponds to $A(0, 0)$ with $\rho = 1$.

5. The main theorems. The most important result in this investigation is obtained via five theorems. Theorem 1 establishes the (strong) consistency of the ρ-p.m.l.e. $\tilde{\theta}_n$. The method of proof is derived from a technique employed by Wald in [8] to establish the consistency of the m.l.e. However, it was found that Wald's technique was not general enough to cope with the random variables arising from randomized experiments in the simple normal prototype example mentioned in Section 1.

The overly restrictive nature of Wald's assumptions were eventually recognized, and in [5], Kiefer and Wolfowitz were able to demonstrate the consistency of the m.l.e. under a substantial relaxation of Wald's conditions.

The assumptions utilized in the present work in order to establish the consistency of $\tilde{\theta}_n$ bear a striking resemblance to the Kiefer-Wolfowitz conditions (although the present work was done independently) and consequently, the reader is referred to [5] for a full motivation for assumptions A1-A7. Assumptions A8-a and A9 represent additional conditions governing the *rate* of convergence of $\tilde{\theta}_n$.

Theorem 2 establishes a bound on the expected sample size under a typical procedure "$A(\gamma_1, \gamma_2)$." Assumptions B1-B6 relate most directly to Theorem 2 and are used primarily in showing that this bound holds *uniformly* over large subsets of the parameter space.

Theorem 3 exhibits an upper bound on the probability of error under $A(\gamma_1, \gamma_2)$ and depends heavily on A8-b, and Theorem 4 merely combines Theorems 2 and 3, yielding a bound on the risk under $A(\gamma_1, \gamma_2)$.

Theorem 5 follows from C1 and some known theorems about convex sets. It establishes the sense in which the proposed class of procedures is optimal.

6. The assumptions. The set of assumptions A1-A7 are the generalization of Wald's assumptions. A8-a and A9 permit us to analyze the rate of convergence of $\tilde{\theta}_n$. A8-b (which is a strengthening of A8-a) is used in establishing bounds on the probability of error under $A(\gamma_1, \gamma_2)$.

A1: The space of (pure) experiments is a finite set consisting of M elements:

$$\mathcal{E} = \{e_1, e_2, \cdots, e_M\}.$$

Associated with each $e \in \mathcal{E}$ is a random variable (r.v.) Y_e, which takes its values in a measure space $(\mathcal{Y}_e, \mu_e)$. μ_e is a measure on $\mathcal{Y}_e$, and Y_e has a density $f(y, \theta, e)$ with respect to μ_e for each $\theta \in \Theta$.

A2:
$$\int_{\mathcal{Y}_e} |\log f(y, \theta, e)|\, f(y, \theta, e)\, d\mu_e(y) < \infty$$

for all $\theta \in \Theta$ and all $e \in \mathcal{E}$.

Hereafter, we will denote the expectation of a Borel function G of Y_e by $E_\theta G(Y_e)$:

$$E_\theta G(Y_e) = \int_{\mathcal{Y}_e} G(y) f(y, \theta, e)\, d\mu_e(y)$$

A3: We assume that Θ can be embedded in a compact topological space $(\Theta^*, \mathfrak{I}^*)$ where $(\Theta^*, \mathfrak{I}^*)$ is T_1, satisfies the first axiom of countability and $\Theta \subseteq \Theta^*$. (A topological space $(\Theta^*, \mathfrak{I}^*)$ is T_1 if, for every pair of points $\varphi, \varphi' \in \Theta^*$, there is a set in $\mathfrak{I}^*$, which contains φ but not φ'. The space satisfies the first countability axiom if there is a countable basis at each point. See [6] for a full discussion of these properties.)

We further assume that the domain of definition of $f(y, \theta, e)$ can be extended from Θ to Θ^* in such a way that

A4: (a). For each $e \in \mathcal{E}$, $\varphi \in \Theta^*$, $f(y, \varphi, e) \geqq 0$ (a.e.μ_e) and

$$\int_{\mathcal{Y}_e} f(y, \varphi, e)\, d\mu_e(y) \leqq 1.$$

(b). If $\theta \in \Theta$, $\varphi \in \Theta^*$ and $\varphi \neq \theta$, then

$$\int_{[f(y,\theta,e) \neq f(y,\varphi,e)]} f(y, \theta, e)\, d\mu_e(y) > 0 \quad \text{for some} \quad e \in \mathcal{E}.$$

A5: If $\varphi_i \to \varphi$ (in $\mathfrak{I}^*$), then for each $e \in \mathcal{E}$, there is a set $D = D(e, \varphi) \subseteq \mathcal{Y}_e$ (which does not depend upon the sequence $\{\varphi_i\}$), for which

$$\int_D f(y, \theta, e)\, d\mu_e(y) = 0 \quad \text{for all} \quad \theta \in \Theta$$

and for which

$$\limsup_{i \to \infty} f(y, \varphi_i, e) = f(y, \varphi, e)$$

whenever $y \notin D$ (upper semi-continuity).

DEFINITION. $w(y, U, e) = \sup_{\varphi \in U} f(y, \varphi, e)$ for each $U \in \mathfrak{I}^*$.

A6: For each $U \in \mathfrak{I}^*$, $w(y, U, e)$ is a (μ_e) measureable function of y.

A7: For each $\theta \in \Theta$ and each $\varphi \in \Theta^*$, $(\varphi \neq \theta)$, there is a set $V = V(\theta, \varphi) \in \mathfrak{I}^*$ containing φ, for which

$$E_\theta \log^+ w(Y_e, V, e) < \infty \text{ for all } e \in \mathcal{E}.$$

(For any function h, $h^+ = \max(h, 0)$.)

Let $\mathfrak{I}$ be the relativization of $\mathfrak{I}^*$ to Θ: $\mathfrak{I} = \{U: U = V \cap \Theta, V \in \mathfrak{I}^*\}$. $\mathfrak{I}$ is a topology on Θ (see [6]).

A8: (a). Given $\theta \in \Theta$ and $\varphi \in \Theta^*$ $(\theta \neq \varphi)$, there is a positive number $t = t(\theta, \varphi)$, a set $V = V(\varphi, \theta) \in \mathfrak{I}^*$ containing φ, and a set $Q = Q(\theta, \varphi) \in \mathfrak{I}$ containing θ, for which

$$E_{\theta'} [w(Y_e, V, e)/f(Y_e, \theta, e)]^t < \infty \text{ for all } e \in \mathcal{E}.$$

and all $\theta' \in Q$.

(b). Given $\theta \in \Theta$, $\varphi \in \Theta^*$ $(\varphi \neq \theta)$ and $\gamma > 0$, there is a set $V = V(\varphi, \theta, \gamma) \in \mathfrak{I}^*$, containing φ, and a set $Q = Q(\theta, \varphi, \gamma) \in \mathfrak{I}$ containing θ, for which

$$E_{\theta'} [w(Y_e, V, e)/f(Y_e, \theta, e)]^{1/1+\gamma} < \infty \text{ for all } e \in \mathcal{E},$$

and all $\theta' \in Q$.

A9: If $E_{\theta'} [w(Y_e, V, e)/f(Y, \theta', e)]^t$ exists and is finite for some $t (0 \leqq t \leqq 1)$, whenever θ' is in some set $Q \in \mathfrak{I}$, then $E_{\theta'} [w(Y_e, V, e)/f(Y_e, \theta', e)]^t$ is upper-semi continuous in θ' (w.r.t. $\mathfrak{I}$) over Q.

(Let $(\Omega, \mathcal{S})$ be a topological space and let g be a real valued function on Ω. The following statements are equivalent:

(a). g is upper-semi-continuous over Ω (in $\mathcal{S}$).

(b). For any real k, the set $\{w: g(w) \geqq k\}$ is closed (in $\mathcal{S}$).

(c). If $w_i \to w$(in $\mathcal{S}$), then $\limsup_{i \to \infty} g(w_i) \leqq g(w)$.

(d). If $w_i \to w$(in $\mathcal{S}$), then for any $\epsilon > 0$, there is an n, such that $g(w_i) \leqq g(w) + \epsilon$ for all $i \geqq n$.

An upper semi-continuous function achieves its maximum over any ($\mathcal{S}$) compact subset of Ω.)

The derivation of a bound on the expected sample size requires an additional terminology which we now develop:

(a). For any $\gamma (0 \leqq \gamma \leqq 1/M)$, Λ_γ is the collection of probability distributions over $\mathcal{E}$ which assign at least probability γ to each element of $\mathcal{E}$. Λ_γ^* is the (finite) set of extreme points of Λ_γ :

$$\Lambda_\gamma^* = \{\lambda_{1,\gamma}, \lambda_{2,\gamma}, \cdots, \lambda_{M,\gamma}\},$$

where:

$$\lambda_{i,\gamma}\{e_j\} = \begin{cases} \gamma & \text{if } i \neq j \\ 1 - (M-1)\gamma & \text{if } i = j. \end{cases}$$

$\Lambda_0 = \Lambda =$ the set of all probability distributions over $\mathcal{E}$.

(b). For $\theta \in \Theta_1 \cup \Theta_2$,

$$a(\theta) = (\Theta_1 \cup \Theta_2) - \Theta_i \quad \text{if} \quad \theta \in \Theta_i \qquad (i = 1, 2),$$

and

$$h(\theta) = \Theta_i \quad \text{if} \quad \theta \in \Theta_i \qquad (i = 1, 2).$$

(c). If $\varphi \in \Theta^*$, $y \in \mathcal{Y}_e$ and $x = (e, y)$ we define

$$f(x, \varphi, \lambda) = f(y, \varphi, e)\lambda\{e\}.$$

This is just an extension of the domain of definition of $f(x, \theta, \lambda)$ (as defined in Section 3) from Θ to Θ^*.

(d). If $e \in \mathcal{E}$, $\theta \in \Theta$ and $\varphi \in \Theta^*$,

$$I(\theta, \varphi, e) = E_\theta \log [f(Y_e, \theta, e)/f(Y_e, \varphi, e)],$$

($I(\theta, \varphi, e)$ may be $+\infty$), and if $\lambda \in \Lambda$, $\theta \in \Theta$ and $\varphi \in \Theta^*$,

$$I(\theta, \varphi, \lambda) = E_\theta \log [f(X_\lambda, \theta, \lambda)/f(X_\lambda, \varphi, \lambda)] = \sum_e \lambda\{e\} I(\theta, \varphi, e).$$

(e). If $\theta \in \Theta_1 \cup \Theta_2$, λ_θ^γ is that element of Λ_γ for which

$$\inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda_\theta^\gamma) = \max_{\lambda \in \Lambda_\gamma} [\inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda)]$$

and

$$I(\theta, \gamma) = \max_{\lambda \in \Lambda_\gamma} \inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda).$$

(f). $I(\theta) = I(\theta, 0)$

(g). $w(x, V, \lambda) = \sup_{\varphi \in V} f(x, \varphi, \lambda)$ for all $V \in \mathfrak{V}^*$.

(h). For $\theta \in \Theta$ and $V \in \mathfrak{V}^*$ and $\lambda \in \Lambda$ we define

$$\tilde{I}(\theta, V, \lambda) = E_\theta \log [f(X_\lambda, \theta, \lambda)/w(X_\lambda, V, \lambda)].$$

($\tilde{I}(\theta, V, \lambda)$ may be $\mp \infty$.)

We now state assumptions B1-B6:

B1: For each $\gamma (0 \leqq \gamma \leqq 1/M)$, $I(\theta, \gamma)$ is continuous (w.r.t. $\mathfrak{V}$) over $\Theta_1 \cup \Theta_2$.

B2: If $\theta \in \Theta_1 \cup \Theta_2$, $0 \leqq \gamma < 1/M$ and $\{\varphi_i\}$ is a sequence in Θ^* converging to φ (in $\mathfrak{V}^*$), then $\lim_{i \to \infty} I(\theta, \varphi_i, \lambda_\theta^\gamma) = I(\theta, \varphi, \lambda_\theta^\gamma)$ (The limit may be $+\infty$).

B3: If $\theta \in \Theta$ and $\tilde{I}(\theta, V, e) < \infty$, then $\tilde{I}(\theta', V, e)$ is continuous in some ($\mathfrak{V}$) neighborhood of θ.

B4: If $\theta \in \Theta$, $0 \leqq \gamma < 1/M$ and $\tilde{I}(\theta, V, e) < \infty$ for all e, then $\tilde{I}(\theta', V, \lambda_{\theta'}^\gamma)$ is continuous in some ($\mathfrak{V}$) neighborhood of θ.

B5: $I(\theta) > 0$ for all $\theta \in \Theta_1 \cup \Theta_2$.

B6: Θ_1 and Θ_2 are in $\mathfrak{V}$.

In order to establish the desired optimality property, we require

C1: If $\theta, \varphi \in \Theta$ and $I(\theta, \varphi, e) < \infty$, then

$$E_\theta \{\log [f(Y_e, \theta, e)/f(Y_e, \varphi, e)]\}^2 < \infty.$$

7. Consistency of the ρ-p.m.l.e. $\tilde{\theta}_n$. Before attempting to establish the consistency of $\tilde{\theta}_n$, we need to investigate the underlying structure associated with the problem as we have formulated it.

LEMMA 1. *If $\theta \in \Theta$, $\varphi \in \Theta^*$ and $\lambda \in \Lambda$, then $I(\theta, \varphi, \lambda) \geqq 0$ with equality if and only if $f(x, \theta, \lambda) = f(x, \varphi, \lambda)$ on a set of P_θ probability measure one.*

PROOF. For any r.v. Z, $\exp E\, Z \leqq E \exp Z$ with equality if and only if Z is constant with probability one (Jensen's inequality). The conclusion follows with $Z = \log [f(x, \varphi, \lambda)/f(x, \theta, \lambda)]$, by applying A4-a.

LEMMA 2. *Given $\theta \in \Theta$, $\varphi \in \Theta^*(\varphi \neq \theta)$, there is a decreasing sequence of sets $\{V_n\}$ such that $V_n \in \mathfrak{I}^*$ for every n, $\bigcap_{n=1}^{\infty} V_n = \{\varphi\}$, and*

$$\lim_{n\to\infty} E_\theta \log w(Y_e, V_n, e) = E_\theta \log f(Y_e, \varphi, e).$$

PROOF. Let U_n be a countable basis for $\mathfrak{I}^*$ at φ and let $V_n = \bigcap_{j=1}^{n} U_j$. Then $\{V_n\}$ is a decreasing sequence of sets in $\mathfrak{I}^*$ and φ lies in every V_n. Since $(\Theta^*, \mathfrak{I}^*)$ is a T_1 space, we have in fact, $\bigcap_{n=1}^{\infty} V_n = \{\varphi\}$. By A7, there is an n_0 such that

$$E_\theta \log w(Y_e, V_n, e) \leqq E_\theta \log w(Y_e, V_{n_0}, e) < \infty$$

for all $e \in \mathcal{E}$ and all $n \geqq n_0$. The conclusion follows by A5 and the monotone convergence theorem.

LEMMA 3. *Given $\gamma (0 < \gamma < 1/M)$, $\theta \in \Theta$ and $\varphi \in \Theta^*$ $(\varphi \neq \theta)$, there is a set $V = V(\varphi, \theta, \gamma) \in \mathfrak{I}^*$, containing φ, and a constant $\beta = \beta(\theta, \varphi, \gamma) < 0$, such that*

$$E_\theta \log \frac{w(X_\lambda, V, \lambda)}{f(X_\lambda, \theta, \lambda)} < \beta \quad \text{for all} \quad \lambda \in \Lambda_\gamma .$$

PROOF. By A4-b and Lemma 1 and 2,

$$\lim_{n\to\infty} E_\theta \log w(Y_e, V_n, e) = E_\theta \log f(Y_e, \varphi, e) \leqq E_\theta \log f(Y_e, \theta, e)$$

with strict inequality for at least one $e \in \mathcal{E}$. Since $E_\theta |\log f(Y_e, \theta, e)| < \infty$ (by A2), it follows that

$$\lim_{n\to\infty} E_\theta \log w(X_\lambda, V_n, \lambda) < E_\theta \log f(X_\lambda, \theta, \lambda)$$

for all $\lambda \in \Lambda_\gamma^*$. Since Λ_γ^* spans Λ_γ, the conclusion follows.

It should be observed that Lemma 3 is not, in general, true for $\gamma = 0$, for if $I(\theta, \varphi, e^1) = 0$ and λ places unit probability on e^1, then for all $V \in \mathfrak{I}^*$ containing φ,

$$E_\theta \log w(X_\lambda, V, \lambda) \geqq E_\theta \log f(X_\lambda, \theta, \lambda).$$

This situation actually occurs in the prototype example of Section 1. If $\theta = (m_1, m_2)$ and $\varphi = (m_1, m_2')$, then $I(\theta, \varphi, e_1) = 0$.

The following lemma permits us to apply the assumptions concerning $f(y, \theta, e)$ and $I(\theta, \varphi, e)$ directly to $f(x, \theta, \lambda)$ and $I(\theta, \varphi, \lambda)$. The proof is left to the reader.

LEMMA 4. A2, A4, A5, A6, A7, A8 *and* A9 *remain true if* $\mathcal{E}$ *is replaced by* Λ, e *by* λ, Y_e *by* X_λ, y *by* x, $\mathcal{Y}_e$ *by* $\mathfrak{X}$ *and* μ_e *by* μ (*see Section* 3) *throughout.*

The following inequality will be used in investigating the rate of convergence of $\tilde{\theta}_n$:

LEMMA 5. *For any* r.v. Z, $P[Z \geqq 0] \leqq E(e^{tZ})$ *for all* $t > 0$.

PROOF. If $E(e^{tZ}) = \infty$, then inequality is trivial. Otherwise,

$$E(e^{tZ}) \geqq E(t^{tZ}) \mid Z \geqq 0)P[Z \geqq 0].$$

Since $E(e^{tZ} \mid Z \geqq 0) \geqq 1$ when $t > 0$, the conclusion follows.

In order to establish the consistency of $\tilde{\theta}_n$, we shall show that for any set $S \varepsilon \mathfrak{Z}$ containing the true parameter θ, the probability (under θ) that $\tilde{\theta}_n \varepsilon S$ for all n sufficiently large, is unity.

DEFINITION. For any set $S \subseteq \mathfrak{Z}$, we define T_S to be the smallest integer m, such that $\tilde{\theta}_n \varepsilon S$ for all $n \geqq m$ if such an integer exists; if no such integer exists, we define $T_S = +\infty$.

We now derive a bound on

$$P_\theta [\tilde{\theta}_m \not\varepsilon S \text{ for some } m \geqq n].$$

THEOREM 1. *Let* γ, $(0 < \gamma < 1/M)$, $S \varepsilon \mathfrak{Z}$ *and* $\theta \varepsilon S$ *be given. If experiments* $\lambda^{(j)}$ *are chosen from* Λ_γ *according to any measurable procedure (i.e., such that* $\lambda^{(j)}$ *is a measurable function of the previous* $(j-1)$ *observations), then there are finite positive constants* k *and* b *and a* $(\mathfrak{Z})$ *neighborhood* Q *of* θ *for which*

$$P_{\theta'}[T_S > m] \leqq k \exp(-bm) \text{ for all } \theta' \varepsilon Q.$$

(k, b and Q depend upon θ, γ and S.)

OUTLINE OF PROOF.

$$P_{\theta'}[T_S > m] \leqq \sum_{n \geqq m} P_{\theta'}[\tilde{\theta}_n \not\varepsilon S].$$

If $\tilde{\theta}_n \not\varepsilon S$, then since $S = S^* \cap \Theta$, (where $S^* \varepsilon \mathfrak{Z}^*$), it follows that $\tilde{\theta}_n \not\varepsilon S^*$. Hence, by definition of $\tilde{\theta}_n$,

$$P_{\theta'}[\tilde{\theta}_n \not\varepsilon S] \leqq P_{\theta'}\left[\sup_{\vartheta \varepsilon \Theta^* - S^*} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \vartheta, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} \geqq \log \rho\right].$$

for all $\theta' \varepsilon S$.

By virtue of the $(\mathfrak{Z}^*)$ compactness of $\Theta^* - S^*$, Lemma 3, A8-a, A9 and the convexity of moment generating functions, we can choose a finite collection of sets $\{V_1, V_2, \cdots, V_p\} \subseteq \mathfrak{Z}^*$, a set $Q \varepsilon \mathfrak{Z}$, containing θ and positive constants b and t, such that

$$\max_{i=1 \cdots p}\left[\max_{\lambda \varepsilon \Lambda_\gamma} E_{\theta'}\left\{\exp t \log \frac{w(X_\lambda, V_i, \lambda)}{f(X_\lambda, \theta', \lambda)}\right\}\right] \leqq e^{-b}$$

and

$$P_{\theta'}[\tilde{\theta}_n \not\varepsilon S] \leqq \sum_{i=1}^{p} P_{\theta'}\left[\sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} \geqq \log \rho\right]$$

for all $\theta' \varepsilon Q$. (p, V_i, t, b and Q depend upon S, γ and θ.) We then apply Lemma 5

and obtain

$$P_{\theta'}[\tilde{\theta}_n \notin S] \leqq \sum_{i=1}^{p} \rho^{-t} e^{-bn}$$

for all $\theta' \in Q$ from which the conclusion follows.

COROLLARY. *For any* $\theta \in \Theta$, $P_\theta[\tilde{\theta}_n \to \theta] = 1$. (*The convergence is relative to* $\mathfrak{I}$.)

PROOF. $\tilde{\theta}_n \to \theta$ if and only if $T_S < \infty$ for every set $S \in \mathfrak{I}$ containing θ.

8. Bounds on the expected sample size under procedure $A(\gamma_1, \gamma_2)$. In this section, bounds on the expected sample size are derived. Lemmas 6, 7, and 8 are building blocks and Lemma 9 is the keystone of the main result of this section. The proof of Lemma 6 follows that of Lemma 2 and is left to the reader.

LEMMA 6. *If* $\theta \in \Theta$, $\varphi \in \Theta^*$ *and* $I(\theta, \varphi, e) = \infty$, *then for any constant* $C > 0$, *there is a set* $V = V(\varphi, \theta, e, C) \in \mathfrak{I}^*$ *containing* φ *for which* $\check{I}(\theta, V, e) > C$.

DEFINITION. For $\theta \in \Theta_1 \cup \Theta_2$, let $\overline{a(\theta)}$ be the ($\mathfrak{I}^*$) closure of $a(\theta)$.

LEMMA 7. *If* $\theta \in \Theta_1 \cup \Theta_2$, $0 < \gamma < 1/M$ *and* $\varphi \in \overline{a(\theta)}$ *then* $I(\theta, \varphi, \lambda_\theta^\gamma) \geqq I(\theta, \gamma)$.

PROOF. If $I(\theta, \varphi, \lambda_\theta^\gamma)$ is infinite, the assertion is trivial. Otherwise, there is a sequence $\{\varphi_i\} \subseteq a(\theta)$ converging to φ. Let $\delta > 0$ be given. By assumption B2 we may choose i so that $I(\theta, \varphi_i, \lambda_\theta^\gamma) \leqq I(\theta, \varphi, \lambda_\theta^\gamma) + \delta$. Since

$$\varphi_i \in a(\theta), I(\theta, \varphi_i, \lambda_\theta^\gamma) \geqq I(\theta, \gamma),$$

and since δ is arbitrary, the conclusion follows.

LEMMA 8. *If* $\theta \in \Theta_1 \cup \Theta_2$ *and* $0 \leqq \gamma < 1/M$, $I(\theta, \gamma) > 0$.

PROOF. By B5, $I(\theta, 0) = I(\theta) = \inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda_\theta^o) > 0$. Let

$$\hat{\lambda}\{e\} = (1 - M\gamma)\lambda_\theta^o\{e\} + \gamma.$$

Then, $\hat{\lambda} \in \Lambda_\gamma$ and $\lambda_\theta^o\{e\} = [\hat{\lambda}\{e\} - \gamma]/[1 - M\gamma]$. Since $\inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda_\theta^o) > 0$, it follows that for some $\delta > 0$, $\inf_{\varphi \in a(\theta)} [I(\theta, \varphi, \hat{\lambda}) - \gamma \sum_e I(\theta, \varphi, e)] > \delta$. Thus, $\max_{\lambda \in \Lambda_\gamma} \inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda) \geqq \inf_{\varphi \in a(\theta)} I(\theta, \varphi, \hat{\lambda}) > \delta > 0$.

In Theorem 2, we show that the expected sample size under $A(\gamma_1, \gamma_2)$ is not much larger than $-(1 + \gamma_2) \log c / I(\theta, \gamma_1)$ when θ is the state of nature and c is small. To do so, we will show in Lemma 9 that $P_\theta[N > n]$ (where N is the sample size) declines rapidly (in fact exponentially) when n is significantly larger than $-(1 + \gamma_2) \log c / I(\theta, \gamma_1)$. The proof of Lemma 9 is complicated, so the general idea is sketched roughly below to help the reader see the forest through the trees:

The event $[N > n]$ is (by definition of the stopping rule) contained in the event

$$\left[\sup_{\theta' \in \Theta_1 \cup \Theta_2} \left\{\inf_{\varphi \in a(\theta')} \sum_{j=i}^{n} \log \frac{f(X^{(j)}, \theta', \lambda^{(j)})}{f(X^{(j)}, \varphi, \lambda^{(j)})}\right\} < -(1 + \gamma_2) \log c\right]$$

and this event is, in turn, contained in the event

$$\left[\sup_{\theta' \in \Theta_1 \cup \Theta_2} \left\{\inf_{\varphi \in a(\theta')} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \theta', \lambda^{(j)})}{f(X^{(j)}, \varphi, \lambda^{(j)})}\right\} < -(1 + \gamma_2) \log c\right].$$

If θ is the true s.o.n., then by appropriately choosing a finite number of ($\mathfrak{I}^*$) neighborhoods $V_1, \cdots, V_r$ so as to cover the ($\mathfrak{I}^*$) compact set $\overline{a(\theta)}$, we will be

able to assert that for all $\theta' \varepsilon h(\theta)$

$$P_{\theta'}[N > n] \leqq \sum_{i=1}^{r} P_{\theta'}\left[\sum_{j=1}^{n} \log \frac{f(X^{(j)}, \theta', \lambda^{(j)})}{w(X^{(j)}, V_i, \lambda^{(j)})} < -(1 + \gamma_2) \log c\right],$$

or more conveniently,

$$P_{\theta'}[N > n] \leqq \sum_{i=1}^{r} P_{\theta'}\left[\sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} > (1 + \gamma_2) \log c\right].$$

If n is larger than $-(1 + \delta)(1 + \gamma_2) \log c/I(\theta', \gamma_1)$, then

$$P_{\theta'}[N > n] \leqq \sum_{i=1}^{r} P_{\theta'}\left[\sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} + I(\theta, \gamma)/1 + \delta > 0\right].$$

The summand can be decomposed into three terms:

$$\log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} + I(\theta', \gamma_1)/1 + \delta = A_{1j} + A_{2j} + A_{3j},$$

where

$$A_{ij} = \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} + \tilde{I}(\theta', V_i, \lambda^{(j)}) - \delta I(\theta', \gamma_1)/2(1 + \delta),$$

$$A_{2j} = -\tilde{I}(\theta', V_i, \lambda^{(j)}) + \tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}),$$

$$A_{3j} = -\tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}) + I(\theta', \gamma_1)(1 - \delta/2(1 + \delta)).$$

Since

$$\tilde{I}(\theta', V_i, \lambda^{(j)}) = -E_{\theta'} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})},$$

A_{1j} has negative mean and hence, $\sum_{j=1}^{n} A_{1j}$ tends to be negative. The particular choice of $\lambda^{(j)}$ (from Λ_{γ_1}) will insure that A_{2j} grows very slowly when n is large. The neighborhoods V_1 will be taken so small that A_{3j} will be approximately $-\delta I(\theta', \gamma_1)/2(1 + \delta)$. Hence, $\sum_{j=1}^{n} (A_{1j} + A_{2j} + A_{3j})$ tends to be negative, and exceeds zero with small probability.

DEFINITION. Let N be the sample size required to reach a terminal decision under $A(\gamma_1, \gamma_2)$.

LEMMA 9. *Let θ be a point in $\Theta_1 \cup \Theta_2$ and let $\delta(0 < \delta < 1)$, $\gamma_1(0 < \gamma_1 < 1/M)$, and $\gamma_2(\gamma_2 > 0)$ be given. Then there are finite positive constants b and k and a (3) neighborhood Q of θ such that for $\theta' \varepsilon Q$ and*

$$n > -(1 + \delta)(1 + \gamma_2) \log c/I(\theta', \gamma_1), \; P_{\theta'}[N > n] \leqq ke^{-bn}$$

under $A(\gamma_1, \gamma_2)$. (k, b and Q depend upon γ_1, θ and δ.)

PROOF. For notational convenience, Q with or without subscripts will denote a generic (3) neighborhood of θ throughout the following discussion. If a result holds in a (3) neighborhood of θ and a second result holds in a second (3) neighborhood of θ, then the results are simultaneously true in the intersection of these

neighborhoods which is itself a (non-vacuous) ($\mathfrak{I}$) neighborhood of θ, and hence, no ambiguity can arise, provided that we only require a finite number of statements (each of which is true in a neighborhood of θ) to be simultaneously true in a neighborhood of θ.

As was mentioned in the introduction, given $\theta \in \Theta_1 \cup \Theta_2$,

$$(8.1)\quad P_{\theta'}[N > n] \leqq P_{\theta'}\left[\inf_{\varphi \in a(\theta)} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \theta, \lambda^{(j)})}{f(X^{(j)}, \varphi, \lambda^{(j)})} < -(1 + \gamma_2) \log c\right]$$

for all $\theta' \in h(\theta)$. (Notice that if $\theta \in \Theta_1 \cup \Theta_2$, then by B6, $h(\theta) \cap \overline{a(\theta)}$ is empty.)

For each $\varphi \in \overline{a(\theta)}$, choose $V = V(\varphi, \theta, \delta) \in \mathfrak{I}^*$ containing φ so that

$$(8.2)\quad \max_{e \in \mathcal{E}} E_{\theta'}\left[\frac{w(Y_e, V, e)}{f(Y_e, \theta', e)}\right]^t < \infty$$

for some $t = t(\theta, \varphi) > 0$ whenever θ' is in some ($\mathfrak{I}$) neighborhood $Q = Q(\theta, \varphi, \delta)$ of θ and

$$(8.3, 4)\quad \tilde{I}(\theta, V, e) > \begin{cases} I(\theta, \varphi, e) - \delta I(\theta, \gamma_1)/2(1 + \delta), & \text{if } I(\theta, \varphi, e) < \infty \\ I(\theta, \gamma_1)/\lambda_\theta^{\gamma_1}\{e\}, & \text{if } I(\theta, \varphi, e) = \infty. \end{cases}$$

((8.2) is possible by A8-a, (8.3) is possible by Lemma 2, and (8.4) is possible by Lemma 6.) Since $\overline{a(\theta)}$ is ($\mathfrak{I}^*$) closed and hence compact, there is a finite set of points $\{\varphi_1, \cdots, \varphi_r\}$ $(r = r(\theta, \delta, \gamma_1))$, for which

$$(8.5)\quad \overline{a(\theta)} \subseteq \bigcup_{i=1}^{r} V_i \text{ (where } V_i = V(\varphi_i, \theta, \delta)).$$

Hence, under $A(\gamma_1, \gamma_2)$,

$$(8.6)\quad P_{\theta'}[N > n] \leqq \sum_{i=1}^{r} P_{\theta'}[\sum_{j=1}^{n} v_{j,i}(\theta') > (1 + \gamma_2) \log c]$$

for all $\theta' \in h(\theta)$, where

$$(8.7)\quad v_{j,i}(\theta') = \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})}.$$

(Keep in mind the fact that $v_{j,i}(\theta')$ depends upon θ implicity through $V_i = V(\varphi_i, \theta, \delta)$, and that, by definition,

$$(8.8)\quad E_{\theta'}[v_{j,i}(\theta') + \tilde{I}(\theta', V_i, \lambda^{(j)}) \mid X^{(1)}, X^{(2)}, \cdots, X^{(j-1)}] = 0.$$

Now, suppose that

$$(8.9)\quad n \geqq -(1 + \gamma_2)(1 + \delta) \log c / I(\theta', \gamma_1).$$

Then

$$(8.10)\quad P_{\theta'}[N > n] \leqq \sum_{i=1}^{r} P_{\theta'}\left[\sum_{j=1}^{n} v_{j,i}(\theta') + I(\theta', \gamma_1)/(1 + \delta) > 0\right].$$

Let

$$(8.11)\quad J(\theta', \theta) = \{i: \tilde{I}(\theta', V_i, e) < \infty \text{ for all } e \in \mathcal{E}\}.$$

By B3, there is a $(\mathfrak{I})$ neighborhood Q of θ such that $\tilde{I}(\theta', V_i, e) < \infty$ if $\tilde{I}(\theta, V_i, e) < \infty$ and $\theta' \varepsilon Q$. Hence, for all $i \varepsilon J(\theta, \theta)$, and all $\theta' \varepsilon Q$,

$$P_{\theta'}\left[\sum_{j=1}^{n} (v_{j,i}(\theta') + I(\theta', \gamma_1)/(1+\delta)) > 0\right] \leqq P_{\theta'}[C_{1i} > 0] + P_{\theta'}[C_{2i} > 0] + P_{\theta'}[C_{3i} > 0], \tag{8.12}$$

where

$$C_{1i} = \sum_{j=1}^{n} [v_{j,i}(\theta') + \tilde{I}(\theta', V_i, \lambda^{(j)}) - \delta I(\theta', \gamma_1)/4(1+\delta)], \tag{8.12a}$$

$$C_{2i} = \sum_{j=1}^{n} [-\tilde{I}(\theta', V_i, \lambda^{(j)}) + \tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}) - \delta I(\theta', \gamma_1)/4(1+\delta)], \tag{8.12b}$$

and

$$C_{3i} = \sum_{j=1}^{n} [-\tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}) + I(\theta', \gamma_1) - \delta I(\theta', \gamma_1)/2(1+\delta)]. \tag{8.12c}$$

Given the first $j-1$ trials, C_{1i} has a finite moment generating function in some neighborhood of θ and a negative mean at θ. Applying Lemma 5 as in Theorem 1, we can show that there is a $(\mathfrak{I})$ neighborhood Q of θ and a positive cinstant b_1, (Q and b_1 depend upon θ and δ) such that

$$P_{\theta'}[C_{1i} > 0] \leqq e^{-b_1 n} \tag{8.13}$$

for all $\theta' \varepsilon Q$, and $i \varepsilon J(\theta, \theta)$. As a direct consequence of (8.3), (8.4) and Lemma 8,

$$\tilde{I}(\theta, V_i, \lambda_\theta^{\gamma_1}) > I(\theta, \gamma_1) - \delta I(\theta, \gamma_1)/2(1+\delta) \tag{8.14}$$

for $i \varepsilon J(\theta, \theta)$ (in fact, for all i, but we don't use this). By B1 and B4, there is a $(\mathfrak{I})$ neighborhood Q of θ (depending upon δ and γ_1) for which

$$\tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}) > I(\theta', \gamma_1) - \delta I(\theta', \gamma_1)/2(1+\delta) \tag{8.15}$$

whenever $\theta' \varepsilon Q$ and $i \varepsilon J(\theta, \theta)$. Consequently,

$$P_{\theta'}[C_{3i} > 0] = 0 \tag{8.16}$$

whenever $\theta' \varepsilon Q$.

In order to deal with the expressions $P_\theta[C_{2i} > 0]$, we recall that B3 and B4 permit us to choose a $(\mathfrak{I})$ neighborhood $Q^* \subseteq h(\theta)$ of θ, (Q^* depends upon θ and δ) for which the following statements are simultaneously true:

$$\Delta = \Delta(\theta, \delta) = \sup_{\theta' \varepsilon Q^*} \{\max_{e,\, e' \varepsilon \mathcal{E}} [\max_{i \varepsilon J(\theta, \theta)} |\tilde{I}(\theta', V_i, e) - \tilde{I}(\theta', V_i, e')|]\} < \infty, \tag{8.17}$$

$$|\tilde{I}(\theta', V_i, \lambda) - \tilde{I}(\theta'', V_i, \lambda)| < \mu_{Q^*}/8 \tag{8.18a}$$

for all $\lambda \varepsilon \Lambda_{\gamma_1}$, and

$$|\tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1}) - \tilde{I}(\theta'', V_i, \lambda_{\theta''}^{\gamma_1})| < \mu_{Q^*}/8 \tag{8.18b}$$

for all θ', $\theta'' \varepsilon Q^*$ and all $i \varepsilon J(\theta, \theta)$,
where

$$\mu_{Q^*} = \inf_{\theta' \varepsilon Q^*} \delta I(\theta', \gamma_1)/2(1 + \delta), \tag{8.19}$$

is positive by virtue of Lemma 8 and B1.

Let T_{Q^*} be as defined in Theorem 1. If $i \varepsilon J(\theta, \theta)$, then for all $\theta' \varepsilon Q^*$,

$$C_{2i} = D_{1i} + D_{2i} + D_{3i} + D_{4i} \tag{8.20}$$

where by (8.17),

$$D_{1i} \equiv \sum_{1 \leqq j \leqq T_{Q^*}} [-\tilde{I}(\theta', V_i, \lambda^{(j)}) + \tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1})] \leqq \Delta T_{Q^*}, \tag{8.21}$$

$$D_{2i} \equiv \sum_{T_{Q^*} \leqq j \leqq n} [-\tilde{I}(\theta', V_i, \lambda^{(j)}) + \tilde{I}(\tilde{\theta}_j, V_i, \lambda_{\tilde{\theta}_j}^{\gamma_1})] \leqq n\mu_{Q^*}/8 \tag{8.22}$$

(since $\lambda^{(j)} = \lambda_{\tilde{\theta}_j}^{\gamma_1}$ and $\tilde{\theta}_j \varepsilon Q^* \subseteq \Theta_1 \cup \Theta_2$ for $j > T_{Q^*}$),

$$D_{3i} \equiv \sum_{T_{Q^*} < j \leqq n} [-\tilde{I}(\tilde{\theta}_j, V_i, \lambda_{\tilde{\theta}_j}^{\gamma_1}) + \tilde{I}(\theta', V_i, \lambda_{\theta'}^{\gamma_1})] \leqq n\mu_{Q^*}/8 \tag{8.23}$$

(since θ' and $\tilde{\theta}_j$ are in Q^* for $j > T_{Q^*}$), and

$$D_{4i} \equiv -n\delta I(\theta', \gamma_1)/4(1 + \delta) \leqq -n\mu_{Q^*}/2. \tag{8.24}$$

Thus, for all $\theta' \varepsilon Q^*$, $i \varepsilon J(\theta, \theta)$,

$$P_{\theta'}[C_{2i} > 0] \leqq P_{\theta'}[T_{Q^*} > n\mu_{Q^*}/4\Delta]. \tag{8.25}$$

By Theorem 1, there are finite positive constants k_2 and b_2, and a (3) neighborhood Q of θ such that

$$P_{\theta'}[T_{Q^*} > n\mu_{Q^*}/4\Delta] \leqq k_2 \exp(-b_2 n) \tag{8.26}$$

for all $\theta' \varepsilon Q$. (k_2, b_2 and Q depend upon θ, δ and γ_1.)

Combining (8.12), (8.13), (8.16), (8.25) and (8.26), we see that there are finite positive constants b_3 and k_3 and a (3) neighborhood Q of θ, such that for $\theta' \varepsilon Q$,

$$\sum_{i \varepsilon J(\theta, \theta)} P_{\theta'}\left[\sum_{j=1}^{n} v_{j,i}(\theta') > -nI(\theta', \gamma_1)/1 + \delta\right] \leqq k_3 \exp(-b_3 n). \tag{8.27}$$

If $i \notin J(\theta, \theta)$ then $E_\theta v_{j,i}(\theta) = -\infty$ under $A(\gamma_1, \gamma_2)$ and the technique of Theorem 1 will establish that there are finite positive constants k_4 and b_4 and a (3) neighborhood Q of θ for which

$$\sum_{i \notin J(\theta, \theta)} P_{\theta'}\left[\sum_{j=1}^{n} v_{j,i}(\theta') > -nI(\theta', \gamma_1)/1 + \delta\right] \leqq k_4 \exp(-b_4 n) \tag{8.28}$$

whenever $\theta' \varepsilon Q$. By adding (8.27) and (8.28), and comparing with (8.10) the conclusion follows. This brings us to the main result of this section:

THEOREM 2. *Let θ be a point of $\Theta_1 \cup \Theta_2$ and let $\epsilon > 0$, $\gamma_1 (0 < \gamma_1 < 1/M)$, and $\gamma_2 (\gamma_2 > 0)$, be given. Then there is a function $\xi(c) = \xi(c; \theta, \epsilon, \gamma_1)$ which (for*

fixed θ, ϵ *and* γ_1) *tends to zero as* c *approaches zero, and a* ($\mathfrak{I}$) *neighborhood* $Q = Q(\theta, \epsilon, \gamma_1)$ *of* θ, *such that for all* $\theta' \varepsilon Q$

$$E_{\theta'}(N) \leqq -(1 + \gamma_2)(1 + \epsilon + \xi(c)) \log c/I(\theta', \gamma_1)$$

under procedure $A(\gamma_1, \gamma_2)$.

PROOF. For any $n^* \geqq 1$

$$E_{\theta'}(N) \leqq n^* + \sum_{j \geqq n^*} P_{\theta'}[N > j].$$

By Lemma 10, there are finite positive constants k and b and a ($\mathfrak{I}$) neighborhood Q of θ (all depending upon θ, ϵ and γ_1) for which

$$P_{\theta'}[N > n] \leqq k \exp(-bn)$$

whenever $\theta' \varepsilon Q$ and $n \geqq n^* = -(1 + \epsilon)(1 + \gamma_2) \log c/I(\theta', \gamma_1)$. Thus, under $A(\gamma_1, \gamma_2)$

$$E_{\theta'}(N) \leqq n^* + k' \exp(-bn^*)$$

whenever $\theta' \varepsilon Q$. By Lemma 9 and B1, we can assure without loss of generality, that Q is chosen so that $I(\theta', \gamma_1)$ is bounded and bounded away from zero in Q. The desired result follows with $\xi(c) = -k''c^{b''}/\log c$, where k'' and b'' are appropriately defined positive constants.

9. Bounds on the probability of error under $A(\gamma_1, \gamma_2)$. In this section we show that for $0 < \gamma_1 < 1/M$ and $\gamma_2 > 0$, the probability of error under $A(\gamma_1, \gamma_2)$ is $O(c)$ (i.e., less than or equal to a constant multiple of c). The essential idea is sketched in

THEOREM 3. *If* γ_1, $(0 < \gamma_1 < 1/M)$, *and* γ_2, $(\gamma_2 > 0)$, *are given and* θ *is a point* $\Theta_1 \cup \Theta_2$, *then there is a constant* W *and a* ($\mathfrak{I}$) *neighborhood* Q *of* θ *for which the probability of error under* $A(\gamma_1, \gamma_2)$ *is*

$$\alpha(\theta') \leqq Wc$$

uniformly for $\theta' \varepsilon Q$. (W and Q depend upon γ_1, γ_2 and θ.)

OUTLINE OF PROOF. Let θ be the true state of nature. By B6, θ is an interior point of its hypothesis and hence, there is a set $A^* \varepsilon \mathfrak{I}^*$ such that

$$\theta \varepsilon A^* \cap \Theta \subseteq h(\theta).$$

If an error is committed on the nth trial, then

$$\sup_{\vartheta \varepsilon \Theta^* - A^*} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \vartheta, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} \geqq -(1 + \gamma_2) \log c$$

for all $\theta' \varepsilon A^* \cap \Theta$.

If we properly combine Lemmas 2 and 4 with A4 and A8-b, and use the compactness of $\Theta^* - A^*$, it is possible to pick a finite collection of sets $\{V_1, V_2, \cdots, V_p\} \subseteq \mathfrak{I}^*$ having the properties that

$$\sup_{\vartheta \,\varepsilon\, \Theta^* - A^*} \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \vartheta, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} \leqq \max_{i=1\cdots p} \sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})}$$

for all $\theta' \,\varepsilon\, A^* \cap \Theta$, and

$$E_{\theta'}\left[\exp\,(1 + \gamma_2)^{-1} \sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})}\right] < g^n$$

(where $0 \leqq g < 1$) for all θ' in some (3) neighborhood of θ.

We now can apply Lemma 5 and obtain

$$\alpha(\theta') \leqq \sum_{n=1}^{\infty} \sum_{i=1}^{p} P_{\theta'}\left[\sum_{j=1}^{n} \log \frac{w(X^{(j)}, V_i, \lambda^{(j)})}{f(X^{(j)}, \theta', \lambda^{(j)})} \geqq -(1 + \gamma_2) \log c\right]$$

$$\leqq pc \sum_{n=1}^{\infty} g^n$$

for all θ' in some (3) neighborhood Q of θ.

10. Bounds on the risk under $A(\gamma_1, \gamma_2)$. At this point, we are almost ready to combine Theorems 2 and 3. However, a lemma concerning the relationship between $I(\theta, \gamma)$ and $I(\theta)$ is required:

LEMMA 12. *Let K be any (3) compact subset of $\Theta_1 \cup \Theta_2$. Then, $\lim_{\gamma \to 0} I(\theta, \gamma) = I(\theta)$ uniformly for $\theta \,\varepsilon\, K$.*

PROOF. By definition, $I(\theta, \gamma) \leqq I(\theta) = I(\theta, 0)$, for all γ. Let $\lambda^*\{e\} = (1 - M\gamma)\lambda_\theta^0\{e\} + \gamma$. ($\lambda^* \,\varepsilon\, \Lambda_\gamma$.) Then

$$I(\theta, \varphi, \lambda^*) = (1 - M\gamma)I(\theta, \varphi, \lambda_\theta^0) + \gamma \sum_{\theta} I(\theta, \varphi, e) \geqq (1 - M\gamma)I(\theta, \varphi, \lambda_\theta^0).$$

But then,

$$I(\theta, \gamma) \geqq \inf_{\varphi \,\varepsilon\, a(\theta)} I(\theta, \varphi, \lambda^*) \geqq (1 - M\gamma) \inf_{\varphi \,\varepsilon\, a(\theta)} I(\theta, \varphi, \lambda_\theta^0) = (1 - M\gamma)I(\theta, 0)$$

So,

$$I(\theta, 0) \geqq \lim_{\gamma \to 0} I(\theta, \gamma) \geqq I(\theta, 0).$$

Since the convergence is monotonic and since $I(\theta, \gamma)$ and $I(\theta, 0)$ are (3) continuous, the convergence is uniform on (3) compacta.

THEOREM 4. *Let K be a (3) compact subset of $\Theta_1 \cup \Theta_2$, over which the regret function is bounded. If $\epsilon(\epsilon > 0)$ is given, then for sufficiently small $\gamma_1(\gamma_1 > 0)$, and $\gamma_2(\gamma_2 > 0)$, the risk under $A(\gamma_1, \gamma_2)$ is*

$$R(\theta) \leqq -(1 + \epsilon + \xi^*(c))c \log c/I(\theta)$$

for all $\theta \,\varepsilon\, K$. (Here, $\xi^*(c)$ depends upon K, ϵ and c, but tends to zero as c approaches zero).

PROOF. Let $\theta \,\varepsilon\, K$ be given and let $\delta = \epsilon/\epsilon + 2$. Choose γ_2 so that $0 < \gamma_2 < (\delta/2)(1 + \delta/2)^{-1}$. By Theorems 2 and 3 there is a (3) neighborhood $Q(\theta)$ of θ such that

$$E_{\theta'}(N) \leqq -(1 + \gamma_2)(1 + \delta/2 + \xi(c; \theta, \gamma_1)) \log c/I(\theta', \gamma_1),$$

and $\alpha(\theta') \leqq Wc$ for all $\theta' \varepsilon Q(\theta)$. The risk is, by definition,

$$R(\theta') \leqq cE_{\theta'}(N) + r_K\alpha(\theta')$$

for all $\theta' \varepsilon K$ (where r_K is the upper bound for the regret over K). Combining Theorems 2 and 3

$$R(\theta') \leqq -(1 + \delta + \xi'(c; \theta, \gamma, \epsilon))c \log c/I(\theta', \gamma_1)$$

for $\theta' \varepsilon Q(\theta)$, (where $\xi'(c; \theta, \gamma_1, \epsilon)$ approaches zero as $c \to 0$).

Since K is compact, a finite number of neighborhoods $Q(\theta)$ cover K: $K \subseteq \bigcup_{i=1}^{s} Q(\theta_i)$. Let

$$\hat{\xi}(c, K, \epsilon, \gamma_1) = \max_{i=1\ldots s} \xi'(c, \theta_i, \gamma_1, \epsilon).$$

For all $\theta \varepsilon K$,

$$R(\theta) \leqq -(1 + \delta + \hat{\xi}(c; K, \epsilon, \gamma_1))c \log c/I(\theta, \gamma_1).$$

By Lemma 12, we can pick γ_1 so small that $I(\theta, \gamma_1) > (1 + \delta)I(\theta)$ for all $\theta \varepsilon K$. Let

$$\xi^*(c) = \hat{\xi}(c)/1 - \delta$$

and the conclusion follows.

11. Comparison with other procedures. This section will serve to establish the optimality of the class of procedures $\{A(\gamma_1, \gamma_2)\}$ in the following sense: If a procedure B has risk $R(\theta') < (1 + o(1))c \log c/I(\theta')$ for some $\theta' \varepsilon \Theta_1 \cup \Theta_2$, then for some other $\theta'' \varepsilon \Theta_1 \cup \Theta_2$, $R(\theta'')$ is of a greater order of magnitude than $-c \log c$:

$$(\text{i.e., } \limsup_{c \to 0} R(\theta'')/(-c \log c) = \infty).$$

Thus, the risk under procedure B is greater (by an order of magnitude) for some $\theta'' \varepsilon \Theta_1 \cup \Theta_2$ if it is significantly smaller than that of $A(\gamma_1, \gamma_2)$ for any other θ'.

Three preliminary lemmas are required. The first is a theorem about convex sets:

LEMMA 13. *Given $\theta \varepsilon \Theta_1 \cup \Theta_2$ and $\delta > 0$, there is a finite set $\Phi_\delta = \Phi_\delta(\theta) \subseteq a(\theta)$ having the property that*

$$\max_{e \varepsilon \mathcal{E}} I(\theta, \varphi, e) < \infty, \qquad \text{for all } \varphi \varepsilon \Phi_\delta$$

and

$$\max_{\lambda \varepsilon \Lambda} \min_{\varphi \varepsilon \Phi_\delta} I(\theta, \varphi, \lambda) \leqq I(\theta) + \delta/2.$$

PROOF. Let $\mathcal{S}_\theta = \{s: s = (I(\theta, \varphi, e_1), \cdots, I(\theta, \varphi, e_M)), \varphi \varepsilon a(\theta)\}$, let $\mathcal{S}_\theta^*$ be the convex hull of $\mathcal{S}_\theta$ and let $\bar{\mathcal{S}}_\theta^*$ be the closure (in M dimensional Euclidean space) of $\mathcal{S}_\theta^*$. Define a function m on $\Lambda \times \bar{\mathcal{S}}_\theta^*$ by: $m(\lambda, \mathbf{s}) = \sum_{i=1}^{M} s_i\lambda\{e_i\}$ (where,

of course, $\mathbf{s} = (s_1, \cdots, s_M)$). By Theorem 2.2.7 of [1], $\max_{\lambda \in \Lambda} m(\lambda, \mathbf{s}) = \max_{1 \leqq i \leqq M} s_i$ is continuous and convex on $\bar{\mathcal{S}}_\theta^*$. Since $I(\theta, \varphi, e) \geqq 0$ for all $\varphi \in a(\theta)$, $\bar{\mathcal{S}}_\theta^*$ is a closed subset of the positive orthant and $\max_{\lambda \in \Lambda} m(\lambda, \mathbf{s}) \geqq 0$ for all $\mathbf{s} \in \bar{\mathcal{S}}_\theta^*$. It follows that $\max_{\lambda \in \Lambda} m(\lambda, s)$ achieves its minimum on $\bar{\mathcal{S}}_\theta^*$ (say at $\bar{\mathbf{s}} = (\bar{s}_1, \cdots, \bar{s}_M)$). Let

$$\hat{I}(\theta) = \min_{s \in \bar{\mathcal{S}}_\theta^*} \max_{\lambda \in \Lambda} m(\lambda, \mathbf{s}) = \max_{\lambda \in \Lambda} m(\lambda, \bar{\mathbf{s}}).$$

In particular, $\hat{I}(\theta) = \max_{1 \leqq k \leqq M} \bar{s}_k$. Since $\bar{\mathbf{s}}$ is a point of closure of $\mathcal{S}_\theta^*$, there is a point $\mathbf{s}^* = (s_1^*, \cdots, s_M^*)$ in $\mathcal{S}_\theta^*$, such that $\max_{1 \leqq k \leqq M} s_k^* \leqq \max_{1 \leqq k \leqq M} \bar{s}_k + \delta/2 = \hat{I}(\theta) + \delta/2$. By Theorem 2.2.2. of [1], $\mathbf{s}^*$ is a convex combination of a set of $M' \leqq M + 1$ points $\mathbf{s}^{(i)}$ of $\mathcal{S}_\theta$:

$$\mathbf{s}^* = \sum_{i=1}^{M'} a_i \mathbf{s}^{(i)} \quad \text{where} \quad a_i > 0, \qquad \sum_{i=1}^{M'} a_i = 1.$$

Let $\varphi^{(i)}$ be such that

$$\mathbf{s}^{(i)} = (I(\theta, \varphi^{(i)}, e_1), \cdots I(\theta, \varphi^{(i)}, e_M)) \quad i = 1, 2, \cdots, M.$$

Let $\Phi_\delta = \{\varphi^{(1)}, \varphi^{(2)}, \cdots, \varphi^{(M')}\}$. Since $s_k^* \leqq \hat{I}(\theta) + \delta/2$, for $k = 1, 2, \cdots M$, it follows that

$$\min_{\varphi \in \Phi_\delta} I(\theta, \varphi, e) \leqq \hat{I}(\theta) + \delta/2, \qquad \text{for each } e,$$

so that

$$\max_{\lambda \in \Lambda} \min_{\varphi \in \Phi_\delta} I(\theta, \varphi, e) \leqq \hat{I}(\theta) + \delta/2.$$

But,

$$\max_{\lambda \in \Lambda} [\inf_{s \in \bar{\mathcal{S}}_\theta^*} m(\lambda, \mathbf{s})] \leqq \max_{\lambda \in \Lambda} [\inf_{s \in \mathcal{S}_\theta} m(\lambda, \mathbf{s})] = \max_{\lambda \in \Lambda} [\inf_{\varphi \in a(\theta)} I(\theta, \varphi, \lambda)] = I(\theta).$$

By Theorem 2.4.2. of [1],

$$\max_{\lambda \in \Lambda} [\inf_{s \in \bar{\mathcal{S}}_\theta^*} m(\lambda, \mathbf{s})] = \min_{s \in \bar{\mathcal{S}}_\theta^*} [\max_{\lambda \in \Lambda} m(\lambda, \mathbf{s})] = \hat{I}(\theta),$$

so that $\hat{I}(\theta) \leqq I(\theta)$. By B1, $I(\theta) < \infty$ for $\theta \in \Theta_1 \cup \Theta_2$, so that $\hat{I}(\theta) < \infty$ Hence,

$$\max_{e \in \mathcal{E}} I(\theta, \varphi, e) < \infty \qquad \text{for all} \qquad \varphi \in \Phi_\delta;$$

for if $I(\theta, \varphi^{(i)}, e_j) = \infty$, then $s_j^{(i)} = \infty$, contradicting the fact that

$$\max_{i \leqq k \leqq M} s_k^* = \max_{i \leqq k \leqq M} \sum_{i=1}^{M'} a_i I(\theta, \varphi^{(i)}, e_k) \leqq \hat{I}(\theta) + \delta/2.$$

This establishes the lemma, since

$$\max_{\lambda \in \Lambda} [\min_{\varphi \in \Phi_\delta} I(\theta, \varphi, e)] \leqq \hat{I}(\theta) + \delta/2 \leqq I(\theta) + \delta/2.$$

DEFINITION

$$S_n(\theta, \varphi) = \sum_{j=1}^{n} \log \frac{f(X^{(j)}, \theta, \lambda^{(j)})}{f(X^{(j)}, \varphi, \lambda^{(j)})}.$$

The next two lemmas establish the sought after optimality property. The first says that $S_N(\theta, \varphi)$ must be large with high probability if the probability of error is to be small at θ and φ. The second shows that the rate of growth of $S_n(\theta, \varphi)$ is such that n must be very large in order to make $S_n(\theta, \varphi)$ large. Together, these lemmas show that the expected sample size must be large if the probability of error is to be kept small. (Here, N is the sample size required to reach a terminal decision.)

LEMMA 14. *Suppose* $\theta \in \Theta_1 \cup \Theta_2$, $\varphi \in a(\theta)$ *and procedure* B *has the property that* $\alpha(\theta) = O(-c \log c)$ *and* $\alpha(\varphi) = O(-c \log c)$. *Then for any* $\delta (0 < \delta < 1)$,

$$P_\theta[S_N(\theta, \varphi) < -(1 - \delta) \log c] = O(-c^\delta \log c).$$

PROOF. Assume (without loss of generality) that $\theta \in \Theta_1$. Let

$$B_n = [H_1 \text{ is accepted on the } n\text{th trial}] \cap [S_N(\theta, \varphi) < -(1 - \delta) \log c].$$

Then,

$$P_\theta[S_N(\theta, \varphi) < -(1 - \delta) \log c] \leqq \sum_n P_\theta[B_n] + P_\theta[H_1 \text{ is rejected}]$$

$$\leqq \sum_n P_\theta[B_n] + O(-c \log c).$$

Since

$$P_\varphi [\text{accept } H_1] = O(-c \log c) \geqq \sum_n P_\varphi[B_n] = \sum_n \int_{B_n} \prod_{j=1}^{n} f(x^{(j)}, \varphi, \lambda^{(j)})\, d\mu(x^{(j)})$$

$$= \sum_n \int_{B_n} e^{-S_N(\theta, \varphi)} \prod_{j=1}^{n} f(x^{(j)}, \theta, \lambda^{(j)})\, d\mu(x^{(j)}) \geqq \exp (1 - \delta) \log c \sum_n P_\theta[B_n]$$

it follows that

$$P_\theta[S_N(\theta, \varphi) < -(1 - \delta) \log c] \leqq e^{-(1-\delta)\log c}\, O(-c \log c) = O(-c^\delta \log c).$$

LEMMA 15. *Given* $\theta \in \Theta_1 \cup \Theta_2$ *and* $\delta > 0$,

$$P_\theta[\max_{1 \leqq m \leqq n} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n[I(\theta) + \delta]] = O(1/n).$$

PROOF. By Lemma 13, $\min_{\varphi \in \Phi_\delta} I(\theta, \varphi, \lambda) \leqq I(\theta) + \delta/2$ for all $\lambda \in \Lambda$. Hence,

$$\min_{\varphi \in \Phi_\delta} \sum_{j=1}^{m} I(\theta, \varphi, \lambda^{(j)}) = m \left(\min_{\varphi \in \Phi_\delta} I(\theta, \varphi, \lambda^*) \right) \leqq m(I(\theta) + \delta/2)$$

for any set of (randomized experiments) $\{\lambda^{(1)}, \cdots, \lambda^{(m)}\}$ where

$$\lambda^*\{e\} = (1/m) \sum_{j=1}^{m} \lambda^{(j)}\{e\}, \qquad \text{(so that } \lambda^* \in \Lambda).$$

Let

$$Z_m^{(1)}(\varphi) = \sum_{j=1}^{m} \left[\log \frac{f(X^{(j)}, \theta, \lambda^{(j)})}{f(X^{(j)}, \varphi, \lambda^{(j)})} - I(\theta, \varphi, \lambda^{(j)} \right]$$

and

$$Z_m^{(2)}(\varphi) = \sum_{j=1}^{m} I(\theta, \varphi, \lambda^{(j)}).$$

If $n \geqq m$, $\min_{\varphi \in \Phi_\delta} Z_m^{(2)}(\varphi) \leqq n(I(\theta) + \delta/2)$, so that if

$$\min_{\varphi \in \Phi_\delta} (Z_m^{(1)}(\varphi) + Z_m^{(2)}(\varphi)) \geqq n(I(\theta) + \delta),$$

then

$$\max_{\varphi \in \Phi_\delta} Z_m^{(1)}(\varphi) > n\delta/2, \qquad \text{for } n \geqq m.$$

Since

$$S_m(\theta, \varphi) = Z_m^{(1)}(\varphi) + Z_m^{(2)}(\varphi),$$

$$P_\theta[\max_{1 \leqq m \leqq n} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n(I(\theta) + \delta)] \leqq \sum_{\varphi \in \Phi_\delta} P_\theta[\max_{1 \leqq m \leqq n} Z_m^{(1)}(\varphi) \geqq n\delta/2].$$

For each φ, $\{Z_m^{(1)}(\varphi)\}$ is a martingale sequence, so that $\{|Z_m^{(1)}(\varphi)|^2\}$ is a semi-martingale. By applying Theorem 3.2 of [4] we obtain

$$P_\theta[\max_{1 \leqq m \leqq n} |Z_m^{(1)}(\varphi)| \geqq n\delta/2] \leqq 4E_\theta |Z_n^{(1)}(\varphi)|^2/n^2\delta^2.$$

Let

$$\sigma^2(\theta, \varphi) = \max_{e \in \mathcal{E}} E_\theta \left| \log \frac{f(Y_e, \theta, e)}{f(Y_e, \varphi, e)} - I(\theta, \varphi, e) \right|^2.$$

Since $\varphi \in \Phi_\delta \subseteq a(\theta)$, we have, by Lemma 12, that $I(\theta, \varphi, e) < \infty$ for all $e \in \mathcal{E}$. By C_1, $\sigma^2(\theta, \varphi) < \infty$ for all $\varphi \in \Phi_\delta$.
Let $\sigma_\delta^2(\theta) = \sum_{\varphi \in \Phi_\delta} \sigma^2(\theta, \varphi)$.

Since

$$E_\theta |Z_n^{(1)}(\varphi)|^2 \leqq n\sigma^2(\theta, \varphi),$$

we conclude that

$$P_\theta[\max_{1 \leqq m \leqq n} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n(I(\theta) + \delta)] \leqq 4\sigma_\delta^2/n\delta^2 = O(1/n).$$

The main theorem of this section now follows readily.

THEOREM 5. *If a procedure* B *has risk* $R(\theta) = O(-c \log c)$ *for each* $\theta \in \Theta_1 \cup \Theta_2$, *then*

$$R(\theta) \geqq -(1 + o(1))c \log c/I(\theta)$$

for all $\theta \in \Theta_1 \cup \Theta_2$.

PROOF. Let $n^* = n^*(c, \delta) = -[(1 - \delta) \log c]/[I(\theta) + \delta]$, $(0 < \delta < 1)$.

$$P_\theta[N \leqq n^*] \leqq P_\theta[\max_{1 \leqq m \leqq n^*} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n^*(I(\theta) + \delta)] + P_\theta[\min_{\varphi \in \Phi_\delta} S_N(\theta, \varphi) \leqq -(1 - \delta) \log c].$$

(This is so because

$$\max_{1 \leqq m \leqq n^*} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n^*(I(\theta) + \delta)$$

whenever $N \leqq n^*$ and $\min_{\varphi \in \Phi_\delta} S_N(\theta, \varphi) \geqq -(1 - \delta) \log c$.)

Since $R(\theta) = O(-c \log c)$ for each $\theta \in \Theta_1 \cup \Theta_2$, Lemma 14 applies to each θ and each $\varphi \in a(\theta)$. (Since $R(\theta) = cE_\theta(N) + r(\theta)\alpha(\theta) = O(-c \log c)$ and since $r(\theta) > 0$ on $\Theta_1 \cup \Theta_2$, $\alpha(\theta) = O(-c \log c)$ on $\Theta_1 \cup \Theta_2$.) In particular,

$$P_\theta[\min_{\varphi \in \Phi_\delta} S_N(\theta, \varphi) \leqq -(1 - \delta) \log c] \leqq \sum_{\varphi \in \Phi_\delta} P_\theta[S_N(\theta, \varphi) \leqq -(1 - \delta) \log c] = O(-c^\delta \log c).$$

By Lemma 15

$$P_\theta[\max_{1 \leqq m \leqq n^*} \min_{\varphi \in \Phi_\delta} S_m(\theta, \varphi) \geqq n^*(I(\theta) + \delta)] = O((-\log c)^{-1}).$$

Hence,

$$E_\theta(N) \geqq n^* P_\theta[N > n^*] = -\frac{(1 - \delta)(1 + o(1)) \log c}{I(\theta) + \delta}.$$

Consequently,

$$R(\theta) \doteq cE_\theta(N) \geqq \frac{-(1 + o(1))(1 - \delta)\, c \log c}{I(\theta) + \delta}$$

for all $\delta (0 < \delta < 1)$. Hence,

$$R(\theta) \geqq -(1 + o(1))c \log c / I(\theta)$$

which was to be proved.

12. Concluding remarks. (a). The optimal properties of the class of procedures $\{A(\gamma_1, \gamma_2)\}$ have been established only for those points θ where the regret is positive. It is quite likely that these procedures are not optimal when the true state of nature lies in a region where the regret is zero. The core of the difficulty lies in the fact that for most meaningful statistical problems, $I(\theta)$ is zero on the boundary between the two hypotheses, rendering Theorem 2 virtually useless. To put it another way, when θ is a boundary point, the likelihood ratio tends to be small in magnitude, causing the expected sample size to be large.

(b). In any particular case, the choice of $(\Theta^*, \mathfrak{I}^*)$ to compactify Θ need not be unique. However, there seems to be a *natural* method of determining a suitable compactification of Θ (if one exists at all): With each point $\theta \in \Theta$, we can associate a point in a function space, $\mathfrak{F}(\theta) = (f(\cdot, \theta, e_1), \cdots, f(\cdot, \theta, e_m))$. If we denote this function space by $\mathfrak{L}$ and let $\mathfrak{L}^*$ be the set of limit points of $\mathfrak{L}$ (in the sense of almost sure convergence), it seems natural to define Θ^* so that

the domain of $\mathfrak{F}(\cdot)$ can be extended in such a way that $\mathfrak{F}(\cdot)$ now takes values in $\mathfrak{L}^*$. The topology on Θ^* will most naturally be one for which component-wise continuity for $\mathfrak{F}$ can be established.

For the prototype example, Θ is Euclidean two space and the function space $\mathfrak{L}$, consists of all functions $\mathfrak{F}(m_1, m_2) = (2\pi)^{-\frac{1}{2}}(e^{-\frac{1}{2}(x-m_1)^2}, e^{-\frac{1}{2}(y-m_2)^2})$, as m_1 and m_2 range over the real line. $\mathfrak{L}^*$ consists of all functions $\mathfrak{F}(m_1, m_2)$ as m_1 and m_2 range over the extended real line. It seems quite natural to take Θ^* to be all points of the form

$$\varphi = (m_1, m_2), \quad -\infty \leqq m_1 \leqq \infty, \quad -\infty \leqq m_2 \leqq \infty,$$

and the obvious topology on the enlarged set Θ^* will satisfy the conditions set forth in our assumptions. The relativization of $\mathfrak{I}^*$ to Θ will be the usual topology on R^2. Alternatively, we could take Θ_0^* to be all points of the form

$$\varphi = (m_1, m_2), \quad -\infty < m_1, m_2 \leqq \infty$$

and let $\mathfrak{I}_0^*$ be the relativization of $\mathfrak{I}^*$ to Θ_0^*. Again, $\mathfrak{I}_0^*$ wll satisfy the required conditions and the relativization of $\mathfrak{I}_0^*$ to Θ is also the usual Euclidean topology.

(c). Finally, a word should be said concerning the apparent complexity of the class of procedures $\{A(\gamma_1, \gamma_2)\}$. The use of the ρ-p.m.l.e. $\tilde{\theta}_n$, (with $0 < \rho < 1$), instead of the seemingly more tractable maximum likelihood estimate $\hat{\theta}_n$, is necessitated by the fact that $\hat{\theta}_n$ may not exist in Θ whereas, $\tilde{\theta}_n$ always will. (When Θ is a finite set, as was the case in [2], $\hat{\theta}_n$ will always exist in Θ and hence, it is permissible to take $\rho = 1$.)

In order to guarantee the consistency of $\tilde{\theta}_n$, it is, in general, necessary that the randomized experiments $\lambda^{(j)}$ put positive weight on each e in $\mathcal{E}$. In our prototype problem for instance, suppose that the experimental rule dictates that e_1 be performed on the first trial and e_2 be performed thereafter. Then $\hat{\theta}(\hat{m}_{1n}, \hat{m}_{2n})$ will not converge to the true parameter. To circumvent this difficulty, we require that the $\lambda^{(j)}$'s be chosen from Λ_γ.

Chernoff recognized this difficulty in [2] and he proposed a different modification of the experimental rule which allowed him to choose his experiments from the larger class Λ. However, it appears that this technique is not readily analyzable in the case where Θ is infinite.

Under the stopping rule given in [2], the probability of error is $O(c)$. When the parameter space is infinite however, this author can prove only that the probability of error is $O(c^{1/1+\gamma_2})$ for any $\gamma_2 > 0$. By modifying the stopping rule so that sampling ceases when the likelihood ratio is greater than $c^{-(1+\gamma_2)}$ (for c between zero and one) instead of c^{-1} (as in [2]) we too can attain a probability of error which is $O(c)$.

(d). It seems natural and desireable to extend the results contained here to the case where $\mathcal{E}$ is an infinite set. Such a result would have many applications to questions arising in connection with statistical inference on time series. It appears as though suitable continuity restrictions on $f(y, \theta, \cdot)$ would permit the application of the techniques employed here to establish the necessary results.

13. Acknowledgments. I wish to thank Professor Herman Chernoff for suggesting this problem and for giving his time so generously when advice was needed. I also wish to thank the Office of Naval Research for their financial support during the course of this investigation.

REFERENCES

[1] DAVID BLACKWELL AND M. A. GIRSCHICK, *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.

[2] HERMAN CHERNOFF, "Sequential design of experiments," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 755–770.

[3] HERMAN CHERNOFF, "Motivation for an approach to the sequential design of experiments," *Decision and Information Processes*, edited by R. E. Machol, McGraw-Hill Book Co., New York, 1959.

[4] J. L. DOOB, *Stochastic Processes*, John Wiley and Sons, New York, 1953.

[5] J. KIEFER AND J. WOLFOWITZ, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 887–906.

[6] JOHN L. KELLEY, *General Topology*, D. Van Nostrand Co., New York, 1955.

[7] SOLOMON KULLBACK, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.

[8] ABRAHAM WALD, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 595–601.

[9] ABRAHAM WALD, "Asymptotic minimax solutions of sequential point estimation problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1951, pp. 1–11.

[10] ABRAHAM WALD, *Sequential Analysis*, John Wiley and Sons, New York, 1947.