

## CONVERGENCE RATES OF ESTIMATORS OF A FINITE PARAMETER: HOW SMALL CAN ERROR PROBABILITIES BE?<sup>1</sup>

BY ANDREW L. RUKHIN

*Purdue University*

In this paper we investigate the rates of convergence to zero of error probabilities in the estimation problem of a finite-valued parameter. It is shown that if a consistent estimator attains Bahadur's bound for the probability of incorrect decision at some parametric point then the error probability does not tend to zero exponentially fast for some other value of the parameter. We evaluate the minimal possible rate of convergence of this probability at a fixed parametric point for all asymptotically minimax procedures, and establish a necessary and sufficient condition for any of these procedures to have a constant risk. A simple example is constructed to demonstrate the asymptotical inadmissibility of the usual maximum likelihood estimator.

**1. Introduction.** Let  $P_\theta$ ,  $\theta \in \Theta = \{1, \dots, m\}$  be a finite collection of (different) probability distributions with positive densities  $f_\theta$ , and let  $x = (x_1, \dots, x_n)$  be a random sample from a member of this family. If  $\delta = \{\delta^n = \delta^n(x), n = 1, 2, \dots\}$  is a sequence of estimators of the finite-valued parameter  $\theta$  then the error probability,  $P_\theta(\delta^n \neq \theta)$ , is an important characteristic of  $\delta$ . The rate at which this probability tends to zero for a consistent estimator  $\delta$  is of interest. Typically, this rate is exponential, so that the following quantities (asymptotic efficiencies),  $G(\theta, \delta) = -\liminf n^{-1} \log P_\theta(\delta^n \neq \theta)$ , are positive and finite.

In this paper we are concerned with how small the error probabilities  $P_\theta(\delta^n \neq \theta)$  can be, i.e., how large the quantities  $G(\theta, \delta)$  can be, for a consistent estimator  $\delta$ .

It follows from results of Bahadur (1960, 1967, 1971) and Bahadur, Gupta and Zabell (1980) that for any consistent  $\delta$

$$(1.1) \quad G(\theta, \delta) \leq \min_{\eta: \eta \neq \theta} K(\eta, \theta),$$

where  $K(\eta, \theta) = E_\eta\{f_\eta(x)/f_\theta(x)\}$  is the Kullback-Leibler information number. However equality in (1.1) cannot be attained for all  $\theta$ , since if  $\delta_m$  is asymptotically maximin, then

$$\min_\theta G(\theta, \delta_m) < \min_{\eta: \eta \neq \theta} K(\eta, \theta).$$

Indeed it is known that if for all  $\delta$

$$\min_\theta G(\theta, \delta_m) \geq \min_\theta G(\theta, \delta),$$

then

$$(1.2) \quad \min_\theta G(\theta, \delta_m) = -\max_{\eta \neq \theta} \inf_{0 < s < 1} \log E_\theta\{f_\eta(x)/f_\theta(x)\}^s$$

(see Krafft and Puri, 1974; and Renyi, 1969).

By convexity

$$\inf_{0 < s < 1} \log E_\theta\{f_\eta(x)/f_\theta(x)\}^s > -E_\eta \log\{f_\eta(x)/f_\theta(x)\} = -K(\eta, \theta).$$

This inequality also follows from the comparison of most powerful test and minimax test

---

Received September 1981; revised July 1982.

<sup>1</sup> This work was supported by National Science Foundation grant MCS-8101670.

AMS 1980 subject classifications. Primary 62F12; secondary 62F10, 62C20, 62B10.

Key words and phrases. Finite-valued parameter, error probabilities, weighted maximum likelihood estimator, maximin procedures.

of hypothesis  $f_\theta$  against alternative  $f_\eta$  (see Chernoff, 1972, page 49).

In this paper it is shown that if (1.1) reduces to an equality for some  $\theta$ , then

$$\min_{\eta: \eta \neq \theta} G(\eta, \delta) = 0.$$

This result is proved in Section 2 by means of a useful lemma generalizing (1.2).

For all  $n$  there exists unique maximin estimator  $\delta^n$ , and it has a constant risk  $P_\theta(\delta^n \neq \theta)$  (see Wald, 1950, Theorems 5.3 and 5.4). However there is a wealth of asymptotically maximin estimators, and it is of interest to find among them a procedure which minimizes the risk, i.e., maximizes  $G(\theta, \delta)$ , at some  $\theta$ .

In Section 3 we study efficiencies  $G(\theta, \delta)$  for an asymptotically maximin estimator  $\delta$ . A sharp upper bound for these quantities is established, and a method of finding an estimator, which attains this bound, is given. The latter procedure belongs to the class of weighted maximum likelihood estimators with prior probabilities depending on the sample size. We also give a necessary and sufficient condition for the constancy of risk of all maximin rules. A simple example, which demonstrates the asymptotical inadmissibility of the usual maximum likelihood estimator for the gain  $G(\theta, \delta)$ , is constructed.

**2. Lower bounds for error probabilities.** We start with the following lemma.

LEMMA 1. *Let  $c_1, \dots, c_m$  be real constants. Then for any estimator  $\delta$*

$$(2.1) \quad \min_{\theta} \{G(\theta, \delta) - c_\theta\} \leq -\max_{\eta \neq \theta} \inf_{s>0} [c_\theta + s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s].$$

*For any subset  $N$  of  $\Theta$*

$$(2.2) \quad \min_{\theta \in N} \{G(\theta, \delta) - c_\theta\} \leq -\max_{\eta \neq \theta, \eta, \theta \in N} \inf_{s>0} [c_\theta + s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s].$$

The first part of the lemma follows from (1.2) by replacing probability distributions  $P_\theta$  in Krafft and Puri (1974) by positive measures  $e^{c_\theta} P_\theta$ . By assigning prior probabilities which give zero mass to elements of  $\Theta$  outside a certain subset  $N$  we obtain (2.2).

There exist procedures for which (2.1) and (2.2) hold with equality. These asymptotically maximin estimators (for the shifted efficiencies  $G(\theta, \delta) - c_\theta$ ) can be taken to be weighted maximum likelihood estimators  $\hat{\delta}_c$ , which, ignoring ties, maximize  $\exp(nc_\theta) \prod_1^n f_\theta(x_j)$ .

It is easy to see that

$$\begin{aligned} \{P_\theta(\hat{\delta}_c \neq \theta)\}^{1/n} &\sim \max_{\eta: \eta \neq \theta} [P_\theta\{e^{nc_\eta} \prod_1^n f_\eta(x_j) > e^{nc_\theta} \prod_1^n f_\theta(x_j)\}]^{1/n} \\ &= \max_{\eta: \eta \neq \theta} \inf_{s>0} e^{s(c_\eta - c_\theta)} E_\theta \{f_\eta(x)/f_\theta(x)\}^s. \end{aligned}$$

Therefore

$$(2.3) \quad G(\theta, \hat{\delta}_c) = -\max_{\eta: \eta \neq \theta} \inf_{s>0} [s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s]$$

and equalities hold in (2.1) or (2.2).

Also, if  $c_\eta - c_\theta + E_\eta \log \{f_\eta(x)/f_\theta(x)\} > 0$ , which means that the derivative of the convex function  $s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s$  is nonnegative at  $s = 1$ , then

$$\begin{aligned} \inf_{s>0} [c_\theta + s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s] \\ = \inf_{0 < s < 1} [c_\eta + s(c_\theta - c_\eta) + \log E_\eta \{f_\theta(x)/f_\eta(x)\}^s]. \end{aligned}$$

It is easy to see that

$$(2.4) \quad \inf_{s>0} [c_\theta + s(c_\eta - c_\theta) + \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s] \leq \min(c_\theta, c_\eta),$$

and that the estimator  $\hat{\delta}_c$  is consistent if and only if for all  $\theta, \eta, \theta \neq \eta, c_\eta - c_\theta + K(\eta, \theta) > 0$ .

THEOREM 1. *Let  $\delta$  be an estimator such that  $G(\theta, \delta) = \min_{\eta: \eta \neq \theta} K(\eta, \theta)$  for some  $\theta$ . Then  $\min_{\eta: \eta \neq \theta} G(\eta, \delta) = 0$ .*

PROOF. Let

$$G(\theta, \delta) = \min_{\eta: \eta \neq \theta} K(\eta, \theta) = K(\xi, \theta),$$

and in (2.1) put  $c_\theta = K(\xi, \delta)$ ,  $c_\eta = 0$  for  $\eta \neq \theta$ . Then the left side of (2.1) is equal to

$$\min\{0, \min_{\eta: \eta \neq \theta} G(\eta, \delta)\} = \min_{\eta: \eta \neq \theta} G(\eta, \delta).$$

But because of (2.4)

$$\max_{\rho \neq \eta} \inf_{s>0} [c_\rho + s(c_\eta - c_\rho) + \log E_\rho \{f_\eta(x)/f_\rho(x)\}^s] \leq \max_{\rho \neq \eta} \min(c_\rho, c_\eta) = 0$$

and

$$\begin{aligned} & \inf_{s>0} [c_\theta + s(c_\xi - c_\theta) + \log E_\theta \{f_\xi(x)/f_\theta(x)\}^s] \\ &= K(\xi, \theta) + \inf_{0<s<1} [-sK(\xi, \theta) + \log E_\theta \{f_\xi(x)/f_\theta(x)\}^s] \\ &= K(\xi, \theta) - K(\xi, \theta) = 0. \end{aligned}$$

Thus

$$\max_{\rho \neq \eta} \inf_{s>0} [c_\rho + s(c_\eta - c_\rho) + \log E_\rho \{f_\eta(x)/f_\rho(x)\}^s] = 0$$

and because of (2.1)  $\min_{\eta: \eta \neq \theta} G(\eta, \delta) = 0$ .

REMARK. A slight modification of the proof shows that if for some positive  $\varepsilon$

$$G(\theta, \delta) = \min_{\eta: \eta \neq \theta} K(\eta, \theta) + \varepsilon = K(\xi, \theta) + \varepsilon,$$

then

$$\min_{\eta: \eta \neq \theta} G(\eta, \delta) \leq \varepsilon^2 / \text{Var}_\xi[\log \{f_\xi(x)/f_\theta(x)\}],$$

where the variance of this inequality is assumed to be finite. Thus  $\varepsilon$ -closeness to the bound (1.1) implies that the function  $G(\eta, \delta)$  is of order  $\varepsilon^2$  for some  $\eta$ .

**3. Maximin procedures and weighted maximum likelihood estimators.** In this section we are interested in the behavior of a maximin procedure  $\delta_m$ , i.e. a procedure such that

$$(3.1) \quad \min_\theta G(\theta, \delta_m) = -\max_{\theta \neq \eta} \inf_{s>0} \log E_\theta \{f_\eta(x)/f_\theta(x)\}^s = g.$$

More exactly how large can be the asymptotic efficiency  $G(\theta, \delta_m)$  under condition (3.1), and how does one evaluate  $\sup_{\delta_m} G(\theta, \delta_m)$  for a particular value of  $\theta$  and find  $\delta_m$  for which this supremum is attained?

To answer these questions we define real  $c(\eta, \theta)$  for  $\theta \neq \eta$  as the real solution of the following equation

$$\inf_{s>0} [sc(\eta, \theta) + \log E_\eta \{f_\theta(x)/f_\eta(x)\}^s] = -g.$$

We also define

$$\bar{c}_\theta = \min_{\eta: \eta \neq \theta} c(\eta, \theta).$$

**THEOREM 2.** For any  $\theta$

$$(3.2) \quad \sup_{\delta_m} G(\theta, \delta_m) = g + \bar{c}_\theta.$$

PROOF. First we show that for any maximin procedure  $\delta_m$ ,  $G(\theta, \delta_m) < g + \bar{c}_\theta$ . For any  $\eta, \eta \neq \theta$

$$\inf_{0 < s < 1} [-s\bar{c}_\theta + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] \geq \inf_{0 < s < 1} [-sc(\eta, \theta) + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] = -c_{\eta\theta} - g$$

and

$$\begin{aligned} & \inf_{0 < s < 1} [-s\bar{c}_\theta + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] \\ & \leq -\bar{c}_\theta + \inf_{0 < s < 1} [s\bar{c}_\theta + \log E_\eta \{ f_\theta(x)/f_\eta(x) \}^s] \\ & \leq -\bar{c}_\theta + \inf_{0 < s < 1} [sc(\eta, \theta) + \log E_\eta \{ f_\theta(x)/f_\eta(x) \}^s] = -\bar{c}_\theta - g. \end{aligned}$$

Combining these inequalities we see that

$$-c(\eta, \theta) - g \leq \inf_{0 < s < 1} [-s\bar{c}_\theta + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] \leq -\bar{c}_\theta - g.$$

Therefore if  $\xi$  is defined by the formula  $c(\xi, \theta) = \bar{c}_\theta$ , then

$$(3.3) \quad \begin{aligned} & \inf_{0 < s < 1} [-s\bar{c} + \log E_\theta \{ f_\xi(x)/f_\theta(x) \}^s] \\ & = \max_{\eta: \eta \neq \theta} \inf_{0 < s < 1} [-s\bar{c}_\theta + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] = -\bar{c}_\theta - g. \end{aligned}$$

Because of Lemma 1 and (3.3)

$$\begin{aligned} \min \{ G(\theta, \delta_m) - \bar{c}_\theta, G(\xi, \delta_m) \} & \leq -\inf_{s \geq 0} [\bar{c}_\theta - s\bar{c}_\theta + \log E_\theta \{ f_\xi(x)/f_\theta(x) \}^s] \\ & = -\bar{c}_\theta + \bar{c}_\theta + g = g. \end{aligned}$$

But because of (3.1),  $G(\xi, \delta_m) \geq g$ , so that  $G(\theta, \delta_m) - \bar{c}_\theta \leq g$ .

Now we exhibit a maximin procedure  $\delta$  such that for a particular value of  $\theta$ ,  $G(\theta, \delta) = g + \bar{c}_\theta$ . Namely, define  $\delta = \hat{\delta}_c$  to be the weighted maximum likelihood estimator with  $c_\theta = \bar{c}_\theta$  and  $c_\eta = 0$  for  $\eta \neq \theta$ .

Because of (2.3) and (3.3)

$$G(\theta, \delta) = -\max_{\eta: \eta \neq \theta} \inf_{s > 0} [-s\bar{c}_\theta + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] = g + \bar{c}_\theta$$

and for  $\xi \neq \theta$

$$\begin{aligned} G(\xi, \delta) & = -\max_{\eta: \eta \neq \xi} \inf_{s > 0} [sc_\eta + \log E_\xi \{ f_\eta(x)/f_\xi(x) \}^s] \\ & \geq -\max_{\eta: \eta \neq \xi} \inf_{s > 0} [sc(\xi, \eta) + \log E_\xi \{ f_\eta(x)/f_\xi(x) \}^s] = g. \end{aligned}$$

Thus  $\delta$  is a maximin procedure, for which the supremum in (3.2) is attained. Theorem 2 is proven.

**COROLLARY.** *Let*

$$\Theta_0 = \{ \theta: \inf_{s > 0} \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s = -g \text{ for some } \eta \}.$$

*Every maximin procedure  $\delta_m$  has a constant asymptotic efficiency  $G(\theta, \delta_m) \equiv g$ , if and only if  $\Theta_0 = \Theta$ .*

This result is clear since  $\bar{c}_\theta = 0$  if and only if  $\theta \in \Theta_0$ . Thus if  $\Theta_0 \neq \Theta$  any maximin procedure with constant risk is inadmissible.

Theorem 2 also provides a method which in some situations (namely when  $\Theta_0 \neq \Theta$ ) allows one to improve upon the traditional maximum likelihood estimator. To illustrate this method let us consider an example in which  $f_\theta$  is a normal density with mean  $a_\theta$  and variance 1.

Easy calculation shows that

$$-\inf_{s > 0} [sc + \log E_\theta \{ f_\eta(x)/f_\theta(x) \}^s] = \{ (a_\eta - a_\theta)^2 - 2c \}^2 (a_\eta - a_\theta)^{-2} / 8$$

for  $c < (a_\eta - a_\theta)^2/2$ . Therefore

$$g = \min_{\eta \neq \theta} (a_\eta - a_\theta)^2/8, c_{\theta\eta} = \{(a_\eta - a_\theta)^2 - 2|a_\eta - a_\theta|(2g)^{1/2}\}/2.$$

Using Theorem 2, for any maximin procedure  $\delta$

$$(3.4) \quad G(\theta, \delta) \leq g + \min_{\eta: \eta \neq \theta} (a_\eta - a_\theta)^2/2 - (2g)^{1/2} \min_{\eta: \eta \neq \theta} |a_\eta - a_\theta|,$$

and for the maximum likelihood estimator  $\hat{\delta}$

$$G(\theta, \hat{\delta}) = \min_{\eta: \eta \neq \theta} (a_\eta - a_\theta)^2/8,$$

which of course does not exceed (3.4). Moreover  $G(\theta, \hat{\delta})$  is equal to the right side of (3.4) if and only if  $\min_{\eta: \eta \neq \theta} (a_\eta - a_\theta)^2/8 = g$ , i.e. when  $\Theta_0 = \Theta$ . If  $\Theta_0 \neq \Theta$ , then for any  $\theta \notin \Theta_0$  there exists a maximin estimator  $\delta$  such that  $G(\theta, \delta) > G(\theta, \hat{\delta})$ .

In fact  $\hat{\delta}$  can be inadmissible with respect to the gain  $G(\theta, \delta)$ . Let  $m = 3, a_1 < a_2 < a_3$ , and, say,  $a_1 + a_3 < 2a_2$ . Then  $\hat{\delta} = 1$  if  $\bar{x} < (a_1 + a_2)/2, \hat{\delta} = 3$  if  $\bar{x} > (a_2 + a_3)/2, \hat{\delta} = 2$  otherwise. The weighted maximum likelihood estimator  $\delta$  has the form  $\delta = 1$  if  $\bar{x} < (3a_2 - a_3)/2, \delta = 3$  if  $\bar{x} > (a_2 + a_3)/2, \delta = 2$  otherwise. Here  $\bar{x} = \sum x_j/n$ . Simple calculation shows that

$$G(1, \delta) = (3a_2 - 2a_1 - a_3)^2/8 > (a_2 - a_1)^2/8 = G(1, \hat{\delta})$$

and for  $\theta \neq 1$

$$G(\theta, \delta) = G(\theta, \hat{\delta}) = (a_3 - a_2)^2/8,$$

which implies inadmissibility of the maximum likelihood estimator.

Since the maximum likelihood estimator is the Bayes estimator with respect to the uniform prior distribution over  $\Theta$ , it is admissible for any finite sample size. Indeed

$$P_1(\hat{\delta}^n \neq 1) = \Phi(-bn^{1/2}/2) > P_1\{\delta^n \neq 1\} = \Phi((a/2 - b)n^{1/2}),$$

$$P_2(\hat{\delta}^n \neq 2) = \Phi(-bn^{1/2}/2) + \Phi(-an^{1/2}) < P_2\{\delta^n \neq 2\} = 2\Phi(-an^{1/2}/2),$$

$$P_3(\hat{\delta}^n \neq 3) = \Phi(-an^{1/2}/2) = P_2\{\delta^n \neq 3\},$$

where  $a = a_3 - a_2, b = a_2 - a_1$  and  $\Phi$  is the standard normal distribution function. Thus  $\delta$  does not improve upon  $\hat{\delta}$  for any  $n$ . However

$$\lim \log P_2(\hat{\delta}^n \neq 2)/\log P_2(\delta^n \neq 2) = 1,$$

while

$$\lim \log P_1(\hat{\delta}^n \neq 1)/\log P_1(\delta^n \neq 1) = (2 - a/b)^2 > 1.$$

Therefore for large values of  $n, \delta$  is preferable to  $\hat{\delta}$ .

In Table 1 the efficiencies  $e_n = -n^{-1} \log P_\theta(\delta^n \neq 0)$  and  $d_n = -n^{-1} \log P_\theta(\hat{\delta}^n \neq \theta)$  of

TABLE 1  
Efficiencies  $e_n$  and  $d_n$  of estimators  $\delta$  and  $\hat{\delta}$  for different sample sizes

$n$	$\theta = 1$		$\theta = 2$		$\theta = 3$
	$e_n$	$d_n$	$e_n$	$d_n$	$e_n = d_n$
2	2.039	1.271	0.368	0.572	0.714
4	1.652	0.946	0.287	0.427	0.460
6	1.506	0.823	0.252	0.357	0.367
10	1.377	0.715	0.217	0.285	0.287
50	1.178	0.558	0.156	0.170	0.170
100	1.155	0.532	0.144	0.150	0.150
1000	1.129	0.504	0.128	0.128	0.128
$\infty$	1.125	0.500	0.125	0.125	0.125

procedures  $\delta$  and  $\hat{\delta}$  are evaluated for different sample sizes when  $a = b/2 = 1$ . In this case the estimator  $\delta$  is considerably better than  $\hat{\delta}$  at  $\theta = 1$  with their efficiencies almost equal at  $\theta = 2$  even for relatively small sample sizes.

## REFERENCES

- BAHADUR, R. R. (1960). Asymptotic efficiency of tests and estimates. *Sankhyā Ser. A* **22** 229–252.
- BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 304–324.
- BAHADUR, R. R. (1971). Some limit theorems in statistics. *Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.
- BAHADUR, R. R., GUPTA, J. C. and ZABELL, S. L. (1980). Large deviations, tests, and estimates. In *Asymptotic Theory of Statistical Tests and Estimation*, edited by I. M. Chakravarti. Academic, New York. 33–64.
- CHERNOFF, H. (1972). Sequential analysis and optimal design. *Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.
- KRAFFT, O. and PURI, M. L. (1974). The asymptotical behavior of the minimax risk for multiple decision problems. *Sankhyā Ser. A* **36** 1–12.
- RENYI, A. (1969). On some problems of statistics from the point of view of information theory. *Proceedings of the Colloquium on Information Theory*, Debrecen 343–357. Also in *Selected Papers of Alfred Renyi*, Akademiai Kiado, Budapest (1976) **3** 560–576.
- WALD, A. (1950). *Statistical Decision Functions*. Wiley, New York.

DEPARTMENT OF STATISTICS  
PURDUE UNIVERSITY  
WEST LAFAYETTE, INDIANA 47907