REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics.* Princeton, N.J.: Princeton Univ. Press, 1946.
[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1949, pp. 379–423, 623–656.
[3] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, pt. 4, Mar. 1959, pp. 142–163.
[4] T. J. Goblick, Jr., "Theoretical limitation on the transmission of information," *IEEE Trans. Inform. Theory*, vol. IT-11, Oct. 1965, pp. 558–566.
[5] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.
[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, 1951, pp. 79–86.
[7] J. Seidler, "Relationships between information theory and decision function theory," *Trans. 2nd Prague Conf.*, Prague, Czechoslovakia, 1960, pp. 579–592.
[8] D. Slepian, "Estimation of signal parameters in the presence of noise," *IRE Trans. Inform. Theory*, vol. IT-3, Mar. 1954, pp. 68–89.
[9] P. Swerling, "Maximum angular accuracy of a pulsed search radar," *Proc. IRE*, vol. 44, Sept. 1956, pp. 1146–1155.
[10] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-15, May 1969, pp. 386–391.
[11] R. M. Fano, *Transmission of Information.* New York: Wiley, 1961.
[12] C. E. Shannon, "Probability of error for optimal codes in Gaussian channels," *Bell Syst. Tech. J.*, vol. 38, 1959, pp. 611–656.

Variance of Bayes Estimates

LOUIS A. LIPORACE, MEMBER, IEEE

***Abstract*—This paper contains an analysis of the performance of Bayes conditional-mean parameter estimators. The main result is that on a finite parameter space such estimates exhibit a mean-square error that diminishes exponentially with the number of observations, the observations being assumed to be independent. Two situations are discussed: true parameter included in the parameter space and true parameter not included in the parameter space. In the former instance only very general assumptions are required to demonstrate the exponential convergence rate. In the latter case the existence of an information function must be invoked. Comments on the continuous-parameter-space realization of the estimator and a discussion of the convergence mechanism are also included.**

I. INTRODUCTION

THE TOPIC discussed in this paper is the performance of Bayes conditional-mean estimators on a finite parameter space. Specifically, the mean-square error in the Bayes estimate is bounded by a quantity that diminishes exponentially in n, the number of observations. A critical element in the variance calculation is the product form of the *a posteriori* density function on the parameter space. This quantity, moreover, brings to mind the calculation of probability of error for a finite-hypothesis detection scheme. That there are numerous such detection schemes that exhibit exponentially decreasing probability of error is a familiar fact. A very basic illustrative example is contained, for example, in [1, p. 162]. Therefore the existence of a type of estimator having exponentially decreasing variance in general is not surprising.

Interestingly, the situation in which the true parameter indexing the density function on the observable is known to be included in the parameter space can be differentiated from the situation in which the true parameter is not included. In the former instance only obvious and general assumptions are required to show that the Bayes estimate converges to the true parameter at exponential rate. In the latter instance the existence of an information criterion must be invoked, the estimator converging to that point in the parameter space at which the information function is maximized. Le Cam [2] and recently Patrick and Costello [3] have discussed the relation of the information function to Bayes estimation. In [3] the authors demonstrated the asymptotic bound $kn^{-s/2}$ on the variance of the Bayes estimate, s being the order of the largest finite moment of the information function. The coefficient k is increasing in s. Another asymptotic statement is contained in [14]. However, the arguments there are valid only for individual sequences of observables; they do not represent an average overall sequences. The distinction between the true parameter being included and excluded was not made in any of these works.

Estimation of the frequency of a sinusoid in noise [5, pp. 272–278] is an instance of engineering significance in which the use of a Bayes estimator on a finite parameter space can be employed, the number of points in the parameter space depending upon the resolution required in the frequency estimate. In this example the true parameter's being contained in the parameter space corresponds to the sinusoid's having one of a number of known carrier frequencies. The true parameter not being included in the parameter set corresponds to the passive estimation of an entirely unknown carrier frequency. The ensuing discussion proceeds as follows.

Section II contains the definition of terms and the calculation of variance of Bayes estimates on a finite parameter space. Comments on Bayes estimates on a continuous parameter space follow in Section III. Section IV contains

Manuscript received February 10, 1971; revised April 27, 1971. This work was supported in part by the U.S. Air Force, Contract F44620-70-C-0113.
The author is with the Department of Electrical Engineering, University of Southern California, Los Angeles, Calif. 90007.

a discussion of the extension of the convergence mechanism to other estimation techniques. That is, the existence of a class of estimators that on finite parameter spaces have exponentially decreasing variance is demonstrated, the Bayes estimator being one member of this class.

II. Definitions and Variance Calculation

Let x denote a random vector having probability density function $f(x;\theta_0)$ indexed by a vector of parameters θ_0. The totality of admissible values of the parameters is denoted by Θ. That is, the density function on x belongs to the family $\mathscr{F} = \{f(x;\theta)\}$, $\theta \in \Theta$. Independent observations of x are indicated as $x_1, x_2, \cdots, x_n$. For notational convenience let

$$X_n = \{x_k\}_{k=1}^n. \tag{1}$$

The parameter estimate $\theta_1(n)$ minimizing Bayes risk for a quadratic loss function (see, e.g., [4]) is the conditional mean of θ given X_n, i.e.,[1]

$$\theta_1(n) = \int_{\theta \in \Theta} \theta p(\theta \mid X_n)\, d\theta. \tag{2}$$

The quantity $p(\theta \mid X_n)$ is the *a posteriori* density function on Θ. With $x_1, \cdots, x_n$ independent this is expressed as

$$p(\theta \mid X_n) = \frac{\prod_{k=1}^n f(x_k;\theta) p_0(\theta)}{\int_{\theta \in \Theta} (\text{numerator})\, d\theta}, \tag{3}$$

where $p_0(\theta)$ is the initial density function on the parameter space. Usually the Bayes estimator is implemented by approximating the parameter space Θ by a finite space Θ_v of v points $\{\theta_j\}_{j=1}^v$ and using the discrete versions of (2) and (3). That is, on Θ_v

$$\theta(n) = \sum_{j=1}^{v} \theta_j p(\theta_j \mid X_n), \tag{4}$$

where

$$p(\theta_j \mid X_n) = \prod_{k=1}^n f(x_k;\theta_j) p_0(\theta_j) \Big/ \sum_{i=1}^{v} \prod_{t=1}^{n} f(x_t;\theta_i) p_0(\theta_i). \tag{5}$$

The convergence properties of $\theta(n)$ calculated according to (4) are specified in Theorems 1 and 2 below, Theorem 1 applying when θ_0 is contained in Θ_v, and Theorem 2 applying otherwise.

A. True Parameter Contained in Parameter Set

Assume now that the true parameter θ_0 is contained in Θ_v, say $\theta_0 = \theta_v$. Denote the mean-square error between $\theta(n)$ and θ_0 by $\sigma^2(n)$, i.e.,

$$\sigma^2(n) \equiv E\{\|\theta(n) - \theta_0\|^2\}, \tag{6}$$

where as usual

$$\|a\|^2 \equiv x^T x, \tag{7}$$

the superscript T indicating the transpose.

[1] With constraints on the family $\mathscr{F}$ the conditional-mean estimator minimizes Bayes risk for more general loss functions; see, e.g., [5].

Theorem 1: If the family $\mathscr{F}$ is identifiable,[2] i.e., if $f(x;\theta_1) = f(x;\theta_2)\,\forall x$ if and only if $\theta_1 = \theta_2$, for all $\theta_1, \theta_2 \in \Theta$, then

$$\sigma^2(n) \le c[(v-1)R]^2 \rho^n, \qquad 0 < \rho < 1,\ R,c \text{ finite}. \tag{8}$$

Proof: Expressing the norm explicitly we rewrite (6) as

$$\sigma^2(n) = \sum_{i=1}^{v-1} \sum_{j=1}^{v-1} (\theta_j - \theta_v)^T(\theta_i - \theta_v) E\{p(\theta_j \mid X_n) p(\theta_i \mid X_n)\}. \tag{9}$$

Equation (9) incorporates the fact that $\theta_v = \theta_0$ and that $p(\theta_j \mid X_n)$ sums to 1. Since $p(\theta_i \mid X_n) \le 1$, $i = 1,2,\cdots,v-1$,

$$p(\theta_j \mid X_n) p(\theta_i \mid X_n) \le p(\theta_j \mid X_n). \tag{10}$$

Letting

$$R = \max_{j \ne v} \{\|\theta_j - \theta_v\|\}$$

and noting (10), we bound (9) as

$$\sigma^2(n) \le [(v-1)R]^2 E\{p(\theta_j \mid X_n)\}, \qquad j \ne v. \tag{11}$$

The proof is completed by demonstrating that the expected value indicated in (11) is bounded by ρ^n, $0 < \rho < 1$.

Since $p(\theta_j \mid X_n) \le 1$,

$$E\{p(\theta_j \mid X_n)\} \le E\{[p(\theta_j \mid X_n)]^\lambda\}, \qquad 0 \le \lambda \le 1. \tag{12}$$

In turn, the right side of (12) can be bounded by deleting from the denominator of (5) all the products except the one indexed by $\theta_v = \theta_0$. Accordingly,

$$\begin{aligned} E\{p(\theta_j \mid X_n)\} &\le cE\left\{\prod_{k=1}^n [f(x_k;\theta_j)/f(x_k;\theta_0)]^\lambda\right\} \\ &= c\{E[f(x;\theta_j)/f(x;\theta_0)]^\lambda\}^n, \end{aligned} \tag{13}$$

where $c = [p_0(\theta_j \mid X_n)/p_0(\theta_v \mid X_n)]^\lambda$. Evaluating the expectation in (13) for $\lambda = \frac{1}{2}$ yields

$$\begin{aligned} E[f(x;\theta_j)/f(x;\theta_0)]^{1/2} &\equiv \int [f(x;\theta_j)]^{1/2}[f(x;\theta_0)]^{1/2}\, dx \\ &= \rho. \end{aligned} \tag{14}$$

By the Schwarz inequality $\rho \le 1$, with equality if and only if $\theta_j = \theta_0$. The coefficient c in (8) and (13) is 1 if the initial density function on the parameter set is uniform.

B. True Parameter Not Contained in Parameter Set

Suppose now that the true parameter θ_0 is not included in the parameter set Θ_v. In that event we expect that under suitable conditions the estimate $\theta(n)$ will converge to that point $\theta_m \in \Theta_v$ such that $f(x;\theta_m)$ is closest in some fitting sense to $f(x;\theta_0)$. Therefore, to quantify the convergence of $\theta(n)$ for the situation in which θ_0 is not contained in Θ_v involves some measure of closeness defined on the parameter set Θ. With this in mind define the information function

[2] The question of identifiability arises when $\mathscr{F}$ is a family of mixture density functions, i.e., when $f(x;\theta)$ is a convex combination of component density functions. A mixture is said to be identifiable if it can be resolved uniquely into its component density functions. This concept has been discussed by Teicher [6], [7] and characterized by Yakowitz and Spragins [8], [9].

$J(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, by

$$J(\boldsymbol{\theta}) \equiv E\{\ln f(\boldsymbol{x};\boldsymbol{\theta})\} = \int \ln [f(\boldsymbol{x};\boldsymbol{\theta})] \cdot f(\boldsymbol{x};\boldsymbol{\theta}_0)\, d\boldsymbol{x}. \tag{15}$$

That $J(\boldsymbol{\theta})$ is a proper measure of similarity is endorsed by the fact that the function is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (see, e.g., [10]). A simple proof is the following. Let $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$ be an arbitrary point in the admissible parameter space of the family $\mathscr{F}$. Then, subject to its existence,

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}_0) = \int \ln [f(\boldsymbol{x};\boldsymbol{\theta}')/f(\boldsymbol{x};\boldsymbol{\theta}_0)] f(\boldsymbol{x};\boldsymbol{\theta}_0)\, d\boldsymbol{x}. \tag{16}$$

By using the inequality $\ln a \leq a - 1$, the right side of (16) can be bounded to yield

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}_0) \leq \int [f(\boldsymbol{x};\boldsymbol{\theta}')/f(\boldsymbol{x};\boldsymbol{\theta}_0) - 1] f(\boldsymbol{x};\boldsymbol{\theta}_0)\, d\boldsymbol{x} = 0,$$

with equality if and only if $f(\boldsymbol{x};\boldsymbol{\theta}') = f(\boldsymbol{x};\boldsymbol{\theta}_0)$ almost everywhere. The Bayes estimator then converges at exponential rate to that point $\boldsymbol{\theta}_m$ such that $J(\boldsymbol{\theta}_m)$ is maximum over the finite parameter space [2], [3]. Obviously $\boldsymbol{\theta}_m = \boldsymbol{\theta}_0$ if $\boldsymbol{\theta}_0$ is included. The convergence is demonstrated as follows.

Theorem 2: If i) $J(\boldsymbol{\theta})$ exists $\forall \boldsymbol{\theta}$ for the identifiable family $\mathscr{F}$, and ii) there is an interval $\pi_T = (0,T] \subset (0,1]$ such that $E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_i)]^t\}$ exists for all $t \in \pi_T$, $\boldsymbol{\theta} \in \Theta_v$, then

$$s^2(n) \equiv E\{\|\boldsymbol{\theta}(n) - \boldsymbol{\theta}_m\|^2\} \leq c[(v-1)R]^2 \eta^n, \qquad 0 < \eta < 1, \tag{17}$$

where

$$J(\boldsymbol{\theta}_m) = \max_{\boldsymbol{\theta} \in \Theta_v} \{J(\boldsymbol{\theta})\}. \tag{18}$$

Proof: The steps in the proof of Theorem 1 through (11) hold here as well. Only an alternative expression for $E\{p(\boldsymbol{\theta}_j \mid \boldsymbol{X}_n)\}$ is needed. Again, bounding $p(\boldsymbol{\theta} \mid \boldsymbol{X}_n)$ by $[p(\boldsymbol{\theta}_j \mid \boldsymbol{X}_n)]^t$, $t \in \pi_T$, and deleting from the denominator of (15) all the products except the one indexed by $\boldsymbol{\theta}_m$ provides the bound

$$E\{p(\boldsymbol{\theta}_j \mid \boldsymbol{X}_n)\} < [E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t\}]^n. \tag{19}$$

The argument is completed by demonstrating that $J(\boldsymbol{\theta}_m) > J(\boldsymbol{\theta}_j)$ implies that $E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t\} < 1$.

Note that

$$\begin{aligned} J(\boldsymbol{\theta}_j) - J(\boldsymbol{\theta}_m) &\equiv E\{\ln [f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]\} \\ &= E\left\{\frac{d}{dt}[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t \Big|_{t=0}\right\} \\ &= E\{\lim_{t \to 0} ([f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t - 1)t^{-1}\}. \end{aligned} \tag{20}$$

Because of assumption ii) the order of averaging and limiting can be interchanged in (20) according to the Lebesgue dominated-convergence theorem [13]. Thus

$$J(\boldsymbol{\theta}_j) - J(\boldsymbol{\theta}_m) = \lim_{t \to 0} t^{-1}(E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t\} - 1). \tag{21}$$

Therefore for any δ, $0 < \delta < 1$, there exists a $t = t(\delta) \in \pi_T$ such that

$$t^{-1}(E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t\} - 1) \leq [J(\boldsymbol{\theta}_j) - J(\boldsymbol{\theta}_m)](1 - \delta)$$

or

$$E\{[f(\boldsymbol{x};\boldsymbol{\theta}_j)/f(\boldsymbol{x};\boldsymbol{\theta}_m)]^t\} \leq 1 - t(1-\delta)[J(\boldsymbol{\theta}_m) - J(\boldsymbol{\theta}_j)] = \eta. \tag{22}$$

Clearly then, $J(\boldsymbol{\theta}_m) > J(\boldsymbol{\theta}_j)$ implies that $\eta < 1$.

Since the variance of $\boldsymbol{\theta}(n)$ diminishes faster than the Cramér–Rao lower bound $0(n^{-1})$, $\boldsymbol{\theta}(n)$ is termed superefficient (a term coined by Le Cam). In [2] Le Cam pointed out that an estimate can be superefficient only on a parameter space of Lebesgue measure zero. Thus one is led to inquire at this point if in part the foregoing convergence properties apply to the continuous-parameter-space version of $\boldsymbol{\theta}(n)$. Comments on this follow in Section III.

III. Continuous Parameter Space

If the true parameter $\boldsymbol{\theta}_0$ is known to be contained in a bounded subset Θ' of Θ, we can conclude that the expected *a posteriori* probability mass (3) lying outside any ε-neighborhood of $\boldsymbol{\theta}_0$ diminishes to zero at exponential rate. Consequently the variance of $\boldsymbol{\theta}_1(n)$ decreases exponentially to ε^2.

Specifically, let $\mathscr{I}(\varepsilon)$ denote an ε-neighborhood of $\boldsymbol{\theta}_0$, i.e.,

$$\boldsymbol{\theta} \in \mathscr{I}(\varepsilon) \Leftrightarrow \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \varepsilon, \tag{23}$$

let $t \in \pi_T$, and assume that $J(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. The probability mass lying outside $\mathscr{I}(\varepsilon)$ can be bounded as

$$\begin{aligned} \int_{\Theta' - \mathscr{I}(\varepsilon)} p(\boldsymbol{\theta} \mid \boldsymbol{X}_n)\, d\boldsymbol{\theta} &\leq \frac{(\int_{\Theta' - \mathscr{I}(\varepsilon)} p(\boldsymbol{\theta} \mid \boldsymbol{X}_n)\, d\boldsymbol{\theta})^t}{(\int_{\mathscr{I}(\varepsilon/2)} p(\boldsymbol{\theta} \mid \boldsymbol{X}_n)\, d\boldsymbol{\theta})^t} \\ &= \left[\frac{\int_{\Theta' - \mathscr{I}(\varepsilon)} \prod_{k=1}^n f(\boldsymbol{x}_k;\boldsymbol{\theta}) \cdot p_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{\mathscr{I}(\varepsilon/2)} \prod_{k=1}^n f(\boldsymbol{x}_k;\boldsymbol{\theta}) p_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\right]^t. \end{aligned} \tag{24}$$

By the mean-value theorem the denominator of (24) can be expressed as

$$\prod_{k=1}^n [f(\boldsymbol{x}_k;\boldsymbol{\theta}'')]^t \cdot [V(\varepsilon/2)]^t, \tag{25}$$

where $\boldsymbol{\theta}''$ is some point interior to $\mathscr{I}(\varepsilon/2)$ and $V(\varepsilon/2)$ is the "volume" of $\mathscr{I}(\varepsilon/2)$, i.e.,

$$V(\varepsilon/2) = \int_{\mathscr{I}(\varepsilon/2)} d\boldsymbol{\theta}. \tag{26}$$

Since Θ' is bounded, the numerator of (24) can be expressed as a sum of M integrals over disjoint regions having volume no greater than $V(\varepsilon/2)$ and covering $\Theta' - \mathscr{I}(\varepsilon)$. By the mean-value theorem each such integral can be expressed as

$$\prod_{k=1}^n [f(\boldsymbol{x}_k;\boldsymbol{\theta}_r)]^t [V_r]^t, \qquad r = 1,2,\cdots,M, \tag{27}$$

where $\boldsymbol{\theta}_r$ is a point interior to the rth region in $\Theta' - \mathscr{I}(\varepsilon)$ and $V_r \leq V(\varepsilon/2)$. Letting $\boldsymbol{\theta}'$ index the maximum of the M quantities in (27), we can bound the right side of (24) as

$$Mc \prod_{k=1}^{n} [f(x_k;\theta'')/f(x_k;\theta')]^t. \tag{28}$$

According to (22), (24), and (28), the expected probability mass lying outside $\mathscr{I}(\varepsilon)$ can be bounded as

$$E\left\{\int_{\Theta'-\mathscr{I}(\varepsilon)} p(\theta \mid X_n)\, d\theta\right\} \le Mc\eta^n, \qquad 0 < \eta < 1. \tag{29}$$

For ε sufficiently small $J(\theta'') > J(\theta')$, i.e., $J(\theta)$ is convex on $\mathscr{I}(\varepsilon)$.

The expected norm-square error in $\theta_1(n)$ can now be computed rather simply. Define

$$\sigma_1^2(n) \equiv E\{\|\theta_1(n) - \theta_0\|^2\}. \tag{30}$$

By incorporating the expression defining $\theta_1(n)$ and expanding the norm, (30) becomes

$$\begin{aligned}\sigma_1^2(n) = E\Bigg\{&\left\|\int_{\Theta'-\mathscr{I}(\varepsilon)} (\theta - \theta_0)p(\theta \mid X_n)\, d\theta\right\|^2 \\ &+ 2\left[\int_{\Theta'-\mathscr{I}(\varepsilon)} (\theta - \theta_0)p(\theta \mid X_n)\, d\theta\right]^T \\ &\cdot \left[\int_{\mathscr{I}(\varepsilon)} (\theta - \theta_0)p(\theta \mid X_n)\, d\theta\right] \\ &+ \left\|\int_{\mathscr{I}(\varepsilon)} (\theta - \theta_0)p(\theta \mid X_n)\, d\theta\right\|^2\Bigg\}. \end{aligned} \tag{31}$$

Letting R denote max $\{\|\theta - \theta_0\|\}$, $\theta \in \Theta'$, and noting (29), we may bound the first term in (31) by $R^2Mc\eta^n$, the second term by $2\varepsilon RMc\eta^n$, and the third term by ε^2. Thus for arbitrarily small positive ε we obtain the bound

$$\sigma_1^2(n) \le R(R + 2\varepsilon)Mc\eta^n + \varepsilon^2. \tag{32}$$

This expression points up that for small n the variance bound is governed, through the number M, by the tightness of the bound on the parameter space, and through the "time constant" of the exponent, by the geometry of the information function near θ_0, the bound being the lower, the more sharply peaked the function $J(\theta)$ is near θ_0.

IV. Discussion of Convergence Mechanism

It becomes apparent at this point that the mechanism governing the convergence of the conditional-mean estimator is two-fold. The exponential rate stems from the product form of $p(\theta \mid X_n)$, whereas convergence to θ_0 is consequent on the maximization of $J(\theta)$ at θ_0. These facts suggest that we might expect the same behavior of other estimators that are formed as weighted averages over the parameter space, if the weighting coefficients are formed as products and the expectation of each factor in a product is a function having a unique maximum at the true parameter. This is in fact the case. Of course, such estimators do not necessarily minimize Bayes risk.

Specifically, define on Θ_v

$$\theta_\phi(n) \equiv \sum_{j=1}^{v} \theta_j \cdot q_\phi(\theta_j \mid X_n), \tag{33}$$

and

$$q_\phi(\theta_j \mid X_n) = \prod_{k=1}^{n} \phi(x_k;\theta_j) / \sum_{i=1}^{v} \prod_{k=1}^{n} \phi(x_k;\theta_i), \qquad i = 1,2,\cdots,v, \tag{34}$$

where $\phi(x;\theta)$ is any statistic such that

$$\max_{\theta \in \Theta} \{E[\ln \phi(\cdot;\theta)]\} = E[\ln \phi(\cdot;\theta_0)]. \tag{35}$$

Then

$$E\{\|\theta_\phi(n) - \theta_m\|^2\} \le [R(v-1)]^2\eta_\phi^n, \qquad 0 < \eta_\phi < 1, \tag{36}$$

where

$$\gamma(\theta) \equiv E\{\ln \phi(x;\theta)\}$$

and

$$\theta_m = \arg\max_{\theta \in \Theta_v} \{\gamma(\theta)\}. \tag{37}$$

The following are illustrative examples.

i) If

$$\phi(x;\theta) = \exp\left\{2f(x;\theta) - \int f^2(x;\theta)\, dx\right\},$$

then

$$\gamma(\theta) = \int f^2(x;\theta_0) - \int [f(x;\theta) - f(x;\theta_0)]^2\, dx.$$

ii) If

$$\phi(x;\theta) = \exp\left\{f(x;\theta) \Big/ \int f^2(x;\theta)\, dx\right\},$$

then

$$\gamma(\theta) = \int f(x;\theta)f(x;\theta_0)\, dx \Big/ \int f^2(x;\theta)\, dx.$$

Interestingly, Saridis *et al.* [11] have fabricated stochastic-approximation estimators that employ $\gamma(\theta)$ of example i) as a regression function, and recently Young and Coraluppi [12] have done likewise with $J(\theta)$. Thus (33)–(35) define a superefficient class of estimators, the Bayes conditional-mean estimator being one member.

Acknowledgment

The author is grateful to L. D. Davisson, C. L. Weber, and J. E. Ohlson for helpful suggestions and for providing the detection-theory interpretation of this discussion.

References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1.* New York: Wiley, 1968.
[2] L. Le Cam, "On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates," *Univ. Calif. Publ. Statist.*, vol. 1, 1953, pp. 277–330.
[3] E. A. Patrick and J. P. Costello, "On unsupervised estimation algorithms," *IEEE Trans. Inform. Theory*, vol. IT-16, Sept. 1970, pp. 556–569.
[4] E. L. Lehman, *Testing Statistical Hypotheses.* New York: Wiley, 1959, pp. 12–23.
[5] A. J. Viterbi, *Principles of Coherent Communication.* San Francisco: McGraw-Hill, 1966, appendix C, pp. 305–310.
[6] H. Teicher, "Identifiability of finite mixtures," *Ann. Math. Statist.*, vol. 34, 1963, pp. 1265–1269.

[7] ——, "Identifiability of mixtures," *Ann. Math. Statist.*, vol. 32, 1961, pp. 244–248.
[8] S. J. Yakowitz and J. D. Spragins, "On the identifiability of finite mixtures," *Ann. Math. Statist.*, vol. 39, 1968, pp. 209–214.
[9] S J. Yakowitz, "Unsupervised learning and the identification of finite mixtures," *IEEE Trans. Inform. Theory*, vol. IT-16, May 1970, pp. 330–338.
[10] S. Kullback, *Information Theory and Statistics.* New York: Dover, 1968, p. 14.
[11] G. N. Saridis, Z. J. Nikolic, and K. S. Fu, "Stochastic-approximation algorithms for system identification, estimation, and decomposition of mixtures," *IEEE Trans. Syst. Sci Cybern.*, vol. SSC-5, Jan. 1969, pp. 8–15.
[12] T. Y. Young and G. Coraluppi, "Stochastic estimation of a mixture of normal density functions using an information criterion," *IEEE Trans. Inform. Theory*, vol. IT-16, May 1970, pp. 258–263.
[13] W. Rudin, *Principles of Mathematical Analysis*, 2nd ed. New York: McGraw-Hill, 1964, pp. 246–247.
[14] E. A. Patrick and L. A. Liporace, "Quasi-Bayes averaging of stochastic-approximation estimators," *Inform. Contr.*, vol. 18, Mar. 1971, pp. 168–182

Barankin Bounds on Parameter Estimation

ROBERT J. McAULAY, MEMBER, IEEE, AND EDWARD M. HOFSTETTER, MEMBER, IEEE

Abstract—**The Schwarz inequality is used to derive the Barankin lower bounds on the covariance matrix of unbiased estimates of a vector parameter. The bound is applied to communications and radar problems in which the unknown parameter is embedded in a signal of known form and observed in the presence of additive white Gaussian noise. Within this context it is shown that the Barankin bound reduces to the Cramér–Rao bound when the signal-to-noise ratio (SNR) is large. However, as the SNR is reduced beyond a critical value, the Barankin bound deviates radically from the Cramér–Rao bound, exhibiting the so-called threshold effect.**

The bounds were applied to the linear FM waveform, and within the resulting class of bounds it was possible to select one that led to a closed-form expression for the lower bound on the variance of an unbiased range estimate. This expression clearly demonstrates the threshold behavior one must expect when using a nonlinear modulation system.

Tighter bounds were easily obtained, but these had to be evaluated numerically. The sidelobe structure of the linear FM compressed pulse leads to a significant increase in the variance of the estimate. For a practical linear FM pulse of 1-μs duration and 40-MHz bandwidth, the radar must operate at an SNR greater than 10 dB if meaningful unbiased range estimates are to be obtained.

I. Introduction

An IMPORTANT problem in communications and radar theory is the estimation of a set of parameters $(\theta_1,\theta_2,\cdots,\theta_m)$ of a signal that has been corrupted by additive white Gaussian noise. In particular, the received waveform is assumed to be of the form

$$r(t) = s(t;\theta) + n(t), \qquad |t| \le T. \tag{1}$$

In radar θ_1 might represent an unknown time delay (target range), θ_2 an unknown Doppler shift (target velocity), and θ_3 an unknown carrier phase angle. The noise term $n(t)$ represents the Gaussian white noise and is assumed to have two-sided spectral density N_0 W/Hz. It is of practical and theoretical interest to determine how well a given estimation scheme can perform with respect to estimating the unknown parameters. In this respect, Barankin [1] has derived a general class of lower bounds on the moments of unbiased estimators. Kiefer [2] has used the Schwarz inequality to obtain lower bounds on the variance of an unbiased estimate of a scalar-valued parameter. Applied to pulse-position modulation communications, this bound was shown to yield considerable information regarding the nonlinear-modulation threshold effect [3]. In this paper, we use the Schwarz inequality to derive lower bounds on the error covariance matrix for unbiased estimates of the vector parameter $\boldsymbol{\theta}$. Then we specialize these results to communications and radar as formulated in (1) and apply the bound to the particular problem of estimating the range of a target by using a linear FM waveform. It is shown that the sidelobes of the corresponding compressed pulse significantly affect the variance of the estimate of the time delay.

Manuscript received May 12, 1969; revised January 22, 1971, and May 20, 1971. This research was supported by the Department of the Army.

The authors are with the M.I.T. Lincoln Laboratory, Lexington, Mass. 02173.

II. Barankin Bound

Let Ω be a sample space of points ω and let $P(\omega \mid \theta)$ be a family of probability measures on Ω indexed by the parameter θ taking values in some index set π. Assume these measures have a density function with respect to some measure μ, i.e., there exists a function $p(\omega \mid \theta)$ such that

$$P(E \mid \theta) = \int_E p(\omega \mid \theta)\, d\mu(\omega)$$

for all measurable sets E.

Let $g(\cdot)$ be a real-valued function defined on π and let $\hat{g}(\cdot)$ be an unbiased estimator of $g(\theta)$, i.e., $\hat{g}(\cdot)$ is a real-valued measurable function defined on Ω with the property that

$$\int \hat{g}(\omega)\, p(\omega \mid \theta)\, d\mu(\omega) = g(\theta), \qquad \forall \theta \in \pi. \tag{2}$$

In Appendix I we use the Schwarz inequality to show that