

A NEW MULTIVARIATE CRAMER-RAO INEQUALITY FOR PARAMETER ESTIMATION

(Application: Input Probing Function Specification)

Thomas H. Kerr
The Analytic Sciences Corporation
6 Jacob Way
Reading, Massachusetts 01867

Abstract

A new form of the Cramer-Rao inequality for the estimator of vector parameter constants is presented. For a scalar Cramer-Rao inequality of the form of the one derived, the so-called Cramer-Rao lower bound does not have a denominator that must be maximized over all components of some matrix as was required in previous multivariate derivations.

For a certain class of maximum likelihood parameter estimation problems, the Cramer-Rao lower bound is the error of estimation. For this class of problems, a denominator having the form exhibited by this lower bound involves a trace and is shown to be a norm squared in a Hilbert space. Minimizing the error of estimation is shown to be equivalent to maximizing the norm in a Hilbert space while constrained to a specific compact set which represents practical constraints. The specification of input probing functions to aid in the estimation of input gain parameters in a linear dynamical system with system process noise is considered as a special case of this class of maximum likelihood parameter estimation problems. The probing functions are bang-bang.

1. Introduction

Previous derivations of a multivariate Cramer-Rao inequality either involved obtaining the denominator of the lower bound by maximizing over all components of some matrix [1],[2], or gave a lower bound in the positive semi-definite matrix sense [3](which is only a partial ordering*). The present derivation does not involve a maximization and yields a scalar inequality which serves as a total ordering.

The particularly nice form of the new Cramer-Rao inequality is applied to a specific class of maximum likelihood problems to show that the problem of minimizing the estimation error is equivalent to maximizing the norm in a Hilbert space.

2. The Vector Cramer-Rao Inequality

The new multivariate Cramer-Rao inequality is derived in Theorem 1. The proof parallels a proof in Nahi [4] before departing, except that vector rather than scalar arguments are used. Theorem 1 (Derivation of New Multivariate Cramer-Rao Inequality): Let θ be the n-vector parameter constant that is being estimated. Let $v(t)$ be the q-vector noise (random process) that affects the measurements. Let $y(t)$ be the

measurement p-vector. Let the function f relate the measurements to the parameters and noises:

$$y(t_i) = f[\theta, v(t_i)] \quad (i = k, \dots, k) \quad (1)$$

Let

$$\psi(\theta) \triangleq E[\hat{\theta}|\theta] = \theta + \phi(\theta) \quad (2)$$

where $\hat{\theta}$ is an estimate of θ , $\phi(\theta)$ is the bias of the estimate $\hat{\theta}$, and

$$\psi(\theta) = \int_{\hat{\theta}} \hat{\theta} p(\hat{\theta}|\theta) d\hat{\theta} \quad (3)$$

A lower bound on the error of estimation is

$$\text{trace } E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T | \theta\} \geq \frac{|\left(\frac{\partial}{\partial \theta}\right)^T \psi(\theta)|^2}{2^{n-1} E\left\{\left[\left(\frac{\partial}{\partial \theta}\right)^T \ln p(\underline{w}|\theta)\right]\left[\left(\frac{\partial}{\partial \theta}\right) \ln p(\underline{w}|\theta)\right]\right\}} \quad (4)$$

Proof: For any probability density function (pdf),

$$\int_{\hat{\theta}} p(\hat{\theta}|\theta) d\hat{\theta} = 1 \quad (5)$$

Differentiating Eq. (5) with respect to θ yields:

$$\int_{\hat{\theta}} \left(\frac{\partial}{\partial \theta}\right)^T p(\hat{\theta}|\theta) d\hat{\theta} = \begin{matrix} 0 \\ (1 \times n) \end{matrix} \quad (6)$$

Post-multiplying both sides by θ yields:

$$\int_{\hat{\theta}} \left(\frac{\partial}{\partial \theta}\right)^T p(\hat{\theta}|\theta) \theta d\hat{\theta} = \begin{matrix} 0 & \cdot & \theta \\ (1 \times n) & (n \times 1) & (1 \times 1) \end{matrix} = \begin{matrix} 0 \\ (1 \times 1) \end{matrix} \quad (7)$$

Differentiating Eq. (3) yields:

$$\left(\frac{\partial}{\partial \theta}\right)^T \psi(\theta) = \int_{\hat{\theta}} \left[\left(\frac{\partial}{\partial \theta}\right)^T p(\hat{\theta}|\theta)\right] \hat{\theta} d\hat{\theta} \quad (8)$$

subtracting Eq. (7) from Eq. (8) yields:

$$\left(\frac{\partial}{\partial \theta}\right)^T \psi(\theta) = \int_{\hat{\theta}} \left[\left(\frac{\partial}{\partial \theta}\right)^T p(\hat{\theta}|\theta)\right] (\hat{\theta} - \theta) d\hat{\theta} = \sum_{i=1}^n \int_{\hat{\theta}} \left[\left(\frac{\partial}{\partial \theta}\right)^T p(\hat{\theta}|\theta)\right] (\hat{\theta}_i - \theta_i) d\hat{\theta} \quad (9)$$

Taking absolute values of both sides of Eq. (9) and squaring yields:

*Standard mathematical jargon, used for conciseness to represent a universally accepted definition, are underscored when first used in this paper.

$$\left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2 = \left| \sum_{i=1}^n \int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right] (\hat{\theta}_i - \theta_i) d\hat{\theta} \right|^2 \quad (10)$$

Applying the inequality $\left| \sum_{i=1}^n a_i \right|^2 \leq 2^{n-1} \sum_{i=1}^n |a_i|^2$ to Eq. (10) yields:

$$\left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2 \leq 2^{n-1} \sum_{i=1}^n \left| \int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right] (\hat{\theta}_i - \theta_i) d\hat{\theta} \right|^2 \quad (11)$$

Applying the Cauchy-Schwarz inequality for integrals n times yields:

$$\begin{aligned} \left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2 &\leq 2^{n-1} \sum_{i=1}^n \left| \int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right] (\hat{\theta}_i - \theta_i) \sqrt{p(\hat{\theta} | \underline{\theta})} d\hat{\theta} \right|^2 \\ &\leq 2^{n-1} \sum_{i=1}^n \left[\int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right]^2 \frac{1}{p(\hat{\theta} | \underline{\theta})} d\hat{\theta} \right] \left[\int_{\hat{\theta}_i} (\hat{\theta}_i - \theta_i)^2 p(\hat{\theta} | \underline{\theta}) d\hat{\theta} \right] \quad (12) \end{aligned}$$

The inequality is all the more true for

$$\left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2 \leq 2^{n-1} \left(\sum_{i=1}^n \int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right]^2 \frac{1}{p(\hat{\theta} | \underline{\theta})} d\hat{\theta} \right) \cdot \left(\sum_{i=1}^n \int_{\hat{\theta}_i} (\hat{\theta}_i - \theta_i)^2 p(\hat{\theta} | \underline{\theta}) d\hat{\theta} \right) \quad (13)$$

or, equivalently,

$$\left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2 \leq 2^{n-1} \left(\sum_{i=1}^n \int_{\hat{\theta}_i} \left[\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right]^2 \frac{1}{p(\hat{\theta} | \underline{\theta})} d\hat{\theta} \right) \cdot \{ \text{trace } E[(\underline{\theta} - \hat{\theta})(\underline{\theta} - \hat{\theta})^T | \underline{\theta}] \} \quad (14)$$

Performing a division yields

$$\text{trace } E[(\underline{\theta} - \hat{\theta})(\underline{\theta} - \hat{\theta})^T | \underline{\theta}] \geq \frac{\left| \left(\frac{\partial}{\partial \underline{\theta}} \right)^T \underline{\psi}(\underline{\theta}) \right|^2}{2^{n-1} \int_{\hat{\theta}_i} \left[\left(\frac{\partial}{\partial \hat{\theta}_i} p(\hat{\theta} | \underline{\theta}) \right) \right]^2 \frac{1}{p(\hat{\theta} | \underline{\theta})} d\hat{\theta}} \quad (15)$$

Eq. (15) is one form of the Cramer-Rao inequality. Now let

$$\underline{W} = [y^T(t_1), y^T(t_2), \dots, y^T(t_k)]^T \quad (16)$$

The denominator of Eq. (15) will be converted from $p(\hat{\theta} | \underline{\theta})$ to $p(\underline{W} | \underline{\theta})$, one which does not depend upon the form of the estimator. Let the joint conditional density function of the observation vector \underline{W} be denoted by $p(\underline{W} | \underline{\theta})$.

Transform the $p \cdot k$ dimensional vector \underline{W} into a $p \cdot k$ dimensional vector \underline{Z} , where

$$\underline{Z} = \underline{Z}(\underline{W}) \quad (17)$$

and

$$\underline{W} = f^*(\underline{Z}) \quad (18)$$

are assumed to exist. Under the above condition we have

$$p(\underline{Z} | \underline{\theta}) = p(\underline{W} = f^*(\underline{Z}) | \underline{\theta}) \cdot \left| \det \left\{ \left(\frac{\partial}{\partial \underline{Z}} \right) f^*(\underline{Z}) \right\} \right| \quad (19)$$

where the determinant enclosed by absolute value signs is the Jacobian of transformation J . By the Jacobian of transformation, the following two equations result:

$$|J| d\underline{Z} = d\underline{W} \quad (20)$$

$$\begin{aligned} &\int_{\underline{Z}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{Z} | \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{Z} | \underline{\theta}) \right] \frac{1}{p(\underline{Z} | \underline{\theta})} d\underline{Z} \\ &= \int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W} | \underline{\theta}) \cdot |J| \right] \cdot \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W} | \underline{\theta}) \cdot |J| \right] \frac{1}{p(\underline{W} | \underline{\theta}) \cdot |J|} d\underline{W} \\ &= \int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W} | \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W} | \underline{\theta}) \right] \frac{1}{p(\underline{W} | \underline{\theta})} d\underline{W} \quad (21) \end{aligned}$$

Let $\underline{Z}(\underline{W})$ be chosen so that the first n components of $\underline{Z}(\underline{W})$ are $\hat{\theta}(\underline{W})$, that is,

$$\underline{Z}^T(\underline{W}) = \left[\hat{\theta}^T(\underline{W}), \underline{\xi}^T(\underline{W}) \right] \quad (22)$$

(1xn) (1x(p-k-n))

where $\underline{\xi}(\underline{W})$ is a $(p-k-n)$ -vector. Making use of Eq. (22) yields:

$$p(\underline{Z} | \underline{\theta}) = p[\hat{\theta}, \underline{\xi} | \underline{\theta}] = p(\hat{\theta} | \underline{\theta}) \cdot p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \quad (23)$$

which upon differentiating with respect to $\underline{\theta}$ yields:

$$\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{Z} | \underline{\theta}) = \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\hat{\theta} | \underline{\theta}) \right] \cdot p(\underline{\xi} | \hat{\theta}, \underline{\theta}) + p(\hat{\theta} | \underline{\theta}) \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] \quad (24)$$

Substituting Eq. (24) into Eq. (21) [and using Eq. (23)] yields:

$$\begin{aligned} &\int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W} | \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W} | \underline{\theta}) \right] \cdot \frac{1}{p(\underline{W} | \underline{\theta})} d\underline{W} \\ &= \left\{ \int_{\hat{\theta}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\hat{\theta} | \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\hat{\theta} | \underline{\theta}) \right] \frac{p(\underline{\xi} | \hat{\theta}, \underline{\theta})}{p(\hat{\theta} | \underline{\theta})} d\hat{\theta} d\underline{\xi} \right. \\ &\quad + \int_{\hat{\theta}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] \frac{p(\hat{\theta} | \underline{\theta})}{p(\underline{\xi} | \hat{\theta}, \underline{\theta})} d\hat{\theta} d\underline{\xi} \\ &\quad + \int_{\hat{\theta}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\hat{\theta} | \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] d\hat{\theta} d\underline{\xi} \\ &\quad \left. + \int_{\hat{\theta}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\hat{\theta} | \underline{\theta}) \right] d\hat{\theta} d\underline{\xi} \right\} \quad (25) \end{aligned}$$

Now since

$$\int_{\underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\xi} | \hat{\theta}, \underline{\theta}) \right] d\underline{\xi} = [0, 0, \dots, 0] \quad (26)$$

(Eq. (26) arises as a result of the fact that:

$$\int_{\underline{\xi}} p(\underline{\xi} | \hat{\theta}, \underline{\theta}) d\underline{\xi} = 1) \quad (27)$$

the last two integrals on the right side of Eq. (25) are zero. The result is

$$\int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W}|\underline{\theta}) \right] \frac{1}{p(\underline{W}|\underline{\theta})} d\underline{W}$$

$$= \int_{\underline{\hat{\theta}}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\hat{\theta}}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\hat{\theta}}|\underline{\theta}) \right] \frac{p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta})}{p(\underline{\hat{\theta}}|\underline{\theta})} d\underline{\hat{\theta}} d\underline{\xi}$$

$$+ \int_{\underline{\hat{\theta}}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta}) \right] \frac{p(\underline{\hat{\theta}}|\underline{\theta})}{p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta})} d\underline{\hat{\theta}} d\underline{\xi} \quad (28)$$

Since the last integral on the right side of Eq. (28) is always non-negative, the following result is obtained:

$$\int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W}|\underline{\theta}) \right] \frac{1}{p(\underline{W}|\underline{\theta})} d\underline{W}$$

$$\geq \int_{\underline{\hat{\theta}}, \underline{\xi}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\hat{\theta}}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\hat{\theta}}|\underline{\theta}) \right] \cdot \frac{p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta})}{p(\underline{\hat{\theta}}|\underline{\theta})} d\underline{\hat{\theta}} d\underline{\xi} \quad (29a)$$

Now since $\int_{\underline{\xi}} p(\underline{\xi}|\underline{\hat{\theta}}, \underline{\theta}) d\underline{\xi} = 1$, Eq. (29a) may be re-written as

$$\int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W}|\underline{\theta}) \right] \frac{1}{p(\underline{W}|\underline{\theta})} d\underline{W}$$

$$\geq \int_{\underline{\hat{\theta}}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\hat{\theta}}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{\hat{\theta}}|\underline{\theta}) \right] \frac{1}{p(\underline{\hat{\theta}}|\underline{\theta})} d\underline{\hat{\theta}} \quad (29b)$$

Now the Cramer-Rao inequality of Eq. (15) may be expressed in terms of the measurements, \underline{W} , as

$$\text{trace } E \left[(\underline{\theta} - \hat{\underline{\theta}})(\underline{\theta} - \hat{\underline{\theta}})^T | \underline{\theta} \right] \geq \frac{\left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{\theta}) \right]^2}{2^{n-1} \int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W}|\underline{\theta}) \right] \frac{1}{p(\underline{W}|\underline{\theta})} d\underline{W}} \quad (30)$$

It is sometimes convenient to re-express the denominator of Eq. (30) by using

$$\int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) p(\underline{W}|\underline{\theta}) \right] \frac{1}{p(\underline{W}|\underline{\theta})} d\underline{W}$$

$$= \int_{\underline{W}} \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) \ln p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) \ln p(\underline{W}|\underline{\theta}) \right] p(\underline{W}|\underline{\theta}) d\underline{W}$$

$$= E \left\{ \left[\left(\frac{\partial}{\partial \underline{\theta}} \right)^T \ln p(\underline{W}|\underline{\theta}) \right] \left[\left(\frac{\partial}{\partial \underline{\theta}} \right) \ln p(\underline{W}|\underline{\theta}) \right] \right\} \quad (31)$$

The final result is Eq. (4). ■

Remark: Notice that for $n = 1$, Eq. (4) reduces to the usual scalar Cramer-Rao inequality.

Remark: It is ^{not} this specific lower bound of the Cramer-Rao inequality, but the form of the denominator that is of interest as a criterion for a total ordering in the input probing function appli-

cation to be discussed in the remainder of this paper. Any other representation of the Cramer-Rao inequality that has the same lower bound denominator, neglecting any constant factors, will suffice and will yield results identical to the ones to be discussed, except that Eq. (46) and (51) will change by that constant factor.

3. An Application in Maximum Likelihood Parameter Estimation

Consider the following problem. Suppose that the vector of measurements \underline{W} , defined in Eq. (16), may be modeled as

$$\underline{W} = H(\underline{u}) \cdot \underline{m} + \underline{V} \quad (32)$$

where \underline{m} is the unknown vector parameter to be estimated, $\underline{m} \in \mathbb{R}^n$, \underline{u} is a vector or scalar deterministic parameter which may be selected, $\underline{u} \in U \subset B$, where U is a compact convex set which contains the origin (the significance of this compact convex set U will be apparent when practical implementation constraints are considered in the model), B is a Banach space, $H(\underline{u})$ is a linear function of \underline{u} , and \underline{V} is vector Gaussian measurement noise.

$$E[\underline{V}] = \underline{d} \quad (33)$$

$$\text{Var}[\underline{V}] = \underline{L} = \underline{L}^T > 0 \quad (34)$$

(Notice that the observations \underline{W} are linear in the parameter \underline{m} and in the noise \underline{V} .) The conditional probability density function of the corresponding \underline{W} given \underline{m} is:

$$p(\underline{W}|\underline{m}) = c \exp \left\{ -\frac{1}{2} [\underline{W} - H(\underline{u}) \cdot \underline{m} - \underline{d}]^T \underline{L}^{-1} [\underline{W} - H(\underline{u}) \cdot \underline{m} - \underline{d}] \right\} \quad (35)$$

where c is the proper normalization constant. The column vector that maximizes Eq. (35) is the maximum likelihood estimate $\underline{\hat{m}}$ of \underline{m} .

By differentiating Eq. (35) with respect to \underline{m} , setting the result equal to the null element of \mathbb{R}^n , and solving for \underline{m} , the following maximum likelihood estimate is obtained:

$$\underline{\hat{m}} = [H^T(\underline{u}) \underline{L}^{-1} H(\underline{u})]^{-1} H^T(\underline{u}) \underline{L}^{-1} [\underline{W} - \underline{d}] \quad (36)$$

Notice that $\underline{\hat{m}}$ is an unbiased linear estimator since $E[\underline{\hat{m}}] = \underline{\hat{m}}$ and $\underline{\hat{m}}$ is linear in \underline{W} .

With this estimate $\underline{\hat{m}}$ are associated all of the advantages of a maximum likelihood estimate, namely:

- (i) the maximum likelihood estimate is asymptotically unbiased [5, p 185] (the estimate of Eq. (36) is strictly unbiased).
- (ii) when the observations, \underline{W} , are linear in the parameter and in the noise (as it is in the present application), then the maximum likelihood estimate is efficient (i. e., it actually achieves the Cramer-Rao lower bound) regardless of the length of the data [4, p 252].

The Cramer-Rao lower bound for this application will now be derived.

Theorem 2: For $\phi(\underline{m}) = 0$ in Eq. (2) (i.e., Eq. (36) is an unbiased linear estimator of \underline{m}), or equivalently

$$\psi(\underline{m}) = \underline{m} \quad (37)$$

and \underline{W} having the pdf of Eq. (35), the Cramer-Rao inequality of Theorem 1, Eq. (4) becomes

$$\text{trace } E \left\{ (\underline{m} - \hat{\underline{m}}) (\underline{m} - \hat{\underline{m}})^T | \underline{m} \right\} \geq \frac{n^2}{2^{n-1} \text{trace } H^T(\underline{u}) L^{-1} H(\underline{u})} \quad (38)$$

Proof: Differentiating Eq. (37) with respect to \underline{m} yields:

$$\left(\frac{\partial}{\partial \underline{m}} \right)^T \psi(\underline{m}) = 1 + 1 + \dots + 1 = n \quad (39)$$

hence the numerator of Eq. (4) is

$$\left| \left(\frac{\partial}{\partial \underline{m}} \right)^T \psi(\underline{m}) \right|^2 = n^2 \quad (40)$$

Taking the natural logarithm of Eq. (35) yields:

$$\ln p(\underline{W} | \underline{m}) = \ln c - \frac{1}{2} [\underline{W} - H(\underline{u}) \underline{m} - \underline{d}]^T L^{-1} [\underline{W} - H(\underline{u}) \underline{m} - \underline{d}] \quad (41)$$

Differentiating both sides of Eq. (41) with respect to \underline{m} yields:

$$\left(\frac{\partial}{\partial \underline{m}} \right)^T \ln p(\underline{W} | \underline{m}) = H^T(\underline{u}) L^{-1} [\underline{W} - H(\underline{u}) \underline{m} - \underline{d}] \quad (42)$$

Applying the following two equalities as lemmas

$$(1) \quad \underline{x}^T H H^T \underline{y} = \text{trace } H^T \underline{xy}^T H \quad (43)$$

$$(2) \quad E[\underline{x}^T H H^T \underline{y}] = \text{trace } H^T E[\underline{xy}^T] H \quad (44)$$

yields:

$$\begin{aligned} E \left[\left(\frac{\partial}{\partial \underline{m}} \right)^T \ln p(\underline{W} | \underline{m}) \right] \left[\left(\frac{\partial}{\partial \underline{m}} \right)^T \ln p(\underline{W} | \underline{m}) \right] \\ = \text{trace } H^T(\underline{u}) L^{-1} E \left\{ [\underline{W} - H(\underline{u}) \underline{m} - \underline{d}] [\underline{W} - H(\underline{u}) \underline{m} - \underline{d}]^T \right\} L^{-1} H(\underline{u}) \\ = \text{trace } H^T(\underline{u}) L^{-1} L L^{-1} H(\underline{u}) \end{aligned} \quad (45)$$

which is the denominator of Eq. (38). The final result of Eq. (38) is obtained by substituting Eq. (45) and Eq. (40) into Eq. (4). ■

Since Eq. (36) is a maximum likelihood estimate and the noise and parameter enter linearly in Eq. (32), equality exists in Eq. (38) since the Cramer-Rao lower bound is achieved. The error of estimation is

$$\frac{n^2}{2^{n-1} \text{trace } H^T(\underline{u}) L^{-1} H(\underline{u})} \quad (46)$$

This error is minimized when $\underline{u} \in U$ is chosen to maximize

$$\text{trace } H^T(\underline{u}) L^{-1} H(\underline{u}) \quad (47)$$

4. Insights Into the Problem's Inner Mathematical Structure

Observe from Theorem 1, Eq. (16), and Eq. (32) that $H(\cdot)$ of Eq. (32) maps B into the space of $(p \cdot k \times n)$ constant matrices. The linear vector space of $(p \cdot k \times n)$ matrices having the inner product

$$(A, C) \triangleq \text{trace } A^T L^{-1} C \quad (48)$$

where L^{-1} is positive definite, is a Hilbert space [6, p. 72]. Further assume that $H(\cdot)$ is a continuous linear map of B into the space of $(p \cdot k \times n)$ matrices. Now

$$\|H(\underline{u})\|^2 \triangleq (H(\underline{u}), H(\underline{u})) = \text{trace } H^T(\underline{u}) L^{-1} H(\underline{u}) \quad (49)$$

and $\|H(\underline{u})\|^2$ is maximized when $\|H(\underline{u})\|$ is maximized.

The above mathematical structure will be used to accomplish the following two objectives:

- (i) to establish that optimal probing functions exist for problems having the structure of Eq. (32)-(34) using function space methods.
- (ii) to determine the nature of the optimal probing functions.

The following Lemmas are arranged in the proper logical order to rigorously accomplish the above two objectives. However, rather than reinvent the wheel, references will be given for proofs to Lemmas representing well known results, the proofs to other Lemmas are sketched to give a feel for what level of mathematical sophistication and manipulation is used, while the proofs of the remaining Lemmas are obvious, once the Lemma is correctly stated.

Lemma 1: $\|H(\underline{u})\|$ is continuous in \underline{u} , where

$$\|H(\underline{u})\| = \sqrt{(H(\underline{u}), H(\underline{u}))}$$

Proof: $\|\cdot\|$ is a continuous function and $H(\cdot)$ was assumed continuous and linear, hence the composite function $\|H(\underline{u})\| \triangleq \|\cdot\| \circ H(\underline{u})$ is continuous in \underline{u} since the composite of two continuous functions is continuous. ■

Lemma 2: $\|H(\cdot)\|$ is a convex functional.

Proof: For any $\underline{u}^A, \underline{u}^B \in B$ and $\lambda \in \mathbb{R}^1, 0 < \lambda < 1$,
 $\|H(\lambda \underline{u}^A + (1-\lambda)\underline{u}^B)\| = \|\lambda H(\underline{u}^A) + (1-\lambda)H(\underline{u}^B)\| \leq$
 $|\lambda| \|H(\underline{u}^A)\| + |1-\lambda| \|H(\underline{u}^B)\| = \lambda \|H(\underline{u}^A)\| +$
 $(1-\lambda) \|H(\underline{u}^B)\|$ by the linearity of $H(\cdot)$. ■

Lemma 3: There exists a $\underline{u}^* \in U \subset B$ such that

$$\max_{\underline{u} \in U} \|H(\underline{u})\| = \|H(\underline{u}^*)\|.$$

Proof: $\|H(\underline{u})\|$ is upper semi-continuous in \underline{u} since it is all the more continuous in \underline{u} . U was assumed to be compact in the Banach space B , a normed linear space. Lemma 3 holds since an upper semi-continuous functional on a compact subset U of a normed linear space B achieves a maximum on U [6, p. 40]. Let \underline{u}^* denote the point where $\|H(\cdot)\|$ achieves this maximum. ■

Corollary 1: In Lemma 3, if neither $U = \{\theta\}$ nor $H(\cdot)$ is identically the null function, then $\|H(\underline{u}^*)\| \neq 0$.

Proof: Obvious. ■

Corollary 2: (Existence of optimal probing function): If neither $U = \{\theta\}$ nor $H(\cdot)$ is identically the null function, then there exists a $u^* \in U \subset B$ such that

$$\max_{u \in U} \|H(u)\|^2 = \|H(u^*)\|^2 \quad \text{and} \\ \|H(u^*)\|^2 \neq 0.$$

Proof: The square of the norm is a maximum where the norm is a maximum since the square is a convex function. ■

Now the convexity of U and the fact that it contains the origin along with the convexity of $\|H(u)\|$ in u are used to establish that the maximum of $\|H(u)\|$ occurs on ∂U , the boundary of U .

Lemma 4: For the conditions of Corollary 2, for $u^* \in U$, where U is compact and contains the origin θ , such that $\|H(u)\| \leq \|H(u^*)\|$ for all $u \in U$, then $u^* \in \partial U$.

Proof: By contradiction, assume that $u^* \notin \partial U$. Since U is compact, U is closed and $\partial U \subset U$. By the convexity of U , there exists a $u^{**} \in \partial U$ and $\lambda^* \in (0, 1)$ such that

$$u^* = \lambda^* u^{**} + (1 - \lambda^*) \theta$$

where θ is the null element of B . Now

$$\|H(u^*)\| = \|H(\lambda^* u^{**} + (1 - \lambda^*) \theta)\| = \|\lambda^* H(u^{**})\| \leq \lambda^* \|H(u^{**})\| \quad (*)$$

Now $\lambda < 1$ (**)

and $\|H(u^*)\| \neq 0$ implies that $\|H(u^{**})\| \neq 0$ by Corollary 2 and (*). Now from (**), we have that

$$\lambda^* \|H(u^*)\| < \|H(u^{**})\|$$

so (*) becomes

$$\|H(u^*)\| < \|H(u^{**})\| \text{ for } u^{**} \in \partial U \subset U$$

which contradicts the result of Lemma 3, therefore $u^* \in \partial U$. ■

An intuitive geometric feel for the conclusions of Lemma 3 and Lemma 4 may be obtained from Figure 1, where the convex functional $\|H(\cdot)\|$ is depicted. The functional is seen to achieve its maximum on the boundary of U .

Figure 2 illustrates another way to solve the problem which makes use of the following lemma.

Lemma 5: The image of U under $H(\cdot)$, $H[U]$, is convex and compact since U is convex and compact.

Proof: Uses Lemma 2 and [7, p. 58]. ■

Finding the y^* , the $(p \times n)$ constant matrix that is contained in $H[U]$ which maximizes the norm

$$\|\cdot\| = \sqrt{\text{trace}(\cdot)^T L^{-1}(\cdot)},$$

then taking the pseudo-inverse $H^\dagger(y^*)$ yields the proper u^* , i.e.,

$$u^* = H^\dagger(y^*) = H^\dagger(\arg \max \|y\|) \quad (50)$$

to maximize the denominator hence minimize the Cramer-Rao lower bound, the error of estimation for the maximum likelihood estimate for this problem is

$$\frac{n^2}{2^{n-1} \text{trace } H^T(u^*) L^{-1} H(u^*)} \quad (51)$$

By a possible decomposition of H into two continuous linear transformations such as

$$H(\cdot) = (H_2 \circ H_1)(\cdot) \quad (52)$$

and obtaining the corresponding adjoint operators for H_1 and H_2 as H_1^* and H_2^* , respectively (e.g., see [6, ex. 1-4, pp. 153-4]), the adjoint of $H(\cdot)$ is

$$H^*(\cdot) = (H_1^* \circ H_2^*)(\cdot) \quad (53)$$

and the pseudo-inverse of H , used in calculating u^* in Eq. (50), is

$$H^\dagger = \{(H_1^* \circ H_2^*) \circ [(H_2 \circ H_1) \circ (H_1^* \circ H_2^*)]^{-1}\}(\cdot) \quad (54)$$

5. A Special Case: Input Probing Function Specification

Consider the following continuous time Bayesian Model [8, p. 28]. System (described by a stochastic differential equation):

$$\dot{\underline{x}}(t) = F \underline{x}(t) + G \underline{w} + \underline{m} u(t) \quad (55)$$

where $u(t)$ is a deterministic control function, each component of $u(t)$ being a member of $L^2[0, T]$ for some finite T (a reasonable physical constraint, since we will only be working with finite length data records), $\underline{w}(t)$ is a zero mean Gaussian white noise (the formal derivative of a Brownian motion process),

$$E[\underline{w}(t)] = 0 \text{ for all } t \quad (56)$$

$$E[\underline{w}(t) \underline{w}^T(\tau)] = I \delta(t - \tau) \text{ for all } t, \tau \quad (57)$$

$\underline{x}(0)$ is a random vector initial condition, independent of $\underline{w}(t)$ for all t ,

$$E[\underline{x}(0)] = 0 \quad (58)$$

$$\text{Var}[\underline{x}(0)] = P \quad (59)$$

Measurement:

$$\underline{y}(t) = H \underline{x}(t) + \underline{d}$$

where F , G , \underline{m} , P , H , \underline{d} are constant

matrices. Eq. (55) is notation for a mathematically rigorous stochastic integral equation where integration with respect to the white noise is replaced with an Ito integral in which integration is with respect to the Brownian motion process. (An apparently more general model than Eq. (55)-(60) would consist of having Q instead of I in Eq. (57) and G_0 instead of G in Eq. (55). By factoring

$$G_0 Q G_0^T \text{ into } G G^T,$$

both models may be shown to be equivalent by demonstrating that the same Fokker-Planck equation is obtained for both system models.)

Assume that $F, G, H, P,$ and d are known from previous calculations [9][10].

Objectives: (1) find m (i. e., obtain an estimate, \hat{m} , of m), the so-called input gain matrix. (2) specify deterministic inputs $u(t)$ to obtain the "most information" about m in the estimate \hat{m} . (A reasonable physical constraint is that only bounded piecewise continuous inputs may be generated so that:

$$U = D[0, t_k] \cap \{u(t) \in L^2[0, t_k] \mid |u(t)| \leq \rho \text{ for } 0 \leq t \leq t_k\} \quad (61)$$

where $D[0, t_k]$ is the function space of piecewise continuous functions and

$$\{u(t) \in L^2[0, t_k] \mid |u(t)| \leq \rho \text{ for } 0 \leq t \leq t_k\}$$

is the function space of functions bounded by $\rho(t_k = T)$. U is the intersection of these two convex sets and is also convex. U is also compact [11, p. 44].)

$$B = L^2[0, t_k] \quad (62)$$

and is both a Hilbert and a Banach space [12]. For measurements taken at times t_1, t_2, \dots, t_k , this problem satisfies the structure requirement of Eq. (32) with

$$H(u) = \begin{bmatrix} \int_0^{t_1} H e^{F(t_1-\tau)} u(\tau) d\tau \\ \int_0^{t_2} H e^{F(t_2-\tau)} u(\tau) d\tau \\ \vdots \\ \int_0^{t_k} H e^{F(t_k-\tau)} u(\tau) d\tau \end{bmatrix} \quad (63)$$

$$V = \begin{bmatrix} (II) \int_0^{t_1} H e^{F(t_1-\tau)} G d\beta_\tau + H e^{F t_1} x(o) + d \\ (II) \int_0^{t_2} H e^{F(t_2-\tau)} G d\beta_\tau + H e^{F t_2} x(o) + d \\ \vdots \\ (II) \int_0^{t_k} H e^{F(t_k-\tau)} G d\beta_\tau + H e^{F t_k} x(o) + d \end{bmatrix} \quad (64)$$

$$E[V] = [d^T, d^T, \dots, d^T]^T \quad (65)$$

$$\text{Var}[V] = \begin{bmatrix} R_{yy}(o) & R_{yy}(t_2, t_1) & R_{yy}(t_3, t_1) & \dots & R_{yy}(t_k, t_1) \\ R_{yy}(t_2, t_1) & R_{yy}(o) & R_{yy}(t_3, t_2) & \vdots & \vdots \\ R_{yy}(t_3, t_1) & R_{yy}(t_3, t_2) & R_{yy}(o) & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{yy}(t_k, t_1) & \dots & \dots & \dots & R_{yy}(o) \end{bmatrix} \quad (66)$$

Now it will be shown that $H(\cdot)$ satisfies all of the properties that were assumed in the last section (i. e., linear and continuous in u).

Lemma 6: $H(\cdot)$ of Eq. (63) is a linear mapping.

Proof: For any $u^A, u^B \in L^2[0, t_k]$, for any $\alpha, \beta \in \mathbb{R}$, it may be seen that $H(\alpha u^A + \beta u^B) = \alpha H(u^A) + \beta H(u^B)$ by using Eq. (63), the linearity of integration and the additivity and scalar multiplication property of matrices.

Lemma 7: $H(\cdot)$ of Eq. (63) is continuous in u .

Proof: $L^2[0, t_k]$ is a metric space with $d(x, y) = \sqrt{\langle x-y, x-y \rangle}$. Since every metric space is first countable [14, p. 102], $L^2[0, t_k]$ is first countable. Now $\{u_n\} \xrightarrow{L^2} \bar{u}$ (strongly) implies that $\{u_n\} \xrightarrow{L^2} \bar{u}$ (weakly) [6, p127]. In a Hilbert space, $\{u_n\} \xrightarrow{L^2} \bar{u}$ (weakly), if and only if $(u_n, y)_2 \rightarrow (\bar{u}, y)_2$ as $n \rightarrow \infty$ for each $y \in L^2[0, t_k]$ [15, p. 111]. Therefore, we have that $H(u_n) \rightarrow H(\bar{u})$ as $n \rightarrow \infty$ by taking the limit of each element of the matrix $H(\cdot)$ and applying it to each partitioned row of Eq. (63). Since this sequential continuity implies continuity in a first countable space [7, p. 44], $H(\cdot)$ is continuous. With these properties of $H(\cdot)$ established and U compact and convex, the conclusions of the previous section hold. Specifically, Eq. (36) is the maximum likelihood estimate of m in Eq. (55). The error of estimation is given in Eq. (51) and the input probing function that yields the most information about m is given by Eq. (50). The probing function falls on the boundary of piecewise continuous functions bounded by ρ , hence the u^* that maximizes the norm-squared is bang-bang. This norm squared is

$$\text{trace } H^T(u) L^{-1} H(u)$$

$$= \int_0^{t_k} \int_0^{t_k} \sum_{m=1}^k \sum_{s=1}^k \text{trace} \{ e^{F^T(t_m-\tau)} H^T(L^{-1})_m H e^{F(t_s-\tau)} [1(\tau)-1(\tau-t_m)] [1(\tau)-1(\tau-t_s)] \} u(\tau) u(s) d\tau ds \quad (67a)$$

$$= \int_0^{t_k} \int_0^{t_k} K(\tau, s) u(\tau) u(s) d\tau ds \quad (67b)$$

where $K(\tau, s)$ is the scalar kernel that is contained within the furthest brackets in Eq. (67a), and $1(\cdot)$ is the unit step function.

Conjecture: For F with all eigenvalues real and negative and (H, F) observable and (F, G) controllable, u^* has at most $(n-1)$ switchings in each of the intervals $(o, t_1), (t_1, t_2), (t_2, t_3), \dots, (t_{k-1}, t_k)$ or at most $k \cdot (n-1)$ switchings total on (o, t_k) . (Paralleling the usual result in time optimal bang-bang control.)

Comment: This conjecture is consistent with a first order example in [4, p. 264] where the u^* was constant, ρ , but $(n-1)=(1-1)=0$ switchings!

Since the optimum input probing function (optimum in the sense that it minimizes the error of estimation associated with the maximum likelihood estimator of m) is bang-bang, the magnitude of u^* is known to be ρ , and all that must be determined for a complete specification are the scalar switching times. Hence, the problem of specifying an infinite dimensional general probing

function has been reduced to determining a finite number of scalar switching times.

One may also be interested in selecting the times at which measurements are taken (t_1, t_2, \dots, t_k), a choice which should be based upon Eq. (6) which reflects the correlation times of the noise.

There have been other treatments of optimal probing function specification for more general problems [3], [4], [13], [16] but the general problem does not have this much nice mathematical structure. Most of the previous emphasis has been on constrained energy of the input function; while what is presented here is constrained magnitude of control on a finite time interval, but the energy is also constrained. References [3], [13] and [16] have extensive bibliographies of the area.

6. Conclusion

A new form of the Cramer-Rao inequality was presented. It was shown that this version of the inequality has a lower bound that is particularly convenient since the denominator is the norm-squared in a Hilbert space of constant matrices. This allows the exploitation of the large body of knowledge on Hilbert spaces that has been accumulated by mathematicians. This lower bound is the error of estimation for a class of maximum likelihood parameter estimation problems. The underlying mathematical structure and the method of solution is detailed for this class of problems. The problem of optimal bounded input probing function selection for identifying input gain matrices in noisy linear systems is shown to be of this special class.

The author gratefully acknowledges a helpful discussion with Dr. Theodore Bick of Union College, Schenectady, N. Y.

References

1. E.A. Patrick, Fundamentals of Pattern Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1972.
2. H. Van Trees, Detection Estimation and Modulation Theory: Part I, John Wiley and Sons, Inc., N.Y. 1968.
3. M. Aoki and M. Staley, On Input Signal Synthesis in Parameter Identification, *Automatica*, Vol. 6, 1970.
4. N. Nahi, Estimation Theory and Applications, John Wiley and Sons, N.Y., 1969.
5. N.L. Johnson and F.C. Leone, Statistics and Experimental Design, Vol. 1, John Wiley and Sons, N.Y., 1964.
6. D.G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, N.Y., 1969.
7. J. Greever, Theory and Examples of Point Set Topology, Brooks/Cole Publishing Co., Belmont, Ca., 1967.
8. F.C. Schweppe, Uncertain Dynamic Systems, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973.
9. E.D. Eyman, T.H. Kerr, and N.K. Loh, Modeling a Particular Class of Multiple-Input/Multiple-Output Black Boxes with Stochastic Integral Equations and Identifying the Required Parameters, *Proceedings of I.F.A.C. Symposium on Identification*, Hague, Netherlands, 1973.

10. T.H. Kerr, Applying Stochastic Integral Equations to Solve a Particular Stochastic Modeling Problem, Ph.D Thesis, University of Iowa, 1971.
11. L.A. Lusternik and V.J. Sobolev, Elements of Functional Analysis, Ungar Publishing Co. N.Y., 1965.
12. W. Rudin, Real and Complex Analysis, McGraw-Hill Book Co., N.Y., 1963.
13. R.K. Mehra, Optimal Inputs for Linear System Identification, *Proceedings of the J.A.C.C.*, Stanford, Ca., 1972.
14. W.J. Pervin, Foundations of General Topology, Academic Press, N.Y., 1964.
15. S.K. Berberian, Introduction to Hilbert Space, Oxford University Press, N.Y., 1961.
16. R.T.N. Chen, "Input for Parameter Identification Part I: A New Formulation and a Practical Solution," *J.A.C.C.*, 1974.

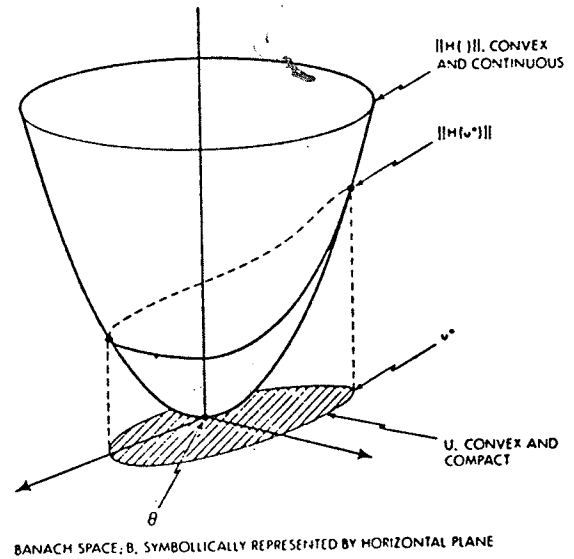


Figure 1 Existence of Optimal Probing Function on Boundary of U

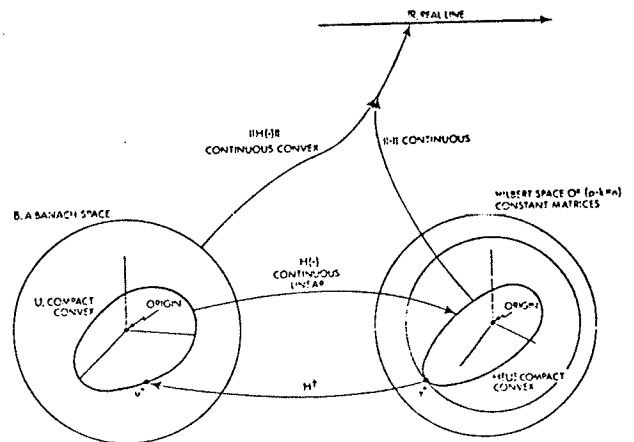
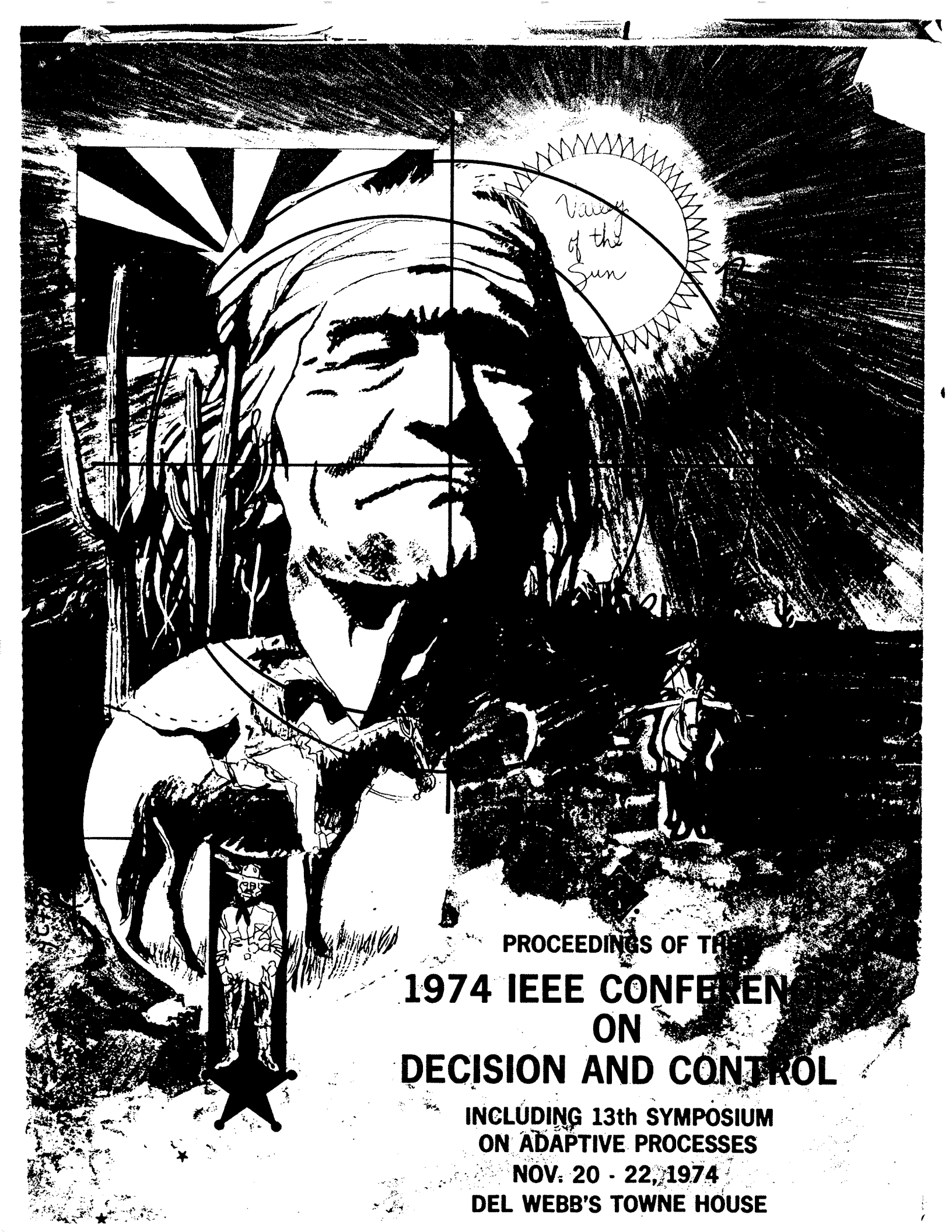


Figure 2 Spaces and Mappings Used in Finding u^* to Maximize $\|H(u)\|$



Valley
of the
Sun

PROCEEDINGS OF THE
**1974 IEEE CONFERENCE
ON
DECISION AND CONTROL**

INCLUDING 13th SYMPOSIUM
ON ADAPTIVE PROCESSES
NOV. 20 - 22, 1974
DEL WEBB'S TOWNE HOUSE