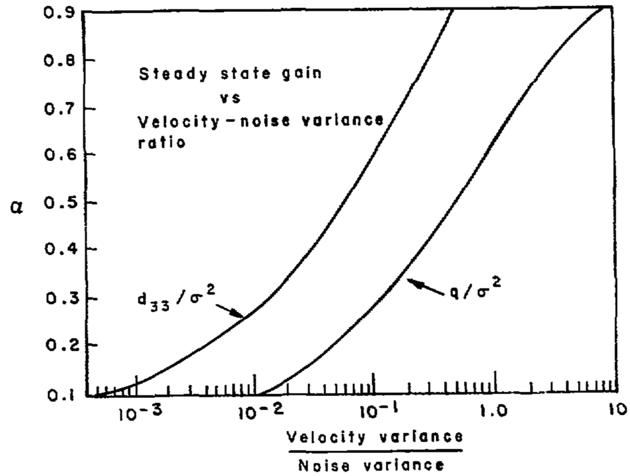

Fig. 1.

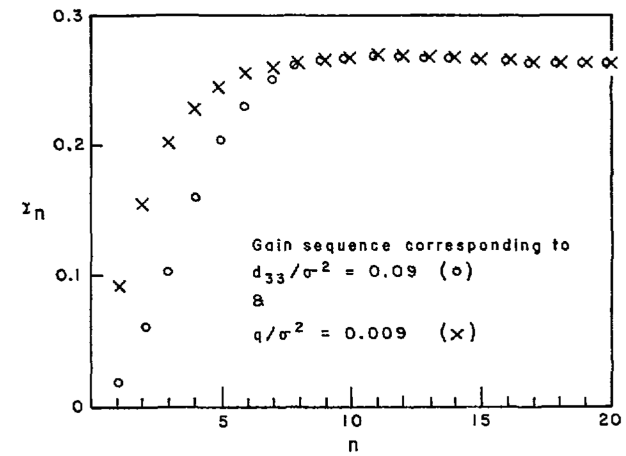

Fig. 2.

The preceding filter gains converge to

$$\alpha_{n,0} \rightarrow \alpha_0 = \frac{2d_{33}}{2d_{33} + \alpha_1^2\sigma^2} \tag{26}$$

where

$$\alpha_{n,1} \rightarrow \alpha_1 = [\rho + \rho\sqrt{1 + \rho/27}]^{1/3} + [\rho - \rho\sqrt{1 + \rho/27}]^{1/3} \tag{27}$$

and

$$\rho = \frac{d_{33}}{\sigma^2}.$$

Hence in steady state, the prediction gain as given by (20) and (21) can, by a proper choice of q, be equal to the optimal gain given by (27). However, the actual magnitude of q must be several times larger than d_{33} to obtain the same steady-state gain. This is illustrated in Fig. 1. Finally, even though the steady-state gains are identical, the gain sequences will generally differ prior to the steady state, as seen in Fig. 2.

V. Conclusion

An approximate solution to the stated problem (6) is given by (11), (12), (15), and (16). Some care must be taken if the solution is programmed in a general fashion, since components of $E(x_n x_n^T)$ and $E(x_n \tilde{x}_n^T)$ may be unbounded when either Φ, $\delta\Phi$, or $\Phi + \delta\Phi$ is unstable. However, the results can be very useful in resolving certain practical aspects of the implementation of linear estimation theory.

In general, (18) can be used to decide whether or not the computations (11), (12), (15), and (16) represent a significant improvement over the usual approach (3) through (5). More significantly, (18) should be helpful in determining the sensitivity of the estimator to model approximations as well as aid in determining the elements of Q in the solution given by (3) through (5).

Finally, a comment on the simplicity of the example. Due to a space constraint, a simple example was chosen in an attempt to illustrate the results with as little confusion as possible. A more useful example can be obtained by considering third-order dynamics for $\Phi + \delta\Phi$ while Φ is second order. This will lead to an analysis of the much studied $\alpha - \beta$ tracker[6] in a maneuvering target environment. This last application is given in Heller.[7]

References

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Engrg.*, ser. D, vol. 82, pp. 34–35, March 1960.
[2] A. G. Carlton, "Linear estimation in stochastic processes," Bumblebee Ser. Rept. 311, JHU/APL, March 1962.
[3] J. Peschon and R. E. Larson, "Analysis of an intercept system," Stanford Research Institute, Stanford, Calif., SRI Project 5188, December 1965, p. E-4.
[4] S. F. Schmidt, "The application of state-space methods to navigation problems," Guidance and Control System Engineering Dept., Philco, Tech. Rept. 4, July 1964.
[5] L. M. Spetner, "Estimation of corrections to estimation matrix and dynamics matrix in discrete linear estimation," Internal Memo. CLM-80, APL/JHU, December 1963.
[6] T. R. Benedict and G. W. Bordner, "Synthesis of an optimal set of radar track-while-scan smoothing equations," *IRE Trans. Automatic Control*, vol. AC-7, pp. 27–32, July 1963.
[7] B. J. Heller, "An alpha-beta tracker filter," U S Naval Ordnance Test Station, China Lake, Calif., Tech. Note 304-134, October 1966.

Maximum Likelihood Estimation of Rational Transfer Function Parameters

ALAN E. ROGERS, MEMBER, IEEE, AND
KENNETH STEIGLITZ, MEMBER, IEEE

***Abstract*—The problem of estimating unknown transfer function parameters from finite input-output records which have been disturbed by additive Gaussian noise with unknown correlation is considered. A rational sampled-data model of preselected order is assumed appropriate, and following the work of Klein,[7] Åström and Bohlin,[10] and Mayne,[11] the likelihood function is generated from the data by numerical filtering. The maximum likelihood criterion leads to nonlinear regression equations for the unknown parameters, which are solved by damped Gauss–Newton iteration. Some computational experiments are described.**

I. Introduction

Many authors have considered the identification problem. Kalman[1] suggested the use of a linear pulse transfer function model and the estimation of its parameters, the coefficients of the transfer function, by linear regression applied to available input–output records. Though numerically simple, the method leads to biased estimates of the denominator coefficients and, therefore, of the system poles, when additive measurement noise is present at the output. The bias occurs because the method minimizes a linear equation error, rather than the difference between the model and the observed outputs. An iterative method has been proposed[2] which does minimize the model plant output error and has, in the presence of uncorrelated Gaussian noise at the output, the probabilistic interpretation of maximum likelihood estimation.[19] Giese and McGhee[3] considered the problem of parameter estimation in the presence of noise with known correlation and showed that the maximum likelihood criterion leads to a weighted least-squares estimation problem. In a similar vein, Levin[4] reduced this weighted least-squares estimation problem

Manuscript received November 17, 1966; revised May 23, 1967. This work was supported by U S Army Research Office, Durham, N. C., under Contract DA-31-124-ARO-D-292. The computer facilities used were supported in part by NSF Grant GP-579. This paper was presented at the 1967 IEEE International Convention, New York, N. Y., March 1967.
A. E. Rogers is with the Dept. of Statistics and Computer Science, University of Delaware, Newark, Dela. 19711.
K. Steiglitz is with the Dept. of Electrical Engineering, Princeton University, Princeton, N. J. 08540.

to an eigenvector problem but, in order to obtain an explicit solution, relaxed certain constraints.[5] Lee[6] presented another least-squares method but, in order to achieve independent measurement errors, some of the information present in the observation records was wasted.

In an econometric context Klein[7] pointed out that true maximum likelihood estimation of transfer function parameters may lead to intractable nonlinear equations. He considered a single-pole transfer function with an additional parameter included to account for unknown serial correlation of the measurement noise. Åström[8] considered process identification when the input signal is unobserved, a problem equivalent to identifying unknown power spectrum parameters.[9] Åström and Bohlin[10] and Mayne[11] discussed parameter estimation from input–output records where the output record contains independent Gaussian noise of unknown spectrum. They observed that the likelihood function of all unknown parameters can be generated by suitably filtering the available data through model filters. Here a similar technique is described; however, it is applied to a transfer function model different from that employed in Åström and Bohlin[10] and Mayne.[11] In addition, the results of computational experiments are presented.

Maximum likelihood estimation of transfer function parameters is discussed under the following conditions.

1) The model transfer function is a rational function of z of preselected order.

2) The estimation uses finite, uninterrupted, normal operating records of one input along with a corresponding output signal.

3) All effects in the output signal, which are not attributable to the input signal, are represented by additive Gaussian noise with a spectral density which, though unknown, is rational and of selected order.

II. Generating the Likelihood Function

The situation and nomenclature are shown in Fig. 1. Use z-transform notation and write $X(z)=\sum x_i z^{-i}$, etc., where summations are over the available records. The model transfer function has $K-1$ zeros and L poles, so that

$$A(z) = \sum_{i=1}^{K} a_i z^{-(i-1)} \qquad B(z) = 1 + \sum_{i=1}^{L} b_i z^{-i}. \tag{1}$$

The "whitening" filter is assumed to have M poles and N zeros, so that

$$C(z) = 1 + \sum_{i=1}^{M} c_i z^{-i} \qquad D(z) = 1 + \sum_{i=1}^{N} d_i z^{-i}. \tag{2}$$

The coefficients of $C(z)$ and $D(z)$ are unknown and are estimated along with those of $A(z)$ and $B(z)$, requiring altogether $P=K+L+M+N$ parameters to be estimated. They form the P vector $\boldsymbol{\theta}$.

If $A(z)$ and $B(z)$ were known exactly, so that the model truly represented the plant, then the model-plant difference signal g would be correlated Gaussian noise. If, in addition, the whitening filter is correct, then the filtered error signal e is uncorrelated Gaussian noise. Hence, the likelihood function for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = (2\pi\lambda^2)^{-T/2} \exp\left(-\frac{1}{2\lambda^2}\sum_i e_i^2(\boldsymbol{\theta})\right) \tag{3}$$

where T is the record length and λ^2 the noise variance. Maximum likelihood estimates of $\boldsymbol{\theta}$ and λ are thus given by solving[19]

$$\sum_i e_i^2(\boldsymbol{\theta}) = \min \qquad \lambda^2 = \min\left(\frac{1}{T}\sum_i e_i^2(\boldsymbol{\theta})\right). \tag{4}$$

If the problem is considered, in z-transform notation the task reduces to finding that $\boldsymbol{\theta}$ which minimizes the average-square value of the signal

$$E(z) = \frac{D(z)}{C(z)}\left[\frac{A(z)}{B(z)}X(z) - W(z)\right] \tag{5}$$

which is a nonlinear numerical problem.

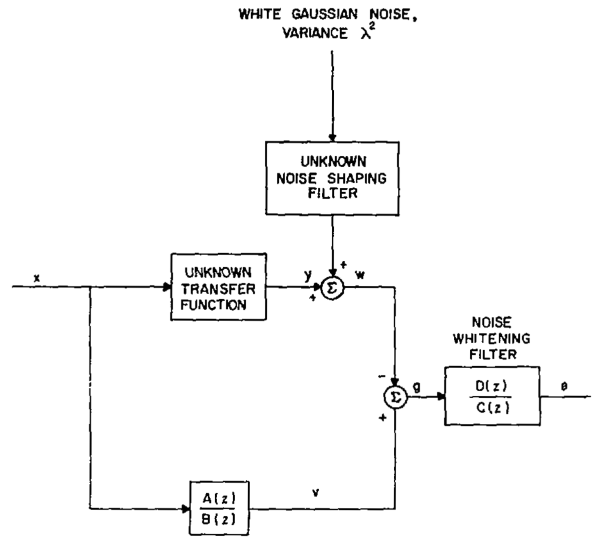

Fig. 1. Block diagram of the parameter estimation problem.

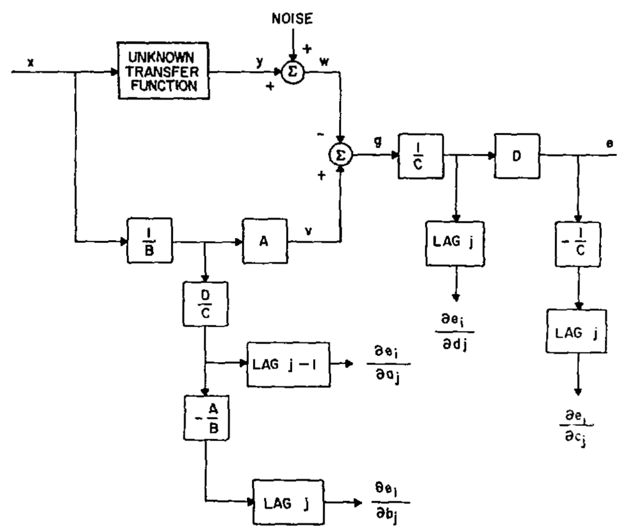

Fig. 2. Generation of the derivative signals by filtering.

III. Numerical Procedure

The Gauss–Newton method[3],[8],[15],[16] has second-order convergence characteristics although it requires the computation of only first-order derivatives. However, it can be unstable, even locally.[16],[17] Hartley's modification[15] employs at each step a one-dimensional search for a minimum and is shown to converge on a quadratic surface. It involves considerably more program complexity than a fixed step-size algorithm; for this reason, a "damped" Gauss–Newton method[17] is used, with the iterative parameter increment

$$\Delta\boldsymbol{\theta} = -k\left[\sum_i \left(\frac{\partial e_i}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial e_i}{\partial \boldsymbol{\theta}}\right)'\right]^{-1} \operatorname{grad}\sum_i e_i^2 \tag{6}$$

where k is a step-size parameter which remains fixed during an iteration. Effective values of k can be established empirically for a class of problems. In the authors' experience, values of the order of 0.2 have resulted in good convergence characteristics for starting points quite far from the final solution. A test for stability of the model at each step was added to the program so that, whenever the model becomes unstable, the step size for the denominator parameters is halved until a stable model is obtained.

The derivative signals $\partial e(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed more efficiently and more accurately by using parameter influence filters,[12]–[14] rather than by the method of finite differences. The necessary expressions can be obtained by differentiating (5) (see Fig. 2). For example,

$$\frac{\partial E(z)}{\partial b_j} = -\frac{D(z)A(z)z^{-j}}{C(z)B^2(z)} X(z), \qquad j = 1, \cdots, L, \text{ etc.} \qquad (7)$$

IV. The Difficulty of Multiple Solutions

Iterative methods locate local maxima of the likelihood function, but do not indicate when a local maximum is the global maximum. There does not, in fact, appear to be any method for establishing with certainty that the global maximum of the likelihood function has been found. Several strategies have been suggested, however, which do appear to increase the probability of finding the global maximum. Among these are the following:

1) Randomly search the parameter space and use as a starting point for the iteration the point corresponding to the greatest value of the likelihood function.[8]

2) Start the iteration at many randomly selected points.[8]

3) Use a suboptimal but consistent estimate as a starting point for the iteration.[20]

4) Having found a local maximum of the likelihood function, search the parameter space for a point which yields a higher value of the function.[21]

The difficulty remains a major theoretical obstacle in using the maximum likelihood criterion when it leads to nonlinear equations in the parameters. In practice, it may not be troublesome because there may be only a small number of local maxima and a fairly large region from which the iteration will converge to the global maximum.

Mantey[18] gives a simple example of a surface with two local minima. The problem he describes is that of fitting in a least-squares sense the impulse response of a second-order rational digital filter with a first-order rational model. In the next section an example which occurred in an experiment with random data is given.

V. Results of Computer Experiments

Experiments were conducted for several different combinations of transfer functions and noise shaping filters on the IBM 7094. Typically, when estimating 5 parameters from records containing 100 data points, 7-place convergence was obtained within 30 to 60 iterations, or 10 to 20 seconds. Occasionally, iterations which were started from different points produced different local maxima of the likelihood function. However, from having solved problems with similar data, it was usually clear when a maximum was not global.

As an example of the experiments a specific case is described. To provide the data that might represent a process input–output record, 100 Gaussian independent pseudorandom numbers were generated for the input signal x. To produce y, these were passed through the filter

$$\frac{A(z)}{B(z)} = \frac{1 + 0.5z^{-1}}{1 - 1.8z^{-1} + 0.85z^{-2}}.$$

Correlated noise was represented by passing 100 more Gaussian independent pseudorandom numbers through the noise shaping filter

$$\frac{C(z)}{D(z)} = \frac{1}{1 - 0.95z^{-1}}.$$

The correlated noise was scaled in amplitude so that, when added to y, the noise-to-signal power ratio was 0.2.

For purposes of comparison, two different models were used, referred to as model A and model B. Model A did not have a noise whitening filter and had $K=2$, $L=2$, $M=0$, and $N=0$. It therefore, provided parameter estimates that minimized the average-square difference between model and observed outputs. Model B included a noise whitening filter and had $K=2$, $L=2$, $M=0$, and $N=1$. In both cases the order of the model transfer function was selected from a priori knowledge of the process.

Using the same sequence of random numbers for x but different noise samples, 12 runs were made. Figs. 3 and 4 show scatter diagrams of b_1 versus b_2, and a_1 versus a_2. As expected, model B, corresponding to maximum likelihood estimation, provided more accurate values on the average. The sample root-average-square distances from the nominal parameter values are

	a_1	a_2	b_1	b_2
Model A	0.3760	0.4937	0.0556	0.0579
Model B	0.1179	0.1311	0.0239	0.0240

Run 2 with model B provided an example of a problem with two local maxima of the likelihood function. The values are

	a_1	a_2	b_1	b_2	d_1	Σe^2
Solution 1	1.187	0.4304	−1.812	0.8622	−0.7968	0.06707
Solution 2	0.9862	1.540	−0.1503	−0.7993	−0.9493	0.1205.

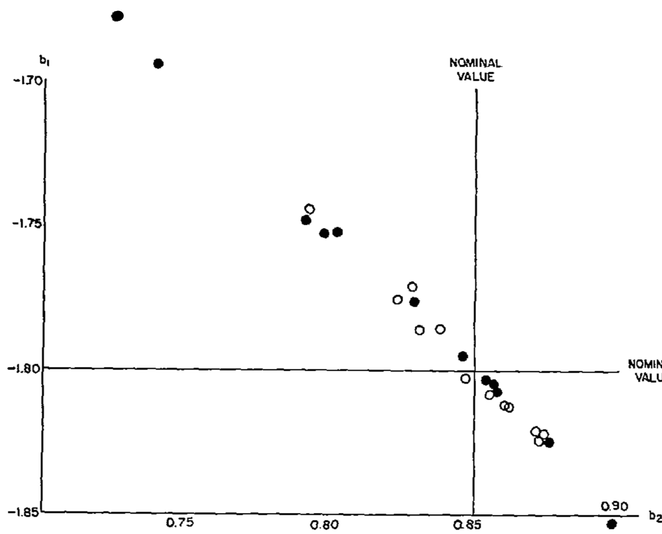

Fig. 3. Scatter diagram of b_1 versus b_2 for the computational experiment. ●—model A, ○—model B.

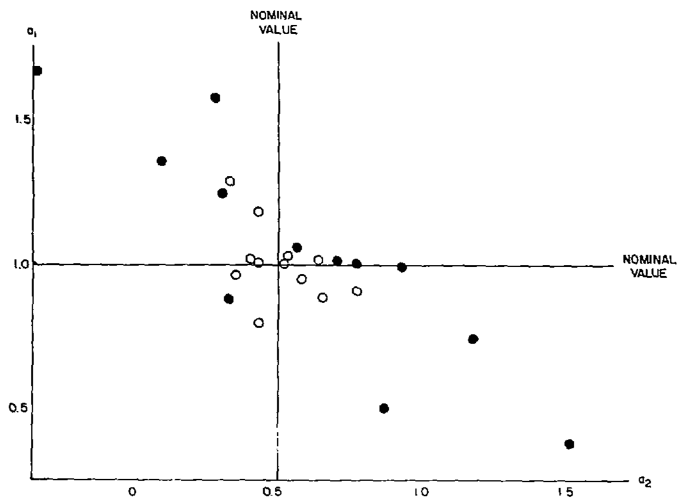

Fig. 4. Scatter diagram of a_1 versus a_2 for the computational experiment. ●—model A, ○—model B.

VI. Conclusions

A method for the maximum likelihood estimation of the parameters of a rational transfer function when the measurement noise is Gaussian and additive at the output, with unknown correlation has been described. It is necessary to estimate the parameters of the noise spectral density along with those of the transfer function. The problem of multiple maxima of the likelihood function stands in the way of developing a completely automatic algorithm, although the method described in this paper appears quite practical.

ACKNOWLEDGMENT

The authors are indebted to Dr. L. E. McBride, and Profs. L. R. Klein and P. J. Dhrymes for many valuable discussions relating to this topic, and to Prof. A. R. M. Noton for bringing to their attention the work of Åström and Bohlin[10] which closely parallels this work.

REFERENCES

[1] R. E. Kalman, "Design of a self-optimizaing control system," *Trans. ASME, J. Basic Engrg.*, ser. D, vol. 80, pp. 468–478, February 1958.

[2] K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems," *IEEE Trans. Automatic Control (Short Papers)*, vol. AC-10, pp. 461–464, October 1965.

[3] C. Giese and R. B. McGhee, "Estimation of nonlinear system states and parameters by regression methods," Preprints, Joint Automatic Control Conf. (Troy, N. Y., June 1965), pp. 46–53.

[4] M. J. Levin, "Estimation of a system pulse transfer function in the presence of noise," *IEEE Trans. Automatic Control*, vol. AC-9, pp. 229–235, July 1964.

[5] L. E. McBride and M. J. Levin, "Discussion of 'Estimation of a system pulse transfer function in the presence of noise'," *IEEE Trans. Automatic Control (Correspondence)*, vol. AC-10, pp. 214–215, April 1965.

[6] R. C. Lee, *Optimal Estimation, Identification, and Control*, Research Monograph 28. Cambridge, Mass.: M.I.T. Press, 1964, sec. 4.32.

[7] L. R. Klein, "The estimation of distributed lags," *Econometrica*, vol. 26, pp. 553–565, 1958.

[8] K. J. Åström, "Control problems in papermaking," *Proc. IBM Sci. Computing Symp.—Control Theory and Applications* (October 1964), pp. 135–167.

[9] S. A. Tretter and K. Steiglitz, "Power spectrum identification in terms of rational models," Preprints, Joint Automatic Control Conf. (Seattle, Wash., August 1966), pp. 720–727.

[10] K. J. Åström and T. Bohlin, "Numerical identification of linear dynamic systems from normal operating records," *Proc. Internat'l Fed. Automatic Control Symp. in Adaptive Control* (Teddington, Middlesex, England, 1965).

[11] D. Q. Mayne, "Parameter estimation," *Automatica*, vol. 3, pp. 245–255, 1966.

[12] H. F. Meissinger, "The use of parameter influence coefficients in computer analysis of dynamic systems," *Proc. Western Joint Computer Conf.* (May 1960), pp. 181–192.

[13] P. Kokotovic, "Simultaneous computation of sensitivity functions," Pupin Control Lab. Rept., July 1962; also in *Sensitivity Analysis of Dynamic Systems*, R. Tomovic, Ed. New York: McGraw-Hill, 1965, pp. 57–78.

[14] K. S. Narendra and L. E. McBride, "Multiparameter self-optimizing systems using correlation techniques," *IEEE Trans. Automatic Control*, vol. AC-9, pp. 31–38, January 1964.

[15] H. O. Hartley, "The modified Gauss-Newton method for the fitting of nonlinear regression functions by least squares," *Technometrics*, vol. 3, pp. 269–280, May 1961.

[16] G. A. Bekey and R. B. McGhee, "Gradient methods for the optimization of dynamic system parameters by hybrid computation," in *Computing Methods in Optimization Problems*, Balakrishman and Neustadt, Eds. New York: Academic Press, 1964, pp. 305–327.

[17] L. E. McBride and K. Steiglitz, "Iterative methods for systems identification," Communications Lab., Dept. of Electrical Engineering, Princeton University, Princeton, N. J., Tech. Rept. 15, June 1966.

[18] P. E. Mantey, "Convergent automatic-synthesis procedures for sampled-data networks, with feedback," Systems Theory Lab., Stanford Electronics Labs., Stanford University, Stanford, Calif., Tech. Rept. 6773-1, October 1964.

[19] H. Cramér, *Mathematical Methods of Statistics*. Princeton, N. J.: Princeton University Press, 1946.

[20] L. R. Klein, private communication.

[21] M. D. Godfrey, private communication.

Input Selection for Parameter Identification in Discrete Systems

ROBERT M. GAGLIARDI, MEMBER, IEEE

***Abstract*—In this paper the problem of selecting an optimal input for identifying an unknown parameter of a known discrete system by observing its output in the presence of Gaussian noise is considered. The system is assumed to be a generalized discrete system in which the inputs and possible parameter values are members of a finite set. The criterion for the optimal input is defined as that which maximizes the probability of correctly determining the true parameter value from a multiple hypothesis test. Although the above criterion totally orders the set of inputs, it is a difficult task to select the best inputs. Some theorems are presented which yield a partial ordering whose extension is the desired total ordering. In the special case of strong noise, it is shown that the ordering of inputs can be related to the perimeter in the output vector space. The results of the paper are applicable to the selection of preset input lengths or to adaptive identification.**

Manuscript received April 15, 1966; revised August 1, 1966, and May 12, 1967. This work was supported by the Joint Services Electronics Program under Contract AF AFOSR-496-64.

The author is with the Dept. of Electrical Engineering, University of Southern California, Los Angeles, Calif.

I. INTRODUCTION

An important problem in system control is that of identifying an unknown parameter inherent to a particular system. When the system input–output characteristics are completely known for every possible parameter value, identification can generally be accomplished by applying some particular input and observing the resulting output. The problem becomes complicated by the presence of noise occurring during the output observation period, and its ability to resolve the parameter values can only be interpreted statistically. In this paper the related problem of selecting the best input for determining which of a finite set of possible parameter values is the true value for a generalized discrete system is investigated. The criterion for best input will herein be interpreted as that which maximizes the probability of correctly identifying the true parameter, based upon noisy output observations. It shall be further assumed that the set of all possible inputs is finite, and that the observable noise is additive and Gaussian. Even though the above criterion totally orders the input set, the probability for a given input is difficult to compute. The objective here is to see if there is an intrinsic property of the optimal inputs that are easier to identify. Once the unknown parameter is identified, within an acceptable probability, the system can be controlled to any desirable output or state by means of a proper subsequent input. Previous work on an analogous type of input selection problem was investigated by Gagliardi[1] using a less meaningful "minimum distance" criterion.

II. SYSTEM MODEL

Let a general discrete system be described by the real difference equation

$$r^{(j)} = F(s_0, u^{(j)}, p), \qquad j = 1, 2, 3, \cdots \tag{1}$$

where $r^{(j)}$ is the system output from the set of reals denoted by R, s_0 is the initial state, $u^{(j)}$ is the input symbol from a set U, p is a system parameter from a set P, and F is a mapping of $s_0 \times U \times P$ into R. An input vector u will correspond to a sequence of input symbols $u = \{u^{(1)}, u^{(2)}, \cdots, u^{(\alpha)}\} \subset U^\alpha$. The system responds to u by yielding the output vector $r = \{r^{(1)}, r^{(2)}, \cdots, r^{(\alpha)}\} \subset R^\alpha$, depending, of course, on s_0 and p. In the following it is assumed that U and P are finite sets, and the parameter p is unknown but remains fixed during the observation period. The observable output is taken to be the sum of r and n where n is an α-dimensional Gaussian noise vector having zero mean and covariance matrix $E(n\, n^T) = \sigma^2 I$, E is the expectation operator, T denotes transpose, I is the $\alpha \times \alpha$ identity matrix, and σ^2 is a scalar. For convenience, we shall consider the observable to be normalized with respect to σ and denote it by y. Thus,

$$y = \frac{r(u)}{\sigma} + n \tag{2}$$

and is an α-dimensional Gaussian vector with a mean of $r(u)/\sigma$ and a covariance matrix I.

In the parameter identification problem, p in (1) must be identified by applying some $u \subset U^\alpha$ and observing y. Since the mapping F is completely known, the objective is to determine the input vector u that best aids in the identification of p. The criterion for "best" is taken as that which maximizes the average probability of correctly determining the true $p \subset P$, where any member of P is equally likely a priori. Note that the basic objective here is to make a multiple hypothesis test on the observable y, using the input u that maximizes the probability of the test being correct. Thus, if $P = \{p_i\}$, $i = 1, 2, \cdots, \beta$, it is desirable to test the hypothesis $p = p_1$ versus $p = p_2$ versus $p = p_3$, etc. Since y is Gaussian and P finite, there is an ample supply of applicable literature[2],[3] describing the optimal test that must be performed on y. In this regard, the test is analogous to a corresponding problem in communication theory, i.e., that of determining which of a set of signals is being received in the presence of Gaussian noise[4] and some of these latter results are applicable here. Trivial situations shall be avoided by assuming that all the parameters of P are observable, i.e., can be identified uniquely with some input u of U^α in the absence of noise.