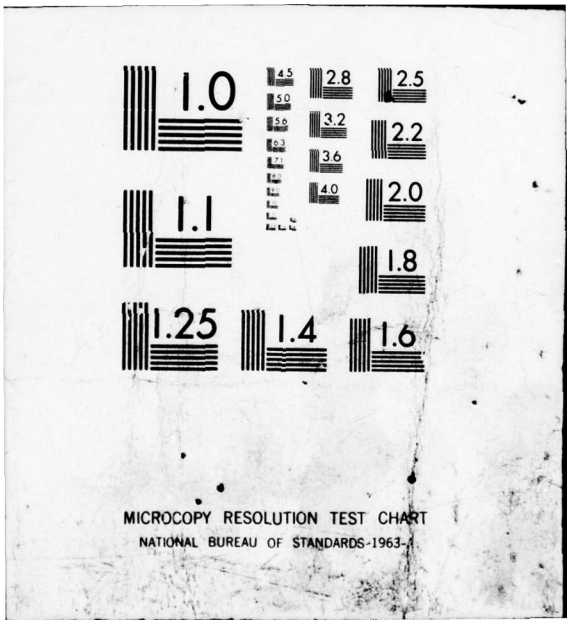
AD-A073 207 WASHINGTON UNIV ST LOUIS MO DEPT OF SYSTEMS SCIENCE --ETC F/G 12/2
DESIGN AND PERFORMANCE EVALUATION FOR SYSTEMS IN AN UNCERTAIN E--ETC(U)
AUG 79 I B RHODES N00014-76-C-0667

UNCLASSIFIED NL

1 OF 2
AD A073207


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

WASHINGTON UNIVERSITY

Department of Systems Science and Mathematics

St. Louis, MO 63130

DESIGN AND PERFORMANCE EVALUATION FOR SYSTEMS IN AN UNCERTAIN ENVIRONMENT

Ian B. Rhodes

Final Report Under Office of Naval Research

Contract N00014-76-C-0667

March 1, 1976 - June 30, 1979

August 23, 1979

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

DESIGN AND PERFORMANCE EVALUATION FOR SYSTEMS IN AN UNCERTAIN ENVIRONMENT.

5. TYPE OF REPORT & PERIOD COVERED

FINAL REPORT.
Mar 1, 1976 - Jun 30, 1979

6. PERFORMING ORG. REPORT NUMBER

1 Mar 76 - 30 Jun 79

7. AUTHOR(s)

Ian B. Rhodes

8. CONTRACT OR GRANT NUMBER(s)

N00014-76-C-0667

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Washington University
Department of Systems Science and Mathematics
St. Louis, MO 53130

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

NR 041-500 (432)

11. CONTROLLING OFFICE NAME AND ADDRESS

Office of Naval Research, Code 432
Mathematical and Information Sciences Division
800 N. Quincy St., Arlington, VA 22217

12. REPORT DATE

August 23, 1979

13. NUMBER OF PAGES

174

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

172p.

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stochastic systems, stochastic processes, filtering, estimation, control, tracking, optimal control, point processes, counting processes, Poisson processes, optical communication, cutoff rate, shortest path problems, decentralized control, optimization, singular estimation, (over)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Included in this report are results of investigations into the following topics: design and performance evaluation of optimal and suboptimal estimation and tracking systems for space-time point-process observations; optimal signal design for coded, direct-detection optimal communication systems; informationally-decentralized shortest path algorithms for networks; singular estimation and control problems; compensator design for polynomial matrix descriptions of linear multivariable systems; a (over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED 409 076
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19. singular control, multi-variable systems, polynomial matrices, innovations, innovations conjecture, controllability, observability, performance evaluation.

20. direct proof of the informational equivalence of the innovations and observations processes for linear estimation or Gaussian processes; and quantitative measures of controllability and observability and their implications in system analysis and performance evaluation.

DESIGN AND PERFORMANCE EVALUATION FOR SYSTEMS IN AN UNCERTAIN ENVIRONMENT

Abstract

Included in this report are results of investigations into the following topics: design and performance evaluation of optimal and suboptimal estimation and tracking systems for space-time point-process observations; optimal signal design for coded, direct-detection optical communication systems; informationally-decentralized shortest path algorithms for networks; singular estimation and control problems; compensator design for polynomial matrix descriptions of linear multivariable systems; a direct proof of the informational equivalence of the innovations and observations processes for linear estimation or Gaussian processes; and quantitative measures of controllability and observability and their implications in system analysis and performance evaluation.

TABLE OF CONTENTS

1. INTRODUCTION

This final report describes the research conducted under Office of Naval Research Contract N00014-76-C-0667 from March 1, 1976 to June 30, 1979.

This research has had as its principal concern the investigation of estimation, decision and control problems for systems operating in an environment of uncertainty, with an increasing emphasis on information-ally-decentralized problems that arise typically in connection with large systems, and particularly C^3-systems, where multiple decision-makers take actions on the basis of their own individual knowledge and data.

Our investigations have been directed both towards finding exact, optimum solutions to various such estimation, decision and control problems and, where these are not available or difficult (or infeasible) to implement, towards providing a basis for designing and assessing the performance of satisfactory suboptimum solutions. An example that encompasses both of these objectives is found in our research, described in Chapter 2, into estimation and tracking problems involving space-time point-process observations. There we derive the optimum estimators and controllers and show them to be nonlinear but finite-dimensional; they are thus implementable, but evaluation of their performance requires infinite-dimensional calculations. We have therefore derived easily-computed upper and lower bounds on the optimum performance; in fact, the upper bounds give the exact performance of a parametrized family of suboptimum designs that are even more-easily implemented than

the optimum, and one of these is identified as providing better performance than any other, thus making it the best design within this class. When significant detector dark current is present, the exact optimum is infeasible to implement, and there is no choice but to examine suboptimum designs.

This example illustrates our view that, because implementable optimum solutions can be found only for relatively few problems, what is frequently needed is a shift in emphasis away from optimal designs and towards implementable designs that achieve satisfactory performance. For this there is needed a setting within which easily-implemented suboptimum designs can be identified, and their performance evaluated through a reasonably complete and computationally-feasible design analysis.

Even so, special cases for which exact, optimum solutions can be found continue to be important, both in their own right and as a basis for suggesting candidate designs for broader classes of problems, and in forming benchmarks with respect to which of these candidate designs can be compared. Problems for which exact solutions have been derived in the course of this research include, in addition to that described above, an optimum signal design problem involving point-process observations; singular estimation and control problems; and shortest path problems with decentralized information and topological requirements.

We now turn to an outline of the contents of the chapters that follow.

In Chapter 2 we review our extended research effort into estimation, detection, and tracking problems involving space-time point-pro-

cess observations. Such problems arise in direct-detection optical communication systems, infrared tracking systems, and star tracking systems, all of which have a requirement for position sensing and active tracking to maintain optical alignment. For these problems we have derived optimal estimation and tracking system designs, and analyzed the performance of these in terms of easily-computed upper and lower bounds. Both the optimum estimator and the optimum controller are of interest in that they are nonlinear but finite-dimensional, and therefore implementable. The upper bounds also give the exact performance of suboptimum estimators and controllers that have certain minimality properties and are even more easily implemented than the corresponding optimum system. The last section of the chapter discusses our recently-begun examination of the case where there is significant detector dark current or significant background radiation, thus superimposing on the problem difficulties akin to those arising in the perhaps more familiar problem of tracking an object in clutter.

Chapter 3 is also concerned with point-process observations, in this case the design of optimum signal waveforms for coded, direct-detection optical communication systems. Significant new results with important practical consequences have followed from this extended project. It is of interest that a modulation scheme we show to be optimum, when an average energy constraint on the transmitted signal is a limiting factor, is the one that has been adopted in the brass-boarded one gigabit-per-second optical communication system currently under development.

Our research into shortest path algorithms with decentralized information and topological requirements is described in Chapter 4. There we present algorithms we have developed that enable each node in a network to calculate its shortest distance to any other node using only local knowledge of the network topology and only local information transfer between adjacent nodes. Shortest path problems arise in many contexts, and algorithms with such decentralized information requirements are of obvious importance in many applications. One area of particular applications interest is that of naval C^3-systems.

These algorithms are based on appropriate modification and reinterpretation of labelling algorithms for shortest path problems in order to extract the desired decentralized properties. All converge in finite time, even if implemented asynchronously. The simplest algorithm was developed with a static network in mind, but it also handles decreasing branch lengths and the introduction of new nodes or branches. Changes are needed, however, to account for increasing branch lengths or the failure of nodes or branches. Three such modifications are presented, each retaining the basic localized information properties. Each has its own characteristics, and its applicability or suitability is a function of the particular network under consideration.

In Chapter 5, we describe the new results we have obtained for singular estimation and control problems. These results have followd from our examination of these problems from a geometric viewpoint, utilizing the ideas first introduced by Wonham and Morse and by Basile and Marro in connection with decoupling and other problems. We show

that certain fundamental subspaces introduced by Wonham and Morse in that context also provide just the right framework for an extremely simple characterization of the solution to the singular estimation or control problem. The solution is as simple for multi-input, multi-output systems as it is for single-input and/or single-output ones, in contrast to available results based on algebraic approaches, where the strongest and simplest solutions are for systems with a single input (in the case of control) or a single output (in the case of estimation). Our geometric characterization reduces easily and directly to known algebraic results in the single-input or single-output case. Both continuous-time and discrete-time problems are included within the same unified development.

Recent years have seen a resurgent interest in frequency-domain methods for the analysis and design of linear multivariable systems, in contrast to the time-domain-based state-space approach that has predominated for the past two decades. These methods are based on polynomial matrix descriptions of multivariable systems and, in fact, have both time-domain and frequency-domain interpretations. Chapter 6 contains a description of our investigation into design methods for systems represented in polynomial matrix form. Our objective has been the development of methods for designing compensators for such systems, and particularly minimum-order compensators. The techniques we have employed draw on the ideas and objects of modern algebra, especially the theory of modules and free modules. The design methods we have developed to date apply primarily to single-output (multi-input) systems, since the simplest available tools from modern algebra turn out to correspond to this

case. It is of interest to observe that this design problem is, for polynomial matrix descriptions, the analog of observer-based compensator design via state-space techniques: there, too, the theory for single-output systems preceded, and is simpler than, that for multi-output systems.

The successful "innovations approach" to estimation rests on the equivalence of the information provided by the innovations process and that provided by the observations process. This is known to hold true in some circumstances, and to fail to hold in others. We have constructed a new, direct proof of its known validity for an important class of linear least-squares estimation problems and problems involving Gaussian processes. This proof is outlined in Chapter 7.

Finally, in Chapter 8, we present some preliminary results of our recently-begun efforts to develop quantitative measures of controllability and observability that have implications in design and performance evaluation methods for large systems. Almost all of linear system theory consists of sharply-defined answers to sharply-posed questions. For example, a system is either controllable or not, or decoupled or not. Disturbances must be completely rejected for disturbance localization to be said to take place. There is no body of theory that allows for approximate achievement of these goals. A disturbance or another input may affect a certain output to an acceptably sight degree, but if the effect is nonzero our present sharp formulations have us conclude that the disturbance is not rejected or that the output is not decoupled from the input. What seems especially needed are quantitative measures of controllability and observability that have implications, in

terms, say, or performance bounds, on such problems of approximate disturbance localization or approximate noninteraction, and on the performance of estimators or controllers designed on a noninteracting basis but implemented on an interacting collection of subsystems. The long-term goal of this research project is the provision of such a framework for decentralized design and performance evaluation.

A number of the chapters that follow are concerned with work that has been already published or is available in a report that has been submitted for publication. In those cases, the appropriate papers or reports are included as appendices, and the presentation is limited to an outline of the results that are established there in detail. Also, in these cases, references are made where possible to the list of references in the relevant appended paper or report.

Chapter 4 was prepared with the assistance of doctoral student Jeffrey M. Abram, and Chapter 6 with the assistance of doctoral student Olive Y. Liu.

2. ESTIMATION AND CONTROL PROBLEMS WITH SPACE-TIME POINT-PROCESS OBSERVATIONS

2.1 Introduction

A major component of our research effort has been concerned with observation models other than the familiar "signal in additive white Gaussian noise" structure. Particular attention has been given to observations that take the form of a doubly-stochastic space-time counting process whose intensity is signal-dependent. Estimation and control problems involving such processes arise in a number of contexts, including quantum-limited optical communication and nuclear medicine. The estimation and tracking problems associated with optical communication systems have been discussed in detail as motivation in the papers that have resulted from this study.

New results of major significance have been obtained for both estimation and control problems involving space-time counting process observations. A sequence of new results culminated in the journal article:

> "Estimation and Control Performance for Space-Time Point-Process Observations," Ian B. Rhodes and Donald L. Snyder, IEEE Transactions on Automatic Control, Vol. AC-22, No.3, June 1977, pp. 338-346.

which is included here as Appendix 3. This paper includes as special cases all earlier results for this class of problems, including our own earlier research under this contract which was reported in the journal article:

> "A Separation Theorem for Stochastic Control Problems with Point Process Observations," D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, Automatica, Vol. 13, No. 1, January 1977, pp. 85-87.

and in the invited conference paper:

> "Estimation and Control Performance for Space-Time Point-Processes," I. B. Rhodes and D. L. Snyder, Proceedings of the Fourteenth Allerton Conference on Circuit and System Theory, University of Illinois, September 1976, pp. 38-51.

These two papers are included here as Appendices 1 and 2.

The next subsection outlines in loose terms the results established in detail in Appendix 3. This is followed by a discussion of our recent efforts to extend these results to the situation where there is significant detector dark current or significant background radiation.

2.2 Summary of Appendix 3

In outline, the paper included as Appendix 3 considers a stochastic system

$$dx_t = F(t)\, x_t dt + G(t)\, u_t\, dt + V(t) dv_t$$

$$dz_t = C(t)\, x_t\, dt + dw_t$$

where u_t is a control variable, v and w are Wiener processes, and the usual assumptions (detailed in the paper) are made. In addition to the observation process z, we assume additional observations of a space-time point process N(t,r) in which each point occurrence has both a temporal coordinate t and a spatial location r. In an optical communication setting, this point process might be thought of as a model for photoelectron conversions on a detector surface, a particular point occurrence corresponding to a conversion taking place at time t and at location r on the detector surface. Associated with N(t,r) is a counting process N_t which simply counts point occurrences regardless of their spatial lo-

cations; N_t is assumed to be a doubly-stochastic process with stochastic intensity μ_t. Given that a point occurrence has taken place, its spatial location r is taken to be a Gaussian random vector with mean $H(t)x_t$ and covariance R. In terms of the photoelectron conversion model mentioned above, the dependence of r on the system state x_t reflects the (random) movement of the center of the incident beam due to vibration, beam steering due to atmospheric turbulence, the motion of the tracking system, etc. The control u_t represents the input to the tracking system, which is included as part of the total state x_t. The randomness of the temporal intensity μ_t includes the transmission of information by modulating μ, as well as randomness due to such effects as fading during propagation.

For this model, we have examined

a) <u>the estimation problem</u> of finding the conditional density of the processes (x_t,μ_t) at time t given observations of both z and the space-time point-process up to time t, and especially to find the associated conditional means and covariances.

b) <u>the control problem</u> of finding the control u_t that depends at most on the past of the space-time point-process and z and minimizes

$$J[u] = E\left\{ \int_0^T [u_t'P(t)u_t + x_t'Q(t)x_t]\, dt + x_T'Sx_T \right\} .$$

Precise statements of these problems, their solutions and some attendant technical assumptions are given in the paper. In simple terms, the resuls we establish there are:

(i) Under assumptions that are reasonable from a practical viewpoint, the joint problem of estimating both the temporal intensity μ_t and the state x_t reduces to two separate problems, one of estimating μ_t from just the temporal component N_t of the space-time point process, and the other of estimating x_t using all available observations. In terms of the optical communication problem, this is of great practical importance since it establishes that demodulation or detection can be carried out independently of tracking (provided, of course, optical boresight is maintained).

The demodulation or detection problem of estimating μ_t from the temporal component N_t of the space-time point process is a standard one that has been solved under a variety of assumptions on μ in the book by Snyder [Ref. 6 in Appendix 3].

We show in Appendix 3 that the conditional density of x_t given all observations up to time t is Gaussian. Furthermore, the conditional mean and covariance satisfy a pair of <u>finite-dimensional</u>, nonlinear stochastic differential equations (see eqs (6) - (8) in Appendix 3). It should be emphasized that although the optimum estimator is <u>nonlinear</u> it is <u>finite-dimensional</u> and therefore implementable in practice.

(ii) The solution to the control problem satisfies a separation theorem analogous to the standard linear-quadratic-Gaussian separation theorem of linear system theory, i.e. the optimum tracking controller separates into two separate and indepen-

dent components: an estimator and a control law. The estimator is the finite-dimensional, nonlinear one described above, while the control law is the certainty-equivalent linear one. Being finite-dimensional, the controller can be easily implemented in practice. This separation theorem is important both theoretically and practically. From a theoretical standpoint, it seems to be the only case beside the standard LQG result where separation of a dual-control problem into two independent problems has been established. Not only is this an important exact result in its own right, but it has the as-yet uninvestigated potential of forming a benchmark for designing and assessing the performance of suboptimal controllers in wider situations, much as we have previously used the standard LQG result to obtain bounds for incrementally conic nonlinear systems. From a practical viewpoint, the separation theorem for space-time point-process observations provides the simple, optimum design for an important class of tracking and other problems.

(iii) Although the optimum estimator is finite-dimensional, its error covariance depends on the point process occurrence times and is thus a random process that is <u>not</u> precalculabe (in contrast to the deterministic, precalculable error covariance of the standard Kalman filter). A natural measure of estimator performance, which also turns out to determine controller performance, then becomes the <u>expected</u> error covariance; however,

although this is deterministic, its calculation is infinite-dimensional. We have, therefore, derived easily-computed upper and lower bounds on the expected error covariance and corresponding bounds on the optimum controller performance. The upper bounds are derived by evaluating exactly the performance of a parametrized family of suboptimum designs; one of these is identified as having smaller performance than any other, thus providing a minimal upper bound within this family. The bound-minimal estimator and controller are thus natural candidates for designs that are even more simply implemented than the optimum, in that they require less on-line computation because the gain coefficients are deterministic and precalculable rather than stochastic and dependent upon the particular realization of the counting process, N_t, as they are in the optimum estimator.

2.3 Extensions When Detector Dark Current is Present

Both estimation and tracking problems are greatly complicated when there is significant detector dark current (or background radiation). This is because of the uncertainty that then exists as to whether an observed point in space-time is due to the signal process or to the dark current. In this respect, the principal difficulties that arise are conceptually similar to those in the more familiar problem of tracking an object in clutter, where again uncertainty exists as to whether an observation corresponds to the object being tracked or to the clutter. A recent summary of the problem of tracking in clutter can be found in [1].

In either case, a solution to the estimation problem can be obtained in principle by constructing a bank of estimators that expands geometrically with successive observations, and appropriately weighting the outputs of these to obtain the conditional mean. In our case, this means that after N point occurrences in space-time, there will be required a bank of 2^N estimators of the type given in Appendix 3, each estimator corresponding to one of the 2^N possible hypotheses as to which points in the observed sequence are due to the signal and which to the dark current. For each such hypothesis, the corresponding estimator satisfies the equations (6) - (8) in Appendix 3, but including only those observation points hypothesized as being due to the signal, and neglecting those hypothesized as being due to the dark current. The state $\hat{x}_{it}$ of the i-th estimator, corresponding to hypothesis H_i as to which observation points are due to the signal and which to dark current, is the conditional mean of the state x_t given <u>both</u> all observed data to time t, <u>and</u> that hypothesis H_i holds. The conditional mean $\hat{x}_t$ of the state is then found as the linear combination

$$\hat{x}_t = \sum_{i=1}^{2^N} p_{it} \hat{x}_{it}$$

where p_{it} is the conditional probability that hypothesis H_i is true given all observation data up to time t. Equations for the p_{it} can be developed under various sets of assumptions on the dark current process. One simple possibility is to assume that the time component of the dark current process is Poisson with rate ν and independent of the signal process, and that, given a dark current point occurrence has taken

place, its spatial location is uniformly distributed over the detector area A (assumed large compared with the covariance of the signal-induced spatial distribution) and is independent of the spatial locations of prior and succeeding points. Even then, the equations for the p_{it} become unwieldy, and they are not given here because, in any event, the requirement of constructing such a rapidly-expanding bank of filters makes this solution impractical in almost any conceivable application, i.e. unless very few point occurrences are expected to take place.

One is, therefore, led to seek more-readily-implemented suboptimal estimators. We have taken the approach of investigating estimators whose dimension is that of the system state x_t, thus bypassing at the outset the expanding-state requirement of the optimum estimator. We adopt the same notation and the same models for the signal and the signal-induced space-time point process as in our paper [Appendix 3], and assume that the dark current satisfies the assumptions given towards the end of the preceding paragraph. Let the first observed point be at time t and at location r. Over [0,t) the optimum estimator is n-dimensional and satisfies eq. (6) in Appendix 3 with the last term identically zero since no points have yet occurred; indeed at time t- the conditional density of the state given the observations is $G(\hat{x}_{t-},\Sigma_{t-})$ i.e. Gaussian with mean $\hat{x}_{t-}$ and covariance Σ_{t-} given via eqs. (6) and (7) in Appendix 3, with, in both cases, the last term identically zero. Under the hypothesis that the observed point is due to the signal, the conditional density of x at t+ is, from eqs. (6) - (9) in Appendix 3, $G(\bar{x}_{t+},\bar{\Sigma}_{t+})$ with

$$\bar{x}_{t+} = \hat{x}_{t-} + \Sigma_{t-}H'[H\Sigma_{t-}H' + R]^{-1}(r-H\hat{x}_{t-})$$

and

$$\bar{\Sigma}_{t+} = \Sigma_{t-} - \Sigma_{t-}H' [H\Sigma_{t-}H' + R]^{-1}H\Sigma_{t-}$$

Under the assumption that the observed point is due to dark current, the conditional density of x at t+ is the same as at t-, viz. $G(\hat{x}_{t-}, \Sigma_{t-})$. It then follows that the conditional density of x at t+ given data to t+, including the observed point, is the convex sum of Gaussian distributions

$$f_{t+} = f(x_{t+}|\text{data to } t+) = p_t G(\bar{x}_{t+},\bar{\Sigma}_{t+}) + (1-p_t)G(\hat{x}_{t-},\Sigma_{t-})$$

$$= G(\hat{x}_{t-},\Sigma_{t-}) + p_t[G(\bar{x}_{t+},\bar{\Sigma}_{t+}) - G(\hat{x}_{t-},\Sigma_{t-})]$$

where p_t is the conditional probability that the observed point at time t is due to the signal. Various equivalent expressions can be given for p_t; one is

$$p_t = [1 + (n/s)^* \exp\tfrac{1}{2}\rho^2]^{-1}$$

where

$$\rho^2 = [r - H\hat{x}_{t-}]'[R + H\Sigma_{t-}H']^{-1}[r - H\hat{x}_{t-}],$$

$$(n/s)^* = (v/A)\Big/(\mu/(2\pi)^{m/2}\det{}^{1/2}(R + H\Sigma_{t-}H')),$$

and we assume for simplicity that the temporal intensity μ_t of the space time point process is constant. It then follows from straightforward calculations that the conditional mean and covariance of x at time t+ given data to t+ are, respectively,

$$\hat{x}_{t+} = p_t \bar{x}_{t+} + (1 - p_t) \hat{x}_{t-}$$

$$\Sigma_{t+} = \Sigma_{t-} - p_t \Sigma_{t-} H' [H\Sigma_{t-}H' + R]^{-1} H\Sigma_{t-} + p_t(1 - p_t)(\bar{x}_{t+} - \hat{x}_{t-})(\bar{x}_{t+} - \hat{x}_{t-})'$$

Although these expressions give the exact conditional mean and covariance immediately following the first point observation, the conditional density is <u>not</u> Gaussian but, rather, the convex combination of Gaussian distributions given above. Thus, in contrast to the situation that obtains in the absence of dark current, conditional Gaussian-ness is not maintained across the first occurrence point, and this procedure cannot be repeated through succeeding observation points. Indeed, after the N-th observation point the conditional density is a convex combination of 2^N Gaussian densities, and it is the generation of these densities that reflects in the 2^N estimators required in the exact solution.

On the other hand, one natural approach to maintaining an n-dimensional filter is to approximate the conditional density f_{t+} following the first observation point by a Gaussian density with the same mean $\hat{x}_{t+}$ and covariance Σ_{t+} given above. This Gaussian approximation then remains Gaussian as it is propagated to just before the next occurrence point using eqs. (6) -(9) in Appendix 3. The above procedure is re-

peated to incorporate the new space-time data point, and the process is repeated.

Evaluation of the performance of this suboptimum estimator is difficult because the resulting equation for the mean-square-error, Σ, depends not only on the occurrence times (as it does in the dark-current-free problem in Appendix 3) but also on the spatial locations of the observation points through both p_t and $\bar{x}_{t+}$. This nonlinear dependence on the spatial locations as well as on the occurrence times greatly complicates an analysis in terms of bounds comparable to that performed in Appendix 3 for the dark-current-free case.

We have begun to investigate parameterized families of suboptimum estimators in which p_t is restricted to being dependent only upon r in a simple way, in combination with the suboptimal estimator eq. (16) in Appendix 3. This means that the family of suboptimum estimators (16) in Appendix 3 is modified to become

$$dx_t^{\#} = Fx_t^{\#}dt + Gu_t dt + L(t)\,[dz_t - Cx_t^{\#}dt] + \int_A p(r)\,M(t)\,[r - Hx_t^{\#}]\,N(dt \times dr)$$

One possibility is to restrict p(r) to being, say,

$$p(r) = \begin{cases} 1 & [r - Hx_t^{\#}]'\,[R + H\Sigma^{\#}H']^{-1}[r - Hx_t^{\#}] \le \alpha \\ 0 & \text{otherwise} \end{cases}$$

where α is a parameter to be chosen and $\Sigma^{\#}$ is the error covariance associated with $x^{\#}$. In simple terms, this means that if r is "sufficiently close" to its expected location $Hx_t^{\#}$, "sufficiently close" being determined by the parameter α, then the point occurrence will be taken as being due to the signal; otherwise, it will be neglected as being due to the dark current. Our objective is to find choices of the gains L(t) and M(t) and of the parameter α that are in some sense optimum, such as leading to a minimum error covariance $\Sigma^{\#}$. Our investigation of this problem is continuing. We remark that a much simpler version of this problem has been examined by simulation in [2], where a simpler criterion for accepting or rejecting points as being due to the signal is employed.

3. OPTIMUM SIGNAL DESIGN FOR CODED, DIRECT-DETECTION, OPTICAL COMMUNICATION SYSTEMS

Important new results have followed from our new approach to the coordinated design of the encoder, optical modulator and demodulator for a digital communication system employing an optical carrier and direct detection. These results are contained in the revised report:

> "Some Implications of the Cutoff-Rate Criterion for Coded, Direct-Detection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Biomedical Computer Laboratory Monograph 363, Washington University, St. Louis, MO, March 1979,

which is included as Appendix 4 and has been submitted for publication in the IEEE Transactions on Information Theory. Individual results from this comprehensive report [Appendix 4] have been presented at two conferences and one workshop, and another conference presentation will take place later this year:

> "Signal Optimization for Random Point Processes," D. L. Snyder and Ian B. Rhodes, AFOSR Workshop in Communication Theory and Applications, Provincetown, Massachusetts, September 17-20, 1978.
>
> "Quantization Loss in Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Sixteenth Allerton Conference on Communication, Control, and Computing, University of Illinois, October 4-6, 1978.
>
> "Some Implications of the Cutoff Rate Criterion for Coded, DirectDetection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, 1979 IEEE International Information Theory Symposium, Grignano, Italy, June 25-29, 1979.

"Quaternary Pulse Modulation is Optimal for Optical Communication at One Gigabit Per Second," Donald L. Snyder and Ian B. Rhodes, National Telecommunications Conference, Washington, DC, November 27-29, 1979.

Because Appendix 4 provides a complete account of the comprehensive collection of results obtained in the course of this extended research effort, we limit ourselves here to a very brief outline of the principal conclusions.

The basis of our new approach has been the reformulation of this signal design problem to use the cutoff rate as a performance measure instead of the usually-employed probability of error. The use of cutoff rate as a design criterion has been eloquently and persuasively argued by Massey in his apparently little-noticed 1974 conference paper [Ref. 8 in Appendix 4]. In this paper he also examined the additive white Gaussian noise channel and was able to prove for the first time a long-standing conjecture on the optimality of a simplex signal set.

We have derived the cutoff rate for a digital communication system employing an optical carrier and direct detection, and we have used this as the performance measure in studying the coordinated design of the optical modulator and demodulator. The choice of modulation that maximizes the cutoff rate has been derived for various relationships between peak amplitude and average energy constraints on the transmitted optical signal, and found to be:

(i) When the average energy constraint is predominant, pulse position modulation is found to be optimum.

(ii) When the peak amplitude constraint predominates, Hadamard matrices can be used to define an optimum choice of modulation.

(iii) When neither constraint predominates, appropriate time sharing of the solutions given in (i) and (ii) above is optimum.

We have also addressed within this framework problems of efficient energy utilization, the choice of input and output alphabet dimensions, and the effect of random detector gain.

Corresponding results are also shown to hold when polarization modulation is employed in the optical modulator as well as temporal modulation. Specifically, for an input alphabet of dimension 4, the optimal modulation when average signal energy constraints predominate employs binary pulse-position and binary polarization modulation; it is of interest to note that such a modulation scheme has been adopted in the one gigabit per second satellite optical communication system reported by Ross et al. in [Ref. 15 of Appendix 4].

4. INFORMATIONALLY-DECENTRALIZED NETWORK PROBLEMS

4.1 Introduction

A major effort has been concentrated on developing shortest path algorithms that enable each node in a network to calculate its shortest distance to any other node using only local knowledge of the network topology and only local information transfer between adjacent nodes. The requirement that information transfer and topological information be localized contrasts sharply with the global information that is required by almost all of the many existing shortest path algorithms; the implementation of these algorithms can be thought of as requiring each node to transmit distance and topology information to a central controller, who is then responsible for solving the problem and sending the appropriate optimal routing information to each of the nodes. In a large network this could involve a significant amount of communication. Additionally, for some networks establishment of a central controller may be expensive, infeasible, or undesirable from a security or reliability viewpoint.

Shortest path problems arise in many contexts, and algorithms with decentralized topological and information transfer requirements are of obvious importance in many applications. In addition to the traditional applications areas, an area of particular applications interest is that of naval C^3-systems, and an algorithm we have developed was presented at the First MIT/ESL-ONR Workshop on Distributed Communication and Decision Problems Motivated by Naval C^3-Systems held at MIT in August, 1978. A more detailed account of this algorithm appears in the conference

paper:

> "A Decentralized Shortest Path Algorithm," Jeffrey M. Abram and Ian B. Rhodes, Proceedings of the Sixteenth Allerton Conference on Communications, Control and Computing, University of Illinois, October 4-6, 1978, pp. 271-277,

which is included here as Appendix 5.

This algorithm was initially developed for a static network in which branch lengths and topology remain constant, though it can accommodate limited changes. We have subsequently made a number of modifications to the algorithm to enable it to operate in a dynamic network in which branch lengths can increase or decrease, and nodes or branches can be added to or removed from the network. The ability of an algorithm to handle such topological changes is essential in most practical applications, including especially those arising in connection with C^3-systems.

A brief outline of the algorithm described in Appendix 5 is given in the next section. This is followed by a description of several modifications of this algorithm to accommodate various types of changes in a dynamic network.

4.2 The Static Algorithm

Consider a directed graph consisting of N nodes, denoted $\{1,2,\ldots,N\}$, and a collection of branches (links), $A=\{(i,j): i,j \in N$ and there exists a branch from i to $j\}$. To each branch $(i,j)\in A$ is associated a length s_{ij}. The lengths are unrestricted in sign, but the sum of the lengths in any closed loop of the network is assumed to be positive.

Very little topological information is needed. Each node needs to know only which of its neighbors are attached to incoming branches, which are attached to outgoing ones, and the lengths of the branches to the outgoing neighbors. For each ultimate destination, a node calculates and stores an assessment of the shortest path via each of its outgoing links; the smallest of these is taken to be its assessment of the shortest path to that destination and is subsequently referred to as the current shortest distance. Also stored is the identity of the outgoing neighbor which achieves this minimal distance. Initially, the current shortest distance is taken to be: for ultimate destinations that are neighboring nodes, the corresponding outgoing branch length; for all other destinations, infinity.

Whenever a node's current shortest distance to a destination changes, either through reinitialization or new information received from a neighbor, this new distance is transmitted to all incoming neighbors. At the conclusion of the algorithm, each node will know the shortest distance to each other node (or that no path exists), the next node in the path that achieves this distance, and the shortest distance via each alternative outgoing node. The algorithm is guaranteed to converge, even if it is implemented in an asynchronous manner.

4.3 Dynamic Network Algorithms

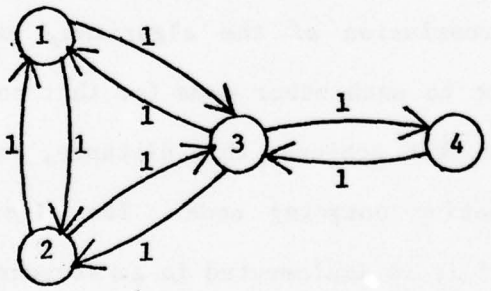
While the above algorithm handles static networks, many problems arise for a dynamic network model. The phenomena that we have investigated include branch lengths decreasing and increasing, branches being introduced into the network, and branches failing or being removed from

the network. It should be noted that, in principle, a node coming up or going down can be treated as the incident set of branches simultaneously doing the same thing.

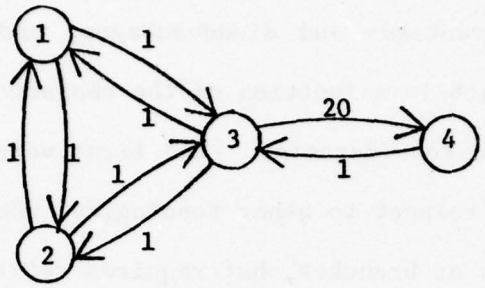
The easiest case to handle is that of decreasing branch lengths. A crucial element of the convergence proof for the static algorithm is that every distance assessment is no smaller than the corresponding true distance, so that monotone convergence applies. For decreasing branch lengths this condition is still met and the convergence proof remains valid.

On the other hand, when a branch length increases, this condition may be violated and the original convergence proof is no longer applicable. The algorithm may or may not still converge for a particular graph; even if it does, convergence may be slow, as illustrated by the following example:

Example 1.



Let node 4 be the destination and assume that the static algorithm has converged to the correct solution. Thus node 3, for example, has a shortest distance of 1, with alternative paths via node 1 or 2 of length 3. Now suppose that branch length s_{34} increases to 20, giving



Notice that all shortest paths to node 4 are affected by the change. When the change occurs, node 3 can immediately adopt 20 as his new direct distance to node 4. However, he now believes that he can achieve a distance of 3 via node 1 or 2. Obviously, this distance of 3 can no longer be physically achieved, but node 3 is unaware that his alternative paths have been affected by the branch increase. He now tells nodes 1 and 2 that his current shortest distance has increased to 3. Node 1 now compares his new distance of 4 via node 3 to his distance via node 2, and decides that his new current shortest distance is 3. Node 2 takes similar action. They then transmit this information to node 3, who now increases to 4 his assessment of the distance to node 4. This process of gradually increasing the distance assessments will continue until the true distances are reached; so even with this small, simple example, convergence will take quite a long time.

Some modification to the algorithm is necessary in order to guarantee convergence when branch lengths increase. When such an increase does occur, the nodes must somehow be enabled to determine which distance assessments can be trusted, and which cannot.

With this goal in mind, we have developed several modified versions of the static algorithm, three of which are described below. Each version has its advantages and disadvantages, and the applicability and suitability of each is a function of the characteristics of the particular network under consideration. The first version is the simplest and most robust with respect to other topological changes, such as introduction of new nodes or branches, but requires total suspension of distance communication for a sufficiently long period that all nodes can be guaranteed to have reinitialized. The second method can effectively handle branch increases and failures but can encounter difficulties when a new link (or node) is introduced into the network. Thus, if links are added rarely and under controlled circumstances, this version could be appropriate. The third modified algorithm was developed in an attempt to improve the second, but as we have often found to be true, a modification which solves one problem can introduce a new one. In this third version the problem of introducing new links has been solved but certain link or node failures cannot be accommodated and require special handling.

Each of these modified algorithms is based primarily on some form of reinitialization of the basic static algorithm; they differ mainly in the mechanics of the reinitialization. It is not sufficient to merely disseminate a reinitialization command throughout the network when a branch increase occurs. Once a node has reinitialized its distance assessments, it needs some guarantee that subsequently received information has also been reinitialized, and some means of doing this must be introduced.

Modification A

Perhaps the simplest and most robust approach is to effectively suspend all communication of distance information for a sufficiently long period of time to insure that all nodes have reinitialized. Several possible mechanisms for achieving this present themselves: one is for the node detecting a branch length increase that affects any of his current shortest distances to decide upon a future time at which communication of distance information based on reinitialization will resume, and to send this to his neighbors who continue to propagate it throughout the network. Implicit here is the existence of a time base common to all nodes, and the availability to each node of (at least an upper bound on) the time it takes for the "reinitialization message" he initiates to propagate throughout the network, which implies a more global knowledge of the network. Since communication of distance information is suspended for this period, it is advantageous to make it as small as possible. This will clearly be aided if a mechanism exists for making these "reinitialization messages" top priority so that they bypass all queues and buffers at each node.

Other mechanisms for achieving the same basic objective have been devised. Together with that above, they share the convergence of the static algorithm and its robustness with respect to other topological changes such as introduction of new nodes or branches. Its feasibility requires that branch length increases occur infrequently relative to the total time information transmission is suspended and the static algorithm subsequently converges.

Modification B

Instead of suspending all communication of distance information until all nodes can be guaranteed to have reinitialized and all distance information can be trusted, a mechanism has been devised for each node to determine which distance information he receives is trustworthy (in that all nodes further down the corresponding path are guaranteed to have reinitialized) and which is not. In simple terms, on hearing that a branch length increase or failure has taken place, a node ignores distance information sent by any questionable neighbor until that neighbor acknowledges that he, too, is aware of the change. As each neighbor in turn so acknowledges, the embargo on his information is removed. In this way, some convergence toward the new solution can be taking place while news of the change is still propagating through the network.

More precisely, this modified algorithm involves the use of a "special action", which, as in the previous algorithm, is initiated by a node detecting a branch increase or failure that affects any of his current shortest distance assessments. When this occurs, the initiator assigns a unique index to the special action (consisting of his node number and a counter), and does the following:

1. Reinitializes
2. Places an embargo, indexed by the special action, on distances received from every neighbor, except for the neighbor at the opposite end of the affected link.
3. Transmits all of his new shortest distances to each neighbor, along with the special action index.

Each neighbor receiving this information takes analogous action, in Step 2 placing an embargo on all his neighbors except for the one he has just heard from, and in Step 3 using the same special action index. He then waits until a message is received from a neighbor, at which point his action is governed by the following:

Case 1. Message contains no special action index.

A. If there is no embargo on this node, calculate the new distances via this neighbor and proceed normally.

B. Else, ignore the message.

Case 2. Message contains a special action index.

A. If there is no embargo on this node with the same index as in the message, perform steps 1 - 3 above.

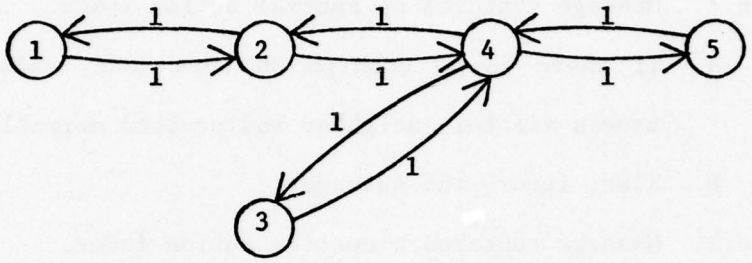
B. Else, remove the matching ban from this particular neighbor. If no other embargo exists, calculate new distance via this node. Otherwise, ignore distance component of message.

Several special actions can exist within the network simultaneously, each distinguishable by its index. A branch failure can be treated as a branch length increasing to infinity. As before, to insure convergence it is necessary for the topology and branch lengths of the network to remain constant for a long enough period of time for the algorithm to find the new solution.

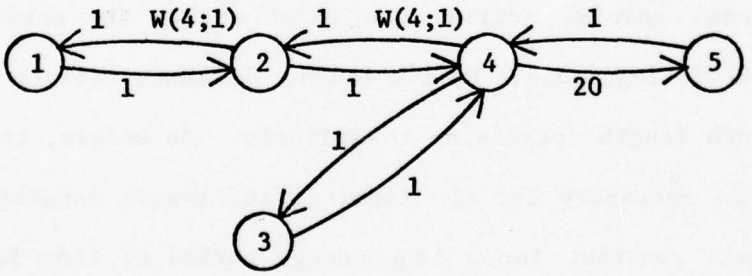
While this algorithm solves the problem of increasing branch lengths, it introduces a new difficulty, viz. that of adding new nodes or links to the network. In the preceding algorithms, bringing up a new node or link causes no trouble. It introduces new paths, which can only serve to decrease shortest paths through the network. Thus, one should

be able to consider this in the category of decreasing branch lengths, discussed earlier. However, in connection with this particular version of the algorithm, adding a new link can cause a serious problem, as illustrated by the following example:

Example 2.



Let node 5 be the destination; suppose branch length s_{45} increases to 20, and apply Modification B. Furthermore, suppose we have reached the point at which node 4 has informed nodes 2 and 3 of special action (4;1), node 3 has acknowledged receipt of this message, but node 2 has not. Pictorially:



where W(4;1), read wait for (4;1), indicates that a ban exists on distances via the specified outgoing neighbor. At this point, every node is aware of special action (4;1) except node 1, who believes that his shortest distance to node 5 is 3. Furthermore, node 3 has fulfilled his duties with respect to action (4;1), and no longer has any memory of it. If node 2, for whatever reason, is

slow in relaying the message concerning action (4;1) to node 1, a problem can develop. Suppose that a pair of links between nodes 1 and 3, each of length 1, is introduced at this point. Node 1 sends its distance of 3 to node 3, who in turn tells node 4 that his distance to the destination has been cut to 4. Node 4, having already removed the ban from node 3, passes this on to node 2. When node 2 finally transmits the special action message to node 1, it is accompanied by the false distance of 6. This is clearly an undesirable situation, brought about by the introduction of the new links between nodes 1 and 3. It happened because node 3 informed all of his neighbors of the special action, and later acquired a new neighbor who was unaware of the action.

Modification B can be used for networks with increasing and decreasing branch lengths and failures, but some other technique must be utilized to add nodes or links to the network.

Modification C

This version of the algorithm is a modification of the previous one, developed to solve the problem of adding links. If a node remembers the indices of special actions after he processes them, he can insure that any new neighbors that he acquires are informed of the most recent special actions. But how long must a node remember which actions have occurred? In a network in which branch increases are a common occurrence, it would be infeasible for a node to remember every special action that it had received. Actually, a node need remember a particular special action only until he can be certain that every node in the network has knowledge of it.

Thus, in order to enable a node to decide when it is safe to forget a given special action, we modify Algorithm B. Now, when a node first hears about a special action, he performs the same steps as before, with one alteration. The node now remembers which neighbor first informed him of the action, and may not send acknowledgement to that neighbor immediately. He tells his other neighbors about the action, waits until they all acknowledge receipt of that message, and then sends his acknowledgement to the node which first sent him word of the special action. The initiator of the action plays the role of a temporary controller. When he has received acknowledgement from all of his neighbors, every node in the network knows of the special action. The initiator can now send the "all clear" signal, allowing each node to erase memory of the action.

Unfortunately, this modification introduces a new difficulty; link and node failures no longer behave the same as branch increases. For instance, certain node or link failures can disrupt the flow of acknowledgement messages, necessitating that some form of emergency action be taken. We have developed a mechanism which enables any node that detects a failure that could interrupt the flow of acknowledgements to initiate this emergency action. The emergency procedure itself must be able to bring the network to a state from which convergence is guaranteed; perhaps the simplest such procedure is to temporarily invoke Modification A in those networks where it is feasible.

4.4 Summary

The static algorithm in Appendix 5 is the foundation upon which each of our algorithms is based. This fundamental algorithm has localized information and communication requirements, operates asynchronously, and is guaranteed to find, in finite time, all shortest paths in a static network; also in networks in which branch length decreases and addition of new nodes and links take place. For networks with increasing branch lengths or branch failures, some means of reinitializing the algorithm is introduced. Modification A is the most versatile of these reinitialization schemes; it can accommodate branch increases, decreases failures and additions, but it requires suspension of all distance communication and shortest path calculations for a sufficiently long period that all nodes can be guaranteed to have reinitialized. However, this algorithm does provide a simple means of reinitialization and is the most robust of the algorithms that we have developed. Modification B is a more complicated algorithm. Its reinitialization mechanism relies upon an acknowledgement system that increases the required information storage capacity of each node. It can accommodate branch length decreases, increases, and failures, but not the introduction of new nodes or links. Its main advantage is the ability to begin converging toward the new solution soon after the reinitialization process begins; its main disadvantage is the need for special treatment in order to insert new links. Modification C contains a more sophisticated acknowledgement system, further increasing the storage requirements of each node. Branch additions and branch length increases and decreases can be accom-

modated, as well as certain branch failures; a node detecting any failure is capable of determining whether it should initiate an emergency procedure in order to guarantee convergence. We note that algorithms utilizing acknowledgement systems are being investigated elsewhere. For instance, Merlin and Segall [3] have developed an algorithm which is more complicated than any of ours and which solves a reduced version of the problem we consider.

The investigation of decentralized network algorithms is continuing as the doctoral research project of Jeffrey M. Abram.

5. A GEOMETRIC APPROACH TO SINGULAR ESTIMATION

The paper presented at the 1976 IEEE International Symposium on Information Theory:

> "A Geometric Approach to Singular Estimation Problems," Ian B. Rhodes, 1976 IEEE International Symposium on Information Theory, Ronneby, Sweden, June 1976.

considered the singular estimation problem characterized by the following question: Given the constant linear system

$$\dot{x}(t) = Ax(t) + Dv(t)$$

with noise-free observations

$$y(t) = Cx(t)$$

where v is white Gaussian noise, what states x(t) can be determined exactly by

(a) differentiation of the current output y(t)?

(b) constructing an appropriate observer that utilizes smoothly the past observations y over [0,t]?

Our answers to these questions concerning singular estimation were based for the first time on a geometric viewpoint drawing on the ideas first introduced by Wonham and Morse [4] and Basile and Marro [5]. Assuming without loss of generality that (A,D) is completely controllable and (C,A) completely observable, the answers we found are:

(a) By differentiation, $x(t)$ can be determined exactly modulo the subspace $\bar{W}$, where $\bar{W}$ is the maximal (A,D)-invariant subspace contained in the nullspace of C. The maximal (A,D)-invariant nullspace in a given subspace was first introduced by Wonham and Morse [4] in their geometric approach to decoupling and other problems. Specifically, $\bar{W}$ is the largest subspace W that is contained in the nullspace, N(C), of C and satisfies $AW \subset W + R(D)$, where R(D) denotes the range of D. It is known [4] that W can be obtained using the iterative formula

$$W_{i+1} = N(C) \cap A^{-1} [W_i + R(D)], \; W_o = N(C)$$

which converges to the limit $\bar{W}$ in at most n-1 steps. It is also known [6] in the single-output case (where C is a row vector) that $\bar{W}$ can be found as

$$\bar{W} = \text{Nullspace} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^d \end{bmatrix}$$

where d is the smallest integer such that $CA^dD \neq 0$. In the context of singular estimation, this reflects the well-known idea that the output is differentiated until the white noise v first appears; because $CA^iD = 0$ for $i < d$, the first d derivatives of y(t) do not contain v and the equation

$$\begin{bmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(d)}(t) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^d \end{bmatrix} \quad x(t)$$

can be solved for x(t) up to $\bar{W}$, the nullspace of the matrix appearing on the right side. Note that the (d+1)-st derivative of y(t),

$$y^{(d+1)}(t) = CA^{d+1}x(t) + CA^{d}Dv(t) \quad ,$$

is the first containing the unknown v(t) and gives only statistical, but not exact, information about x(t).

Thus our result that x(t) can be determined exactly modulo $\bar{W}$ generalizes to multiple-output systems a well-known result for single-output ones.

(b) Using an observer, such as

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K[y(t) - C\hat{x}(t)],$$

the set of states x(t) that can be determined exactly depends on the covariance of the initial state x_o, but the largest such set is the subspace $\bar{V}$, where $\bar{V}$ is the maximal (A',C')-invariant subspace in N(D'). Again, $\bar{V}$ can be found in at most n-1 steps using the analog of the iterative formula given above. Here the comparable explicit formula is for single input systems, viz. if D is a vector and d is as above,

$$\bar{V} = N \begin{bmatrix} D' \\ D'A' \\ \vdots \\ D'A'^{d} \end{bmatrix}$$

For either single-input or multi-input systems, $\bar{V}$ can be interpreted in terms of the observer error, $\tilde{x}(t) = x(t) - \hat{x}(t)$, which satisfies

$$\dot{\tilde{x}}(t) = (A - KC)\,\tilde{x}(t) + Dv(t)$$

The subspace $\bar{V}$ is the largest set of x's that can be made, by appropriate choice of the observer gain K, unaffected by the noise input v. If the covariance of the initial state is zero along $\bar{V}$ and K is chosen appropriately, the covariance of x(t) remains zero along $\bar{V}$ for all future t.

In summary, by differentiation we can determine x(t) modulo $\bar{W}$ and by an observer, assuming appropriate initial noise covariance, we can determine x(t) exactly along $\bar{V}$. Now, if we take $(\bar{W})^{\perp}$ to be the prototypical set that can be determined exactly by differentiation and $(\bar{V})^{\perp}$ to be the prototypical subspace modulo which x(t) can be determined using an observer, then

(i) the set of states that can be determined exactly <u>both</u> by differentation <u>and</u> by an observer is $\bar{V} \cap (\bar{W})^{\perp} \triangleq \bar{R}$. It turns out that this subspace is another fundamental subspace type introduced by Wonham and Morse: $\bar{R}$ is the maximal (A',C')-<u>controllability</u> subspace contained in N(D'); for a definition of this, see [4].

(ii) The set of states that cannot be determined either by differentiation or by an observer is $(\bar{V})^{\perp} \cap \bar{W} \triangleq \bar{S}$, i.e. using both differentiation and an observer, the state can be determined only modulo $\bar{S}$. In this case, $\bar{S}$ is the maximal (A,D)-controllability subspace in N(C).

Of course, if $\bar{S}$ is simply the origin, then the entire state can be determined exactly using together differentiation and an observer. A sufficient (but not necessary) condition for this is that the system have a single "input", i.e. the noise v is real-valued, and not vector-valued. Similarly, if $\bar{R}$ is simply the origin, then no state can be determined both by differentiation and by using an observer. A sufficient (but not necessary) condition for this is that the system have a single output. If both $\bar{R}$ and $\bar{S}$ are simply the origin, then all states can be determined but none by both differentiation and by an observer. In other words, the state space then can be separated into a direct sum of two sub-spaces, one of these being the set of states determinable by differentiation and the other the set of states determinable by an observer. This will be so for single-input, single-output systems for which, as has been noted above, explicit algebraic formulas are available for both $\bar{V}$ and $\bar{W}$.

We have thus provided an alternative interpretation of the special properties enjoyed by single-input, single-output systems insofar as singular estimation is concerned. At the same time, we have shown that the ideas and objects of geometric systems theory are convenient and natural for solving the general multi-input, multi-output singular estimation problem.

For the discrete-time system

$$x_{k+1} = Ax_k + Dv_k$$
$$y_k = Cx_k$$

the states that can be determined exactly using an observer

$$\hat{x}_{k+1} = A\hat{x}_k + K[y_k - C\hat{x}_k]$$

is exactly as in the continuous-time case, viz. it depends on the covariance of the initial state, but is at most the subspace $\bar{V}$ defined earlier. However, the analog of differentiation is differencing, and the result of waiting for the output data so that differencing can be performed is that smoothing becomes involved. It is found that the continuous-time result holds true for deducing x_k from future and present observation data y_j, $j \geq k$, i.e. x_k can be determined modulo the sub-space $\bar{W}$. Because the useful data will in fact extend at most n-1 steps into the future, we can fix data availability to time k and determine x_{k-j} modulo $\bar{W}$ using data to time k, for an appropriate j, $0 \leq j \leq n-1$. It has been our longstanding conjecture that suitable forward propagation of $\bar{W}$ or $(\bar{W})^{\perp}$ is intimately connected to the constant directions of the Riccati equation [7] - [9], but the exact form of this relationship has yet to be established.

For both continuous-time and discrete-time systems, corresponding new results for the singular control problem follow by standard duality arguments.

6. COMPENSATOR DESIGN FOR POLYNOMIAL MATRIX SYSTEM DESCRIPTIONS

6.1 Introduction

A major conceptual development in system theory over the last couple of decades has been the replacement of the classical transfer function approach in the frequency domain by the state-space approach in the time domain. However, in the past few years, there has been a resurgent interest in frequency-domain methods. This is due in large part to the development of an alternative time-domain technique, namely the differential operator approach to the analysis and synthesis of time-invariant linear multivariable dynamical systems. Rosenbrock [10] has shown how to derive many state-space results through the analysis of certain polynomial matrices. Independently, Popov [11] has shown how such seemingly state-space-theoretic problems as the realization of systems in controllable canonical form and the determination of the controllability indices could be elegantly solved by polynomial matrix methods, starting from the transfer matrix. More recently, a significant number of investigators [12]-[20] have used polynomial matrix methods to solve other problems, employing the fact that the (pxm) transfer matrix, T(s), of a linear time-invariant multivariable system can be factored as

$$T(s) = R(s)P^{-1}(s) = P_Q^{-1}(s)\, Q(s) \tag{1}$$

where R(s) and P(s) [$P_Q(s)$ and Q(s)] are relatively right[left] prime polynomial matrices in the Laplace operator s, and P(s) [$P_Q(s)$] is column[row] proper. Such a factorization directly implies a minimal time domain realization of T(s) in differential operator form, namely

$$P(D)z(t) = u(t)$$
$$y(t) = R(D)z(t) \tag{2}$$

or

$$P_Q(D)y(t) = Q(D)u(t) \tag{3}$$

where P(D) and R(D)[P_Q(D) and Q(D)] are polynomial matrices of dimensions mxm and pxm[pxp and pxm] in the differential operator D=d/dt with P(D)[P_Q(D)] column[row] proper and nonsingular, z(t) is a p-vector called the partial state, u(t) is the m-vector input, and y(t) is the p-vector output.

The equivalent state-space representation of (2) is just

$$(DI-A)x(t) = Bu(t)$$
$$y(t) = Cx(t) \tag{4}$$

where x(t) is the state vector and A,B,C are real matrices of appropriate dimensions.

In view of (1) and (2) or (3), the Laplace operator s and the differatial operator D can be, and will be, freely interchanged in our subsequent discussion. It should be noted that a differential operator description of the dynamical behavior of a physical system often follows as a direct r sult of applying well-known physical laws to model the system.

One of the most important features of the controllable differential operator representation (2) can be seen if we consider the effect of the linear state variable feedback (lsvf) on a compensated system defined by

the control law

$$u(t) = Fx(t) + Jv(t) \tag{5}$$

in the case of state-space representation of the form (4), or

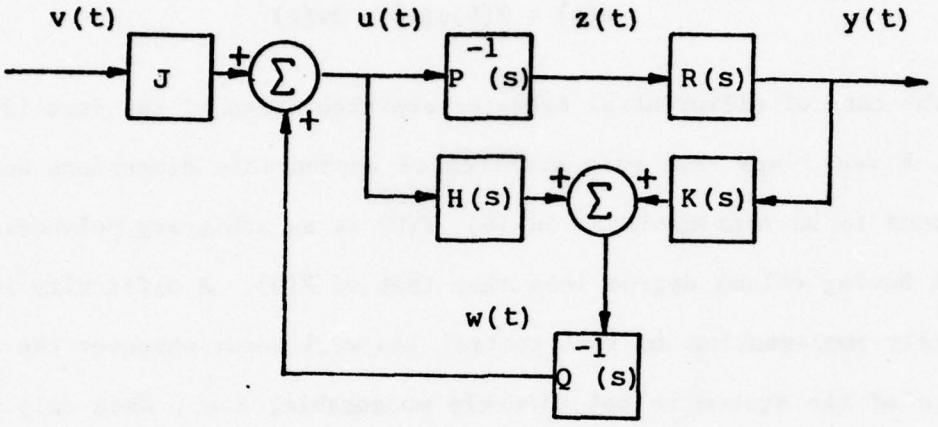
$$u(t) = F(D)z(t) + Jv(t) \tag{6}$$

in the case of differential operator representation of the form (2). In (5), F and J are real gain matrices of appropriate dimensions and J is assumed to be nonsingular. In (6), F(D) is an arbitrary polynomial matrix having column degree less than that of P(D). A difficulty in physically implementing an lsvf control law will occur whenever the entire state of the system is not directly measurable; i.e., when only $y(t) = R(D)z(t)$ is available for direct measurment. This problem can be circumvented in the case of an observable state-space system through the employment of a Luenberger observer. An entirely analogous result can be obtained by the differential operator approach. In this regard, the following result [12, Theorem III] is important:

> Consider the differential operator representation (2). For any F(D) of lower column degree than P(D), there exists a triple $\{Q(D),H(D),K(D)\}$ of polynomial matrices satisfying the following two properties:
>
> (a) $K(D)P(D) + H(D)R(D) = Q(D)F(D)$
>
> (b) $Q^{-1}(D)[H(D)\ K(D)]$ is an asymptotically stable proper transfer matrix.

The significance that this result has for compensator design can be seen

by considering the compensation, depicted below, of a given system (P(D), R(D)) of the form (2):



It is seen from the diagram that

$$u(t) = w(t) + Jv(t)$$

If w(t) exponentially approaches F(D)z(t) then the compensation scheme (6) will be realized asymptotically. A little algebra using (2) and the relation from the block diagram

$$Q(D)w(t) = K(D)u(t) + H(D)y(t)$$

shows that this is precisely what conditions (a) and (b) of the theorem ensure; furthermore, the compensator is stable and realizable because its transfer functions $Q^{-1}(s)H(s)$ (from y to w) and $Q^{-1}(s)K(s)$ (from u

to w) are so specified by condition (b). A procedure for constructing a triple {H(D), K(D), Q(D)} with the two properties in the theorem have been given by Wolovich [12], using the classical "eliminant matrix" of two polynomials. There are, however, no methods available for constructing a triple with these two properties and the additional property that the determinant of Q(D) has minimum degree. This corresponds to the design of a stable, minimum-order compensator. Our research efforts have been directed towards the development of such methods, with a view to subsequent application of these ideas to other system problems.

6.2 Problem Formulation

Consider a linear time-invariant multivariable system defined by the equations

$$P(D)z(t) = u(t)$$

$$y(t) = R(D)z(t)$$

where P(D) and R(D) are polynomial matrices of dimensions mxm, pxm, respectively, in the differential operator D=d/dt and P(D) is column proper and nonsingular. Our goal is, for any polynomial matrix F(D) of dimension mxm having lower column degree than Q(D) to find a triple {H(D), K(D), Q(D)} of polynomial matrices of appropriate dimensions which satisfies the following properties:

i) $H(D)P(D) + K(D)R(D) = Q(D)F(D)$ (7a)

ii) $Q^{-1}[H(D)\ K(D)]$ is an asymptotically stable proper transfer function (7b)

iii) det Q(D) is of minimal degree (7c)

For simplicity, the argument s or D will be omitted hereafter, since the two are interchangeable. Equation (7a) can be rewritten as

$$HP + KR - QF = 0$$

or, equivalently,

$$[P' \; R' - F'] \begin{bmatrix} H' \\ K' \\ Q' \end{bmatrix} = [0] \tag{8}$$

Let

$$T = [P' \; R' - F'],$$

$$G = \begin{bmatrix} H' \\ K' \\ Q' \end{bmatrix},$$

where P' is a row proper, $(RP^{-1})'$ is proper, and the row degree of F' is less than that of P'. Hence the polynomial matrix T is row proper and of full rank. Therefore, instead of solving (7a) for the triple {H, K, Q}, we may solve the linear equation on free polynomial modules,

$$TG = [0] \tag{8'}$$

where the elements of the composite polynomial matrix G are to satisfy conditions (7b) and (7c).

6.3 The Free Modular Approach

Consider the linear equation

$$T\underline{m} = \underline{n} \tag{9}$$

where T is a linear map from the free R[s]-module M of rank r(M) to the

free R[s]-module N of rank r(N). It is clear that (9) has a solution if and only if $\underline{n} \in$ Im T. Because R[s] is a free principal ideal domain, Im T and ker T are free modules. Equation (9) thus involves, for complete analysis, the computation of two bases, one for Im T, and one for ker T. We also have the customary identity

$$r(\text{Im } T) + r(\ker T) = r(M).$$

Our procedure for solving (8') is first to determine a minimal reduced basis for ker T and then use this basis to construct a polynomial matrix G which meets all the requirements.

Let $v_1,\ldots,v_n$, n=r(ker T), be a basis for ker T, where each element can be expressed in the manner

$$v_i = \sum_{j=0}^{r_i} v_{i,j} s^j \quad , \qquad v_{1,r_i} \neq 0$$

where r_i is the column degree of v_i. We say that $\{v_1,\ldots,v_n\}$ is a minimal reduced basis for ker T if the rank of the constant matrix formed from the last m rows of

$$[v_{r_1} \ \ldots \ v_{r_m}] \tag{10}$$

is equal to m and $\sum_{i=1}^{m} r_i$ is minimal.

For a row proper and full rank matrix T, a minimal reduced basis for ker T always exists, though it is not unique. There are algorithms for constructing such bases, although we shall not present them here. Without loss of generality, we may assume that our basis is always minimal and reduced.

Returning to equation (8'), we have an equation immediately recognized as (9) with $\underline{n} = 0$ and T: $R[s]^{2m+p} \rightarrow R[s]^m$ given by matrix [P' R' - F']. The matrix G = [H K Q]' can then be constructed using the basis $\{v_1,\ldots,v_m\}$ for ker T. In view of (8), n = rank(ker T) = m + p, and the column elements of G can then be expressed as linear combinations of the basis elements $v_1,\ldots,v_n$; i.e.

$$g_i = \sum_{j=1}^{n} a_{i,j} v_j \qquad i = 1,\ldots,m \tag{11}$$

where $G = [g_i]$, and the g_i and v_i are partitioned as

$$g_i = \begin{bmatrix} h_i \\ k_i \\ q_i \end{bmatrix} \qquad v_i = \begin{bmatrix} n_i \\ d_i \end{bmatrix}$$

The determinant of Q' can be written as the exterior product of the q_i; i.e.,

$$\text{Det } Q' = q_1 \Lambda \ldots \Lambda q_m \tag{12}$$

By rewriting (12) with the aid of (11), while taking due account of the multi-linear and skew-symmetric properties of the expression, it is possible to obtain the revealing form

$$\text{Det } Q' = \sum b_{k(1),\ldots,k(m)}(s)[d_{k(1)}(s) \Lambda \ldots \Lambda d_{k(m)}(s)] \tag{13}$$

where the sum is taken over all integer arrangements satisfying

$$1 \leq k(1) \leq \ldots. \leq k(m) \leq n \tag{14}$$

It then becomes a matter of determining the b's in (13).

Let $Q_{m,n}$ denote the set of all strictly increasing sequences of positive integers r, of length m, chosen from 1,....,n, i.e., $1<r(1)<....<r(m)<n$. Because the $d_{k(i)}(s)$ are known, finding the b's in (13) is essentially a decomposition problem. Select a basis $v_1,....,v_n$ for an F-vector space V; then an m-vector has the general form (13) for the sum taken over all integer arrangements in $Q_{m,n}$ and with the b's in F. Decomposition means that m vectors

$$\tilde{v}_i = \sum_{j=1}^{n} a_{i,j} v_j \qquad a_{i,j} \in F \tag{15}$$

can be found in V such that the exterior product

$$\tilde{v}_1 \wedge ..... \wedge \tilde{v}_m \tag{16}$$

is equal to the m-vector in question.

The mathematical literature dealing with the construction $b \rightarrow a$ tends to be framed in a vector space, rather than a free-module, context. For example, the following result is known [21, p. 568]:

Proposition: In an (m+1)-dimensional F-vector space V, every m-vector is decomposable.

An extension of this result to free R[s]-modules has been given by Sain[13]. Using this extension, a solution to the compensator design problem has been constructed for single-output systems.

6.4 Single-Output Systems

For single-output systems, p=1 and so m=n-1, and a polynomial matrix $A(s) = [a_{i,j}(s)]$ can be found such that

i) $b_r = \text{Det } A_r$, $\quad r \in Q_{m,n}$,

ii) $G = [v_1 \ldots\ldots v_n]A$,

iii) $Q^{-1}[H\ K]$ is proper,

where A_r is the submatrix of A consisting of the r(1),...,r(m)-th columns of A.

The matrix A(s) can be determined in one or two steps depending on the solution to (13). Suppose that we have a solution b to (13) for a desirable polynomial Det Q'. Step 1 is to define a matrix B as follows

$$B = \begin{bmatrix} b_2 - b_3/b_1 & \cdot & \cdot & (-1)^{n+1}b_{n+1}/b_1 \\ b_1 & & & \\ & 1 & & \\ & & \cdot & \\ & & & \cdot \\ & & & 1 \end{bmatrix} \qquad (17)$$

where b_i is the shorthand notation for b_{r_i}, and r_i is the i-th element of $Q_{m,n}$ when its elements are ordered lexicographically. If b_1 is a constant, we may simply let A be equal to B; otherwise, step 2 is necessary to convert the rational matrix B to a polynomial matrix. Step 2 is to determine the matrix A from the fact that

$$(\tilde{v}_1 \wedge \ldots\ldots \wedge \tilde{v}_m) \wedge \tilde{v}_i = 0 \ , \qquad i = 1,\ldots,m$$

where $\tilde{v}_i$ is defined by (15) except that $a_{i,j}$ are elements of R[s] and $\{v_1,\ldots,v_n\}$ is a minimal reduced basis for V = ker T(s); i.e.,

$$\Big(\sum_{r_i \in Q_{m,n}} b_{r_i} v_{r_i}^{\wedge}\Big) \wedge \Big(\sum_{i=1}^{n} a_{i,j} v_i\Big) = 0 \qquad (18)$$

where $v^{\wedge}_{r_i}$ denotes the exterior product of the m elements of $\{v_1,\ldots,v_n\}$ corresponding to $r_i \in Q_{m,n}$. Because $v^{\wedge}_{r} \wedge v_i = 0$ if i is in the sequence r for all $r \in Q_{m,n}$ and $v_1 \wedge \ldots \wedge v_n$ cannot vanish because $\{v_i\}$ is a basis for V, we have

$$\sum_{i=1}^{N} (-1)^{i+1} b_i a_{i,j} = 0, \quad j = 1,\ldots,m$$

where $N = \binom{n}{m}$ = number of sequences in $Q_{m,n}$. We are then essentially looking for a polynomial basis for

$$\ker [b_1, -b_2, \ldots, (-1)^{N-1} b_N] \tag{19}$$

in the free R[s]-module sense. If one of the non-zero b is a constant, then, by reordering $\{v_i\}$ such that b_1 is that constant, the column elements of B defined in (17) form a basis for the kernel (19). Otherwise B is a rational function and the column vectors of B form a basis for that kernel in an R(s)-vector space context but not necessarily a free modular basis. An algorithm has been developed to construct a free modular basis from the B matrix.

6.5 Multi-Output Systems

For multi-output systems, $p > 1$ and so $m < n-1$, and the free R[s] module extension of the Proposition given at the end of Section 3 is no longer applicable. Generalizations of that Proposition in a vector space setting are known; e.g., Theorem 1.4 in [22], but we have so far been unable to establish any extension to a free R[s]-module setting. That such an extension should be possible, at least under certain circumstances, is illustrated by the following alternative approach.

Consider the possibility of triangularizing the transfer function matrix by precompensation; since the set of proper, stable, real, rational transfer functions forms a Euclidean domain, it is possible to triangularize a transfer function matrix by postmultiplying it by a unimodular matrix of rational functions [23], thus transforming a multi-input, multi-output system into a sequence of multi-input, single-output systems. Postmultiplication by a unimodular matrix amounts to allowing "input dynamics" in the compensation scheme.

It can be shown that, for a minimal system described by (2), the exterior products d_r^{Λ}, for all $r \in Q_{m,n}$, are relatively prime polynomials. Hence, the determination of minimum degree of Q' matrix is essentially to find $a_r(s)$, $r \in Q_{m,n}$, and the smallest integer k such that

i) $c_0 + \ldots + c_k s^k = \sum_{r \in Q_{m,n}} a_r(s) d_r(s)$ is a stable polynomial of degree k, and

ii) $\deg a_r(s) + \deg d_r(s) \leq k$.

This problem seems to be related to the generalization of the so-called "eliminant matrix" of two polynomials.

Further investigation of these questions is being carried on as the doctoral research project of Olive Y. Liu.

7. ESTIMATION AND STOCHASTIC CONTROL: THE INNOVATIONS CONJECTURE

7.1 Introduction

The informational equivalence of the "signal in additive white noise" observation process

$$z_t = \int_0^t y_s ds + w_t \tag{1}$$

and the innovations process

$$v_t = z_t - \int_0^t \hat{y}_s^Z \, ds = \int_0^t \tilde{y}_s^Z \, ds + w_t \tag{2}$$

is known to hold under certain conditions and to fail to hold under others; an account of these can be found in Benes [24]. Two of the most important sets of conditions under which informational equivalence holds are:

1. The "signal" y is a second-order process with $E\int_0^t |y_s|^2 ds < \infty$, the "noise" w is an uncorrelated-increment second-order process, future noise is uncorrelated with past signal, and our concern with with <u>linear</u> least-squares estimation. Specifically, let Z_t be the closed-linear subspace of $H = L_2(\Omega,F,P)$[1] generated by $\{z_s^i, s\in[0,t],\ i = 1,2,\ldots,m = \dim z_s\}$.

---

[1] For the complete probability space (Ω,F,P), $H=L_2(\Omega,F,P)$ is the Hilbert space of real-valued, zero-mean, finite-variance random variables on (Ω,F,P) with inner product $\langle u,v\rangle = E(uv)$.

Then $\hat{y}_s^Z$ in (2) is the projection $E^*[y_s|Z_s]$ of y_s on Z_s, $\tilde{y}_s^Z \triangleq y_s - \hat{y}_s^Z$, and informational equivalence means that $Z_t = N_t$ for all t, where N_t is the closed linear subspace of H generated by $\{v_s^i, s\in[0,t],\ i = 1,2,\ldots,m\}$.

2. The "signal" y and the "noise" w are jointly-Gaussian second-order processes, y is almost-surely square-integrable on any finite interval, w is a Wiener process whose future increments are independent of past y, and our concern is with least-squares estimation. Specifically, let Z_t by the sub-σ-algebra of F generated by $\{z_s^i, s\in[0,t],\ i = 1,2,\ldots,m\}$. Then $\hat{y}_s^Z$ in (2) is the conditional expectation $E[y_s|Z_s]$, $\tilde{y}_s^Z = y_s - \hat{y}_s^Z$, and informational equivalence means $Z_t = N_t$ (mod P) for all t, where N_t is the sub-σ-algebra of F generated by $\{v_s^i, s\in[0,t],\ i = 1,2,\ldots,m\}$.

We have constructed a new direct proof of innovations equivalence for these two cases using only elementary facts from stochastic processes and estimation theory such as those in the book [25]. This contrasts with most of the existing proofs which involve deep and sophisticated results in the theory of stochastic processes. In terms of directness, generality and assumed background, our proof is comparable to that recently published in [26]; the principal argument in the proof, however, is entirely different.

A paper presenting these proofs in detail is in preparation. We summarize here the essential arguments involved.

7.2 Linear Least-Squares Estimation

The essential ingredient of the proof is to introduce the process μ defined by

$$\mu_t = z_t - \int_0^t \hat{y}_s^N \, ds \tag{3}$$

where $\hat{y}_s^N = E^*[y_s | N_s]$. Recalling that $N_t \subset Z_t$, it is immediately clear from (3) that $M_t \subset Z_t$, and thus that $M_t + N_t \subset Z_t$, where M_t is the subspace of H generated by $\{\mu_s^i, s\ [0,t],\ i = 1,2,\ldots,m\}$. On the other hand, rewriting (3) as

$$z_t = \mu_t + \int_0^t \hat{y}_s^N \, ds \quad ,$$

it is seen immediately that $Z_t \subset M_t + N_t$. Thus

$$Z_t = M_t + N_t \quad . \tag{4}$$

This fact, that knowledge of the past of both μ and ν is equivalent to knowledge of the past of z, is the very reason the process μ is introduced.

Now, substitution of (2) into (3) gives

$$\mu_t = \int_0^t \hat{y}_s^Z - \hat{y}_s^N) ds + \nu_t \tag{5}$$

It is easily proved that the processes $(\hat{y}^Z - \hat{y}^N)$ and v appearing on the right side of (5) are uncorrelated, and by an easily proved result [25, Lemma 4.3.2] the subspace M_t can then be represented as the set of integrals

$$M_t = \left\{ \int_0^t b'(s)[\hat{y}_s^Z - \hat{y}_s^N]ds + \int_0^t b'(s)dv_s ; b \epsilon L_2^m[0,t] \right\} \tag{6}$$

This, in fact, is a generalization of the standard result that N_t can be represented as the set of Wiener integrals on v, i.e.,

$$N_t = \left\{ \int_0^t a'(s)dv_s ;\ a \epsilon L_2^m[0,t] \right\} \quad . \tag{7}$$

In view of (4), any vector in Z_t can then be represented as a vector sum $m_t + n_t$, where m_t and n_t have representations as above, i.e.,

$$Z_t = M_t + N_t = \left\{ \int_0^t b'(s)[\hat{y}_s^Z - \hat{y}_s^N]ds + \int_0^t c'(s)dv_s ;\ b, c \epsilon L_2^m[0,t] \right\}$$

In particular, $(\hat{y}_t^Z - \hat{y}_t^N)$ is in Z_t and has a representation of this form: also, because $\hat{y}_t^N$ is the projection of $\hat{y}_t^Z$ on N_t, $(\hat{y}_t^Z - \hat{y}_t^N)$ is orthogonal to N_t and $c' \equiv 0$. Thus

$$\hat{y}_t^Z - \hat{y}_t^N = \int_0^t B'(t,s)[\hat{y}_s^Z - \hat{y}_s^N]ds$$

After it is shown that

$$\int_0^T \int_0^t \sum_{i,j} \left| B_{ij}(t,s) \right|^2 ds\,dt < \infty$$

using $E\int_0^T |y_t|^2 dt < \infty$, it then follows by any of a number of arguments (Contraction Mapping Theorem, Picard Iterations, Inversion of Volterra Operators) that $\hat{y}_s^Z - \hat{y}_s^N = 0$ for all s. Then, immediately from (6), (7) and (4),

$$M_t = N_t = Z_t \quad ,$$

which is the desired result.

7.3 Estimation of Gaussian Processes

Once causal equivalence has been established for linear least-squares estimation, the corresponding result for the case where y and w are jointly Gaussian follows readily by arguments such as those in Section III of [26]. The proof uses the fact that, for y Gaussian,

$$\int_0^t |y_s|^2 ds < \infty \text{ a.s.} \quad \text{iff} \quad E\int_0^t |y_s|^2 ds < \infty$$

to provide a bridge between the two cases.

As mentioned above, a paper presenting these arguments in detail is in preparation.

8. QUANTITATIVE MEASURES OF CONTROLLABILITY AND OBSERVABILITY

8.1 Introduction

Our long-term objective in this recently-begun research effort is to develop quantitative measures of controllability and observability, and to use these as a basis for providing an analytical framework for design and performance evaluation, especially for large-scale or decentralized systems.

The large and highly-developed body of knowledge concerning the structural properties of linear systems is framed almost entirely in terms of "yes" or "no" questions and answers; a state is either reachable or it is not, an input disturbance is either localized away from an output or it is not, a system is decoupled or it is not; in each case, available characterizations afford conditions that can be checked to determine which of the two holds true. Almost all of these involve, directly or indirectly, the controllability and observability properties of the system. There is, however, no body of knowledge relating to the approximate achievement of these goals, or, more generally, of the degree to which they are achieved. For many practical purposes it is sufficient if, for example, an input has an acceptably small influence on an output, and it is not necessary for this influence to be zero. Especially in large systems, some measure of the degree of interaction or noninteraction between subsystems seems essential for analyzing the system, for designing decentralized estimation or compensation schemes, and for assessing the performance of these estimators and controllers (in terms, say, of performance bounds).

We present here some preliminary results from our initial investigation of these questions. For simplicity of presentation, we concentrate on discrete-time constant systems, with occasional references to the continuous-time or time-varying counterparts of these. The next section introduces the measures of controllability and observability we have been working with to this point, while Section 8.3 presents some implications of these in measuring interaction between input and output, and in providing lower bounds for smoothing problems, in order to provide simple illustration of some of the consequences of these controllability and observability measures.

8.2 Measures of Reachability and Observability

Consider the constant, discrete time system

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k \end{aligned} \tag{1a}$$

with $x_k \in R^n$, $u_k \in R^p$ and $y_k \in R^m$. As in [27], we apply the input u_k over [-n,-1], starting with $x_{-n} = 0$, and observe the output y_k over [0,n-1]. Letting

$$\begin{aligned} \underline{u}' &= [u'_{-1}\ u'_{-2}\ \cdots\ u'_{-n}], \\ \underline{y}' &= [y'_0\ y'_1\ \cdots\ y'_{n-1}], \\ F &= [B,\ AB,\ \ldots.,\ A^{n-1}B], \\ H &= \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}, \end{aligned} \tag{2}$$

we then have that the state x_0 at time 0 is given by

$$x_0 = F\,\underline{u} \tag{3}$$

while the output over [0,n-1] is, in terms of x_0,

$$\underline{y} = Hx_0 \tag{4}$$

or, in terms of $\underline{u}$,

$$\underline{y} = HF\,\underline{u} \triangleq G\,\underline{u} \tag{5}$$

where

$$G = HF = \begin{bmatrix} CB & CAB & \cdots & CA^{n-1}B \\ CAB & CA^2B & & \vdots \\ \vdots & & & \vdots \\ CA^{n-1}B & \cdots & & CA^{2n-1}B \end{bmatrix} \tag{6}$$

is the Hankel matrix associated with the system.

One natural way to measure the reachability of a given state x at time 0 is as the maximum value of the inner product $x'x_0$ over all x_0 of the form (3) with $\|\underline{u}\|^2 = \underline{u}'\underline{u} \leq 1$, i.e.,

$$r(x) = \sup_{\underline{u}'\underline{u} \leq 1} x'Fu = \sqrt{x'FF'x} = \|F'x\| \tag{7}$$

where the second equality follows as a result of performing the simple maximization. If x has norm 1, this reduces to maximizing the projection along x of all states reachable with at most one unit of input energy.

Iff $r(x) = 0$ then x is unreachable in the standard sense of the term; i.e., orthogonal to the set of reachable states, so the inner product in (7) is identically zero. Large values of $r(x)$ correspond in some sense to states that are easily reached, though not in the classic sense of the term, since such states may have unreachable components.

It is easily seen that $r(x)$ is a sublinear function of x. It is also convex, so that, in particular, the set of states whose reachability measure is no greater than α, i.e., $\{x: r(x) \leq \alpha\}$, is a convex set for all $\alpha \geq 0$.

The corresponding dual observability measure of a state x at time 0 is simply the norm of the output sequence $\underline{y}$ that is produced by x, i.e.,

$$o(x) = \left\| \underline{y} \right\| = \left\| Hx \right\| = \sqrt{x'H'Hx} \tag{8}$$

We note that $o(x)$ is zero iff the state x is unobservable in the standard sense of the term. Large values of $o(x)$ mean that x gives rise to an output sequence $\underline{y}$ with large norm, and in this sense x is highly observable. The observability measure $o(x)$ is sublinear and convex; the set of all states whose observability measure is less than a given β is convex; i.e., $\{x: o(x) \leq \beta\}$ is convex.

These measures, $r(x)$ and $o(x)$, preserve suitable duality conditions. It is easily checked that the reachability measure, $r^d(x)$, of the system dual to (1) is simply the observability measure $o(x)$ of (1). Conversely, the observability measure, $o^d(x)$, of the system dual to (1) is the reachability measure of (1).

It is convenient, in order to avoid the square roots in (7) and (8), to work with the squares

$$R(x) = r^2(x) \quad , \qquad 0(x) = o^2(x) \tag{9}$$

Another reachability measure is the conjugate functional of R. The conjugate functional of a convex functional R is defined by [see, e.g., 28],

$$R^*(z) = \sup_x [z'x - R(x)] \tag{10}$$

Performing the indicated maximization with R(x) given by (7) and (9), we find

$$R^*(z) = \tfrac{1}{4}\, z'(FF')^{-1}z \ , \tag{11}$$

assuming that FF' is invertible; if it is not, $R^*(z)$ is still defined by (10) and takes the value ∞ if z lies in the nullspace of F'. For simplicity of notation, we assume here that FF' is invertible. Henceforth, we also neglect the constant factor $\frac{1}{4}$ in (11) and take as our alternative reachability measure

$$R^*(z) = z'\,(FF')^{-1}z \tag{11'}$$

This reachability measure (11') has a simple interpretation in terms of the system (1): it is the minimum amount of control energy $\underline{u}'\underline{u}$ needed to reach state z at time 0 from state 0 at time -n, i.e. min $\{\underline{u}'\underline{u}:F\underline{u}=z\}$. This minimum energy is known from standard least-squares linear system theory to be just

$$\underline{u}^{o\prime}\underline{u}^{o} = R^{*}(z) = z'(FF')^{-1}z \ .$$

Thus the measure R^{*} given by (11') is in fact a measure of unreachability: small values of R^{*} correspond to small values of required input energy, and thus to relatively easily reached states; large values of R^{*} correspond to large required input energies and thus to states that are more difficult to reach; in particular, states in the nullspace of F' (and thus unreachable according to the "standard" definition) have $R^{*}(z) = \infty$. We observe that R^{*} is convex, as are all conjugate functionals of convex functionals.

The corresponding conjugate functional of 0 is, again neglecting the constant factor of ¼ as in (11'),

$$0^{*}(z) = z'(H'H)^{-1}z \tag{12}$$

An interpretation of this in terms of the system (1) follows by considering the problem of minimizing the norm of the linear functional on $\underline{y}$ that produces the projection (or, more generally, the inner product) of the initial state x_o on the vector z. Suppose $x_o'z$ is found as the linear functional $\underline{w}'\underline{y}$; the vector $\underline{w}$ of minimum norm that accomplishes this is $\underline{w}^{o} = H(H'H)^{-1}z$, so that $\underline{w}^{o\prime}\underline{w}^{o} = z'(H'H)^{-1}z = 0^{*}(z)$. We thus see that 0^{*} gives, in fact, a measure of unobservability: Small values of $0^{*}(z)$ correspond to a small effort to determine $x_o'z$ and thus to a "more observable" z than do large values of $0^{*}(z)$. Note that $0^{*}(z) = \infty$ if z is in the nullspace of H, corresponding to a state that is unobservable in the standard sense of the term. As with all our measures, 0^{*} is convex.

The measures 0^* and R^* also preserve appropriate duality relations: the observability measure 0^{*d} of the dual to system (1) is the reachability measure R^* of system (1), and vice versa.

As a means of making more concrete the relationship between R and R^*, observe that if FF' is diagonalized by the similarity transformation T, so that $T'FF'T = \Lambda$, then

$$R(x) = x'FF'x = (T'x)'\Lambda(T'x) = \sum_{i=1}^{n} \lambda_i (T'x)_i^2$$

whereas

$$R^*(x) = x'(FF')^{-1}x = (T'x)^1\Lambda^{-1}(T'x) = \sum_{i=1}^{n} \lambda_i^{-1}(T'x)_i^2$$

This makes clearer the earlier observation that R is a measure of reachability, whereas R^* is a measure of unreachability.

Finally, we note that the continuous-time analog of

$$FF' = \sum_{i=1}^{n} A^i BB'A^{i'}$$

is the "controllability Grammian"

$$\int_0^T e^{At} BB' e^{At} dt$$

8.3 Some Implications

We present here two simple implications of the above measures of controllability and observability. The first of these concerns the interaction between an input and an output of a system of the form (1). We emphasize that, in this context, u in (1) may be simply a part of the

total input and y in (1) may be simply part of the total system output, so that we are thinking in terms of possible applications to disturbance rejection or decoupling. We adopt as a measure of interaction between input and output the maximum norm of the output sequence $\underline{y}$ that can be produced by an input sequence $\underline{u}$ with $\underline{u}'\underline{u} \leq 1$, i.e.

$$\max \quad \{\underline{y} : \underline{y} = Hx_o, x_o = F\underline{u}, \underline{u}'\underline{u} \leq 1\}$$
$$= \max_{\|\underline{u}\| \leq 1} \|HFu\| = \|HF\|$$

(Recall that HF is the Hankel matrix of the system (1).) A number of equivalent expressions for this can be given in terms of the above measures; each, upon reflection, has its own intuitive interpretation and provides some insight into the underlying measures. For example

$$\|HF\|^2 = \max_{\underline{u}'\underline{u} \leq 1} \underline{u}'F'H'HF\underline{u} = \max_{\substack{\underline{u}'\underline{u} \leq 1 \\ x=F\underline{u}}} x'H'Hx = \max_{R^*(x) \leq 1} O(x)$$

where the last equality follows after a few lines of algebra and the use of the definition of R^*. One intuitive interpretation of this expression is that large interaction between input and output is a consequence of at least some states whose observability measure $O(x)$ is large also being reasonably reachable (in the sense that the unreachability measure $R^*(x)$ is no greater than 1). If all states having high observability, as measured by $O(x)$, also have low reachability (as measured by a high unreachability $R^*(x)$), then input-output interaction will be small. That interaction between input and output should depend on the reachability and observability of the states is to be expected. What is im-

portant about the above expression for $\|HF\|^2$ is that it quantifies this dependence in terms of specified measures of reachability and observability.

The corresponding dual expression is

$$\|HF\|^2 = \max_{0^*(x) \leq 1} R(x) \tag{13}$$

and this has a dual interpretation to that above.

Taking a Lagrange multiplier approach to performing the maximization in (13) yields, after a few lines of algebra,

$$\|HF\|^2 = \sup \{\underline{\mu}: FF' + \underline{\mu} (H'H)^{-1} \text{ is singular}\}$$

while the corrsponding dual expression is

$$\|HF\|^2 = \sup \{\underline{\nu}: H'H + \underline{\nu} (FF')^{-1} \text{ is singular}\} \quad .$$

In either case, the eigenvector corresponding to the zero eigenvalue so created might be thought of as the state through which maximum input-output interaction takes place.

An alternative expression for $\|HF\|^2$ can be given in terms of the measures o and r. It can be shown after some algebra that

$$\|HF\|^2 = \max_{\|x\| \leq 1} \{o(x)\ r(x)\}$$

Again, this provides the interpretation that low input-output interaction requires easily-reached states (in the sense of having large r(x)) to have low observability (in the sense of small o(x)) and easily-observed states to have low reachability. The quantification of this

expected intuition, and the measures involved, are quite different from those given earlier.

We conclude this section with the second of our two simple applications of the measures of controllability and observability, in this case to providing bounds on the error variance in smoothing problems. Consider the discrete-time stochastic system

$$x_{k+1} = Ax_k + Dv_k$$
$$y_k = Cx_k + w_k$$

where v and w are independent white noise sequences, and the initial state is taken to be unknown. Let the error covariance in the Gauss-Markov (smoothing) estimate of x_0 given $y_0,\ldots,y_{n-1}$ be Σ. Then it is easily shown that

$$\Sigma^{-1} \leq H'H \quad , \qquad \Sigma \geq (H'H)^{-1}$$

where the matrix inequalities denote the usual partial ordering: $P \geq Q$ iff P-Q is nonnegative definite. The above bounds on Σ and Σ^{-1} are, in fact, simply the Cramer-Rao bound for this problem.

In particular,

$$x_0' \Sigma^{-1} x_0 \leq 0(x_0) \ ,$$

and

$$z' \Sigma z \geq 0^*(z) \ ;$$

$0^*(z)$ thus provides (if z is a unit vector) a bound on the error variance in the direction z: the smaller is $0^*(z)$, and thus the more ob-

servable is z, the smaller is the lower bound on the error variance in this direction.

These two simple examples are intended merely to illustrate the kind of results that follow from an analysis in terms of quantitative measures of controllability and observability. Research in this area is continuing.

9. SUMMARY

Two of the research topics included in this report are concerned with estimation, decision, and control problems for observation models other than the familiar "signal in additive white (Gaussian) noise" one. Both involve observations of a doubly-stochastic point process. In one problem we derive and examine the performance of optimal and suboptimal estimation and tracking systems when available observations include a space-time point process; the optimal estimators and controllers are shown to be nonlinear but finite-dimensional (and therefore implementable), and their performance is analyzed in terms of upper and lower bounds, the upper bounds giving the performance of suboptimal schemes that are even more-easily implemented than the corresponding optimum. In the second problem, we derive optimal modulation and demodulation systems for coded, direct-detection optical communication systems under various conditions on the average energy and peak amplitude of the transmitted optical signal; here the received data is a point-process whose intensity is signal-development. A modulation scheme we show to be optimum when average energy constraints are a limiting factor is, in fact, the one employed in a one-gigabit-per-second satellite optical communication system currently under development.

Algorithms have been derived that enable each node in a network to compute its shortest distance to any other node using only local topological information and decentralized information transfer between adjacent nodes. Shortest path algorithms with such decentralized informa-

tion requirements are of obvious importance in many applications, including C^3-systems. A number of modifications of our basic algorithm have been derived, all retaining the basic informationally-decentralized characteristics, but each with its own advantages and limitations in handling various topological changes in the network.

Singular estimation and control problems have been examined from a geometric viewpoint, and various subspaces that are fundamental to the geometric approach to system theory are shown to provide simple, concise solutions to the singular estimation and control problems. These solutions are the same for multi-input, multi-output systems as for single-input, single-output ones; the geometric solution reduces directly to well-known algebraic solutions in the latter case.

Compensator design methods have been investigated for multivariable systems represented in polynomial matrix form. Recent years have seen a reawakening of interest in frequency-domain design techniques, in contrast to the time-domain methods that have predominated for the past two decades, and these have been based principally on polynomial matrix system descriptions. The theory underlying our design procedures draws on the ideas and results of modern algebra, especially multilinear algebra.

A new, direct proof has been derived of the known causal equivalence of the innovations and observations processes for linear estimation and for Gaussian processes.

Finally, we have presented some preliminary results from our recently-begun research effort to develop quantitative measures of controllability and observability that have consequences in terms of mea-

suring such properties as the degree of interaction or noninteraction between input and output and in deriving bounds on estimator or controller performance. Of particular interest in the longer term are applications to large, decentralized system problems.

10. CHRONOLOGICAL LIST OF PUBLICATIONS, CONFERENCE PRESENTATIONS AND WORKSHOP PRESENTATIONS UNDER CONTRACT N00014-76-C-0667

"A Geometric Approach to Singular Estimation Problems," Ian B. Rhodes, 1976 IEEE International Symposium on Information Theory, Ronneby, Sweden, June 1976.

"Estimation and Control Performance for Space-Time Point-Processes," I. B. Rhodes and D. L. Snyder, Proceedings of the Fourteenth Allerton Conference on Circuit and System Theory, University of Illinois, September 1976, pp. 38-51.

"A Separation Theorem for Stochastic Control Problems with Point Process Observations," D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, Automatica, Vol. 13, No. 1, January 1977, pp. 85-87.

"Estimation and Control Performance for Space-Time Point-Process Observations," Ian B. Rhodes and Donald L. Snyder, IEEE Transactions on Automatic Control, Vol. AC-22, No.3, June 1977, pp. 338-346.

"Decentralized Information and Control: Shortest Path Problems," Ian B. Rhodes, Proceedings of the First MIT/ESL-ONR Workshop on Distributed Communication and Decision Problems Motivated by Naval C^3-Systems: Communications and Computer Issues in C^3-Problems, MIT, Cambridge, Massachusetts, August 1-18, 1978, pp. 40-51.

"Signal Optimization for Random Point Processes," D. L. Snyder and Ian B. Rhodes, AFOSR Workshop in Communication Theory and Applications, Provincetown, Massachusetts, September 17-20, 1978.

"Quantization Loss in Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Sixteenth Allerton Conference on Communication, Control, and Computing, University of Illinois, October 4-6, 1978.

"A Decentralized Shortest Path Algorithm," Jeffrey M. Abram and Ian B. Rhodes, Proceedings of the Sixteenth Allerton Conference on Communications, Control and Computing, University of Illinois, October 4-6, 1978, pp. 271-277.

"Some Implications of the Cutoff-Rate Criterion for Coded, Direct-Detection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Biomedical Computer Laboratory Monograph 363, Washington University, St. Louis, MO, March 1979. Submitted to IEEE Transactions on Information Theory.

"Some Implications of the Cutoff Rate Criterion for Coded, Direct-Detection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, 1979 IEEE International Information Theory Symposium, Grignano, Italy, June 25-29, 1979.

"Quaternary Pulse Modulation is Optimal for Optical Communication at One Gigabit Per Second," Donald L. Snyder and Ian B. Rhodes, National Telecommunications Conference, Washington, DC, November 27-29, 1979.

In Preparation

"A Direct Proof of the Innovations Conjecture for Linear Estimation or Gaussian Processes," Ian B. Rhodes.

Doctoral Theses in Progress

Jeffrey M. Abram Topic: Decentralized Shortest Path Algorithms

Olive Y. Liu Topic: Compensator Design for Polynomial Matrix Descriptions of Linear Multivariable Systems

11. REFERENCES

[1] Y. Bar-Shalom "Tracking Methods in a Multitarget Environment," IEEE Transactions on Automatic Control, Vol. AC-23, No. 4, August 1978, pp. 618-626.

[2] John M. Santiago, Jr., "Fundamental Limitations of Optical Trackers," M.S. Thesis, Department of Electrical Engineering, Air Force Institute of Technology, Dayton, Ohio; Report AFIT/GEO/EE/78-4, December 1978.

[3] P. M. Merlin and A. Segall, "A Failsafe Algorithm for Loop-Free Distributed Routing in Data-Communication Networks," EE Pub. No. 313, Faculty of Electrical Engineering, Technion Institute of Technology, Haifa, Israel, September 1977.

[4] W. M. Wonham and A. S. Morse, "Decoupling and Pole Assignment in Linear Multivariable Systems: A Geometric Approach," SIAM Journal Control, Vol. 8, No. 1, February 1970, pp. 1-18.

[5] G. Basile and G. Marro, "Controlled and Conditioned Invariant Subspaces in Linear System Theory," JOTA, Vol. 3, No. 5, 1969, pp. 306-315.

[6] A. S. Morse and W. M. Wonham, "The Status of Non-Interacting Control," IEEE Transactions on Automatic Control, Vol. AC-16, No. 6, December 1971, pp. 568-481.

[7] R. S. Bucy, D. Rappaport, and L. M. Silverman, "Correlated Noise Filtering and Invariant Directions of the Riccati Equation," IEEE Transactions on Automatic Contrl, Vol. AC-15, 1970, pp. 535-540.

[8] M. Gevers and T. Kailath, "Constant, Predictable, and Degenerate Directions of the Discrete-Time Riccati Equation," Automatica, Vol. 9, 1973, pp. 699-711.

[9] D. J. Clements and B. D. O. Anderson, "Linear-Quadratic Discrete-Time Control and Constant Directions," Automatica, Vol. 13, 1977, pp. 255-264.

[10] H. H. Rosenbrock, State-Space and Multivariable Theory, John Wiley, New York, 1970.

[11] V. M. Popov, "Some Properties of Control Systems with Irreducible Matrwix Transfer Functions, Seminar on Differential Equations and Dynamic Systems, Lecture Notes in Mathematics, No. 144, Berlin, 1969, pp. 169-180.

[12] W. A. Wolovich, Linear Multivariable Systems, Applied Mathematical Sciences Series, Vol. II, Springer, New York, 1974.

[13] W. A. Wolovich, "The Differential Operator Approach to Linear System Analysis and Design," J. Franklin Institute, Vol. 33, 1976.

[14] W. A. Wolovich and P. Ferreira, "Output Regulation and Tracking in Linear Multivariable Systems, IEEE Trans. Automatic Control, Vol. AC-24, No. 3, pp. 460-465, 1979.

[15] A. E. Eckberg, Jr., "A Characterization of Linear Systems via Polynomial Matrices and Module Theory," MIT Report ESL-R-528, 1974.

[16] G. D. Forney, Jr., "Minimal Bases of Rational Vector Spaces with Applications to Multivariable Linear Systems," SIAM J. Control, Vol. 13, No. 3, 1975.

[17] S. H. Wang and E. J. Davidson, "A Minimization Algorithm for the Design of Linear Multivariable Systems," IEEE Trans. Automatic Control, Vol. AC-18, No. 3, pp. 220-225, 1973.

[18] L. Cheng and J. B. Pearson, "Frequency Domain Synthesis of Multivariable Linear Regulators," IEEE Trans. Automatic Control, Vol. AC-23, No. 1, pp 3-15, 1977.

[19] M. K. Sain, "A Free Modular Algorithm for Minimal Design of Linear Multivariable Systems," Proc. IFAC Sixth World Congress, Part IB, 1975.

[20] M. K. Sain, "Pole Assignment and A Theorem From Exterior Algebra," Proc. Decision and Control Conference, 1977.

[21] S. MacLane and G. Birkhoff, Algebra, London: MacMillan, 1967.

[22] M. Marcus, Finite Dimensional Multilinear Algebra, Part II," Marcell Dekker, 1975.

[23] N. T. Hung and B. D. O. Anderson, "Triangularization Technique for the Design of Multivariable Control Systems, IEEE Trans. Automatic Control, Vol. AC-24, No. 3, pp. 455-460, 1979.

[24] V. E. Benes, "On Kailath's Innovations Conjecture," Bell System Technical Journal, Vol. 55, 1976, pp. 981-1001.

[25] M. H. A. Davis, Linear Estimation and Stochastic Control, Chapman and Hall, London, 1977.

[26] M. H. A. Davis, "A Direct Proof of Innovations/Observations Equivalence for Gaussian Processes," IEEE Transactions on Information Theory, Vol. IT-24, No. 2, March 1978, pp. 252-254.

[27] R. E. Kalman, P. L. Falb, and M. A. Arbib, Topics in Mathematical System Theory, New York: McGraw-Hill, 1969, Chapter 10.

[28] D. G. Luenberger, Optimization by Vector Space Methods, New York: John Wiley and Sons, 1969, Chapter 7.

APPENDIX 1

Reprint of Paper:

"A Separation Theorem for Stochastic Control Problems with Point Process Observations," D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, Automatica, Vol. 13, No. 1, January 1977, pp. 85-87.

(Pages 81 - 84)

Automatica, Vol. 13, pp. 85–87. Pergamon Press, 1977. Printed in Great Britain

Brief Paper

A Separation Theorem for Stochastic Control Problems with Point-Process Observations†

D. L. SNYDER,‡ I. B. RHODES§ and E. V. HOVERSTEN¶

Key Word Index—Communications control applications; control theory; filtering; Kalman filters; nonlinear filtering; point process; state estimation; stochastic control; tracking systems.

Summary—The exact solution is derived for a stochastic optimal control problem involving a linear stochastic plant, quadratic costs, and nonlinear, nongaussian observations. The observations are in the form of a point process in which each point has both a temporal and a spatial coordinate. The state of the stochastic plant influences the intensity of the observed time–space point process. The solution to this dual control problem can be realized with a separated estimator–controller in which the estimator is nonlinear, mean-square optimal, and finite dimensional, and the controller is the certainty equivalent linear controller. Motivation for the stochastic optimal control problem studied here is given in terms of position sensing and tracking for quantum-limited optical communication problems.

1. *Introduction*

THE MOST general stochastic optimal-control problem is a so-called dual control problem which has been solved only under very restrictive conditions. Of special importance is the *separation theorem* which demonstrates that for a linear stochastic plant, quadratic costs, and linear observations in additive Gaussian noise, the optimal control law can be determined by solving separately and independently a causal stochastic estimation problem and a deterministic control problem. In this paper, we demonstrate that a similar separation holds for the exact solution to a dual control problem involving a linear stochastic plant, quadratic costs, and nonlinear, nongaussian observations. The observations are in the form of a point process in which each point has both a temporal and a spatial coordinate. The state of the stochastic plant influences the intensity of the observed time–space point process. We show that the solution to this dual control problem can be realized with a separated estimator–controller in which the estimator is nonlinear, mean-square optimal, and finite-dimensional, and the controller is the certainty-equivalent linear controller. Motivation for the dual control problem is given in terms of optical position sensing and tracking.

*Received 5 March 1976; revised 21 July 1976. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by associate editor Y. Sunahara.

†This work was supported by the National Science Foundation under Research Grant Number ENG74-07800, the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources, the Air Force Office of Scientific Research, Air Force Systems Command, under Grant AFOSR-73-2496, and the Office of Naval Research under Contract N00014-76-C-0667.

‡Department of Electrical Engineering and Biomedical Computer Laboratory, Washington University, St. Louis, MO 63130, U.S.A.

§Department of System Science and Mathematics, Washington University, St. Louis, MO 63130, U.S.A.

¶COMSAT Corporation, Washington, D.C. 20024, U.S.A.

The point-process observation-model we adopt here is a generalization of that in [1] to include feedback interactions between the observed points and the state of the linear stochastic plant. This interaction implies that the plant state is not a Gaussian process. Even so, we find as in [1] that at any time the state is conditionally Gaussian given observations of the time–space point process up to that time.

2. *Model and problem statement*

We adopt the following model. Denote by $[t_0, \infty)$ and R^m a semi-infinite time interval and an m-dimensional Euclidean space, respectively. We consider as observations a point process on $[t_0, \infty) \times R^m$; thus, each observed point is identified by a temporal coordinate $t \in [t_0, \infty)$ and a spatial coordinate $\mathbf{r} \in R^m$. Let T and A be Borel sets in $[t_0, \infty)$ and R^m, respectively, and denote by $N(T \times A)$ the number of points occurring in $T \times A$. We define $N(t) = N([t_0, t) \times R^m)$ as the number of points up to but not including time t regardless of their spatial location. We use $\mathcal{N}_t$ to denote the sequence of points up to time t; $\mathcal{N}_t$ consists of the number $N(t)$ and time–space coordinates $(t_1, \mathbf{r}_1)$, $(t_2, \mathbf{r}_2) \ldots$, $(t_{N(t)}, \mathbf{r}_{N(t)})$ of all points in $[t_0, t) \times R^m$.

We assume that

$$\lim_{\tau, \rho \downarrow 0} (\tau \rho^m)^{-1} Pr\{N([t, t+\tau) \times c(\mathbf{r}, \rho)) = 1 | \mathcal{N}_t, x(\sigma); \sigma \geq t_0\}$$

$$= \lim_{\tau, \rho \downarrow 0} (\tau \rho^m)^{-1} Pr\{N([t, t+\tau) \times c(\mathbf{r}, \rho)) \geq 1 | \mathcal{N}_t, x(\sigma); \sigma \geq t_0\}$$

$$= \Lambda(t) \exp\{-\tfrac{1}{2}[\mathbf{r} - H(t)x(t)]' R^{-1}(t)[\mathbf{r} - H(t)x(t)]\}, \qquad (1)$$

where: $c(\mathbf{r}, \rho) = [r_1, r_1 + \rho) \times \cdots \times [r_m, r_m + \rho)$ is a cube in R^m; $\Lambda(t)$ is a known function of t; $H(t)$ is a known, $m \times n$ matrix-valued function of t; $R(t)$ is a known, symmetric, positive-definite, $m \times m$ matrix-valued function of t; and $\{x(t); t \geq t_0\}$ is the n-dimensional state of a linear stochastic plant as defined below. Thus, the conditional probability that a single point will be observed in a small time–space volume $[t, t+\tau) \times c(\mathbf{r}, \rho)$ given $\mathcal{N}_t$ and $\{x(\sigma); \sigma \geq t_0\}$ is approximated to order $\tau \rho^m$ by $\lambda(t, r, x(t)) \tau \rho^m$, where we define

$$\lambda(t, r, x(t)) = \Lambda(t) \exp\{-\tfrac{1}{2}[\mathbf{r} - H(t)x(t)]' R^{-1}(t) \times [\mathbf{r} - H(t)x(t)]\}.$$

We assume that the process $\{x(t); t \geq t_0\}$ is defined for $t \geq t_0$ by the following linear stochastic differential equation

$$dx(t) = F(t)x(t)\,dt + G(t)u(t)\,dt + V(t)\,dv(t), \quad x(t_0) = x_0, \qquad (2)$$

where: $F(t)$, $G(t)$, and $V(t)$ are known $n \times n$, $n \times k$, and $n \times \ell$ matrix-valued functions of t, respectively; $\{u(t); t \geq t_0\}$ is a k-dimensional control input; $\{v(t); t \geq t_0\}$ is a standard, ℓ-dimensional Wiener process such that for any time $t \geq t_0$, the future $\{v(\sigma); \sigma \geq t\}$ of v is independent of $\mathcal{N}_t$; and the random initial state, x_0, is assumed to be normal with mean-value vector $\mathbf{x}_0$ and covariance matrix Σ_0.

Consider now the stochastic optimal control problem involving the linear dynamic system (2) and the average quadratic cost functional

$$J[u] = E\left\{\int_{t_0}^{t} [u'(t)P(t)u(t) + x'(t)Q(t)x(t)]\,dt + x'(T)Sx(T)\right\}, \tag{3}$$

where P, Q, and S are given matrices such that $P(t)$ is positive definite for $t \in [t_0, T]$, $Q(t)$ is non-negative definite for $t \in [t_0, T]$, and S is non-negative definite. Attention is restricted to the so-called classical information pattern in which the control input $u(t)$ at each time $t \in [t_0, T]$ depends on the observations $\mathcal{N}_t$ up to that time. In other words, we consider control laws $\mu(\cdot,\cdot)$ that map pairs of the form $(\mathcal{N}_t, t)$ into $u(t) \equiv \mu(\mathcal{N}_t, t)$. In view of this, the symbol $u(t)$ will henceforth denote the control law μ evaluated at $(\mathcal{N}_t, t)$. In order to emphasize that the cost functional (3) therefore depends on the choice of control law μ we shall write $J[\mu]$ instead of $J[u]$. We seek the control law μ_0 that minimizes $J[\mu]$.

It is well-known that when observations of x have the linear, additive form

$$dz(t) = H(t)x(t)\,dt + dw(t), \tag{4}$$

where w is a standard Wiener process, the control law that minimizes $J[\mu]$ is defined by $u_0(t) = -P^{-1}(t)G'(t)K(t)\hat{x}(t)$, where $\hat{x}(t)$ is the causal minimum mean-square-error estimate of $x(t)$ in terms of past data $\{z(\sigma);\ t_0 \le \sigma < t\}$, and $K(t)$ is the precomputable solution to the matrix Riccati equation

$$dK(t)/dt = -K(t)F(t) - F'(t)K(t) + K(t)G(t)P^{-1}(t)G'(t)K(t) - Q(t) \tag{5}$$

with the final condition $K(T) = S$. This result is usually called the *separation theorem* because it shows that the solution to this special version of the stochastic control problem can be obtained by solving separately and independently a least-squares control problem and a least-squares estimation problem. In the next section, we shall show that an analogous separation holds when the observations are in the form of the time–space point process defined above. We note in this instance that neither the plant state nor the observations are normally distributed.

We will need the following lemma, which can be proven in exactly the same manner as Lemma 1 in [1].

Lemma 1. Denote by $p_t(X|\mathcal{N}_t)$ the conditional probability density of $x(t)$ given $\mathcal{N}_t$ for $t \ge t_0$. Then

$$dp_t(X|\mathcal{N}_t) = L[P_t(X|\mathcal{N}_t)]\,dt + p_t(X|\mathcal{N}_t)\int_{R^m} [\lambda(t, r, X) - \hat{\lambda}(t, r)]\hat{\lambda}^{-1}(t, r)N(dt \times dr), \tag{6}$$

where we define

$$\hat{\lambda}(t, r) = E[\lambda(t, r, x(t))|\mathcal{N}_t] = \int \cdots \int_{R^n} \lambda(t, r, X)p_t(X|\mathcal{N}_t)\,dX_1 \cdots dX_n,$$

and where $L[(\cdot)]$ is the following partial-differential operator

$$L[(\cdot)] = -\sum_{i=1}^{n} \partial[FX + Gu]_i(\cdot)/\partial X_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^2[VV']_{ij}(\cdot)/\partial X_i \partial X_j.$$

3. *Estimator–controller solution*

The optimal control law μ_0 that minimizes the average quadratic cost $J[\mu]$ when the observations are of the nonlinear, nongaussian type described above is given in the following proposition.

Proposition 1. Under the above assumptions, the optimal control law μ_0 that minimizes $J[\mu]$ is defined by

$$u_0(t) = -P^{-1}(t)G'(t)K(t)\hat{x}(t), \tag{7}$$

where $K(t)$ satisfies (5), and $\hat{x}(t) = E[x(t)|\mathcal{N}_t]$ satisfies

$$d\hat{x}(t) = F(t)\hat{x}(t)\,dt + G(t)u_0(t)\,dt + \int_{R^m} M(t)[r - H(t)\hat{x}(t)]N(dt \times dr); \quad \hat{x}(t_0) = \bar{x}_0, \tag{8}$$

$$d\hat{\Sigma}(t) = F(t)\hat{\Sigma}(t)\,dt + \hat{\Sigma}(t)F'(t)\,dt + V(t)V'(t)\,dt - \int_{R^m} M(t)H(t)\hat{\Sigma}(t)N(dt \times dr); \quad \hat{\Sigma}(t_0) = \Sigma_0, \tag{9}$$

$$M(t) = \hat{\Sigma}(t)H'(t)[H(t)\hat{\Sigma}(t)H'(t) + R(t)]^{-1}. \tag{10}$$

Furthermore, $x(t)$ is conditionally Gaussian with mean $\hat{x}(t)$ and covariance $\hat{\Sigma}(t)$ given $\mathcal{N}_t$.

Proof of Proposition 1. According to Åström[2], the cost function $J[\mu]$ can be rewritten as

$$J[\mu] = \int_{t_0}^{T} E\{\|u(t) + A(t)\hat{x}(t)\|^2_{P(t)}\}\,dt + E\{x_0'K(t_0)x_0\} + \int_{t_0}^{T} tr[B(t)E(\hat{\Sigma}(t)) + V(t)V'(t)K(t)]\,dt$$

where $\hat{x}(t) = E[x(t)|\mathcal{N}_t]$ is the causal minimum mean square-error estimate of $x(t)$ given $\mathcal{N}_t$, $\hat{\Sigma}(t)$ is the corresponding conditional error covariance given $\mathcal{N}_t$, $A(t) = P^{-1}(t)G'(t)K(t)$, $B(t) = A'(t)P(t)A(t)$, $\|v\|_P^2 = v'Pv$ for any vector v, and $tr\,[\cdot]$ denotes trace. It is evident by virtue of the non-negativity of the first term on the right in this expression that the optimal control law μ_0 is defined by $u_0(t)$ in (7) provided that $\hat{\Sigma}(t)$, and so $E[\hat{\Sigma}(t)]$, is independent of the choice of control law. We now demonstrate this independence by arguing first, that for any causal control law, $x(t)$ is conditionally normal given $\mathcal{N}_t$. This can be verified by using (6) and paralleling the proof by induction of Proposition 1 in [1]; it is found for a control law defined by $u(t) = \mu(\mathcal{N}_t, t)$, that $x(t)$ is conditionally normal given $\mathcal{N}_t$ with a conditional mean and covariance which satisfy (8)–(10) with $u_0(t)$ replaced by $u(t)$. It follows as a special case that these assertions hold for the particular choice $u(t) = u_0(t)$. Now, examination of (9) shows that the only way $\hat{\Sigma}(t)$ can be influenced by the choice of control law is through the point process

$$N(t) = \int_{t_0}^{t}\int_{R^m} N(dt \times dr). \tag{11}$$

However, it is evident (see [3] or [4] for a proof) that $\{N(t);\ t \ge t_0\}$ is a point process with rate function

$$\int_{R^m} \lambda(t, r, x(t))\,dr = (2\pi)^{m/2}\Lambda(t)\det^{1/2}[R(t)]. \tag{12}$$

As this rate function is independent of both $\mathcal{N}_t$ and $\{x(\sigma);\ \sigma \ge t_0\}$, it follows that $\{N(t);\ t \ge t_0\}$ is a Poisson process with a rate that is independent of the choice of control. Hence, $\hat{\Sigma}(t)$ is independent of the control law, and Proposition 1 is then established.

4. *Application to optical tracking*

Communication systems that employ a narrow beam of light as a carrier, star-tracking systems, and infra-red tracking systems all have a requirement for position sensing and active tracking to maintain optical boresight in the presence of a variety of disturbances[5]. The requirements can be quite stringent with a design goal of a few microradians of angular tracking accuracy not being uncommon. The above estimator–controller solution provides a possible tool for the design of an optical tracking system under the following idealized conditions.

Let $I(t, r)$ denote the light intensity at time $t \in [t_0, \infty)$ and position $r \in \mathcal{R}$ of an optical field incident on the photoemissive surface of a two-dimensional photodetector on boresight

and without any motions. Here, $\mathcal{R}$ is a subregion of R^2 corresponding to the photoemissive surface. We assume a Gaussian intensity profile

$$I(t, \mathbf{r}) = I_0(t) \exp\{-\tfrac{1}{2}\mathbf{r}' R^{-1}(t)\mathbf{r}\}.$$

Vibration, beam steering due to propagation of the light beam through atmospheric turbulence, and other effects cause the spot of light on the photoemissive surface to move about in a random fashion. In this case, the intensity profile becomes

$$I(t, \mathbf{r}, y_m(t)) = I_0(t) \exp\{-\tfrac{1}{2}[\mathbf{r} - y_m(t)]' R^{-1}(t)[\mathbf{r} - y_m(t)]\},$$

where $y_m(t)$ models the random motions. We assume that $\{y_m(t);\ t \geq t_0\}$ is derived from a Gaussian diffusion satisfying

$$\begin{aligned} dx_m(t) &= F_m(t)x_m(t)\,dt + V_m(t)\,dv_m(t), \\ y_m(t) &= H_m(t)x_m(t), \end{aligned}$$

where $\{v_m(t);\ t \geq t_0\}$ is a standard Wiener process. The purpose of the tracking controller is to compensate for these random motions in order to maintain optical boresight. Thus, in the presence of a controller to position telescopes, mirrors, or other pointing devices, the intensity becomes

$$\begin{aligned} &I(t, \mathbf{r}, y_m(t), y_p(t)) \\ &\quad = I_0(t) \exp\{-\tfrac{1}{2}[\mathbf{r} - y_m(t) + y_p(t)]' R^{-1}(t)[\mathbf{r} - y_m(t) + y_p(t)]\} \end{aligned}$$

where $y_p(t) - y_m(t)$ is the tracking error. Ideally, this error should be zero, but this cannot be accomplished for two reasons: the position error $y_m(t)$ is unknown and must be estimated from data available at the photodetector output, and the tracking devices will have some inertia so that $y_m(t)$ cannot be tracked instantaneously even if it were known. We model the tracking devices by a linear stochastic plant

$$\begin{aligned} dx_p(t) &= F_p(t)x_p(t)\,dt + G_p(t)u(t)\,dt + V_p(t)\,dv_p(t) \\ y_p(t) &= H_p(t)x_p(t), \end{aligned}$$

where $u(t)$ is the input to the tracking devices from the tracking controller, and $\{v_p(t);\ t \geq t_0\}$ is a standard Wiener process modeling local disturbances such as those due to vibration.

Photoelectron conversions take place in the photoemissive surface at a rate proportional to the incident light intensity [6]. Thus, the photoelectron conversion rate has the form of $\lambda(t, \mathbf{r}, x(t))$ for $(t, \mathbf{r}) \in [t_0, \infty) \times \mathcal{R}$ with $\Lambda(t) = \eta I_0(t)/h\nu$, where η is the quantum efficiency of the photoemitter, h is Planck's constant, and ν is the optical frequency. Here, x is the vector obtained by adjoining x_m and x_p, and H is obtained from H_m and H_p in an obvious way.

The problem of optical tracking is to follow the position of maximum light intensity at time t in terms of photoelectron conversions observed on $[t_0, t) \times \mathcal{R}$. Except for the finiteness of $\mathcal{R}$, this problem is identical to the control problem studied above when photoelectron conversions are identified as points. An approximation that appears reasonable when the beam is small and the tracking errors are small (i.e. fine tracking mode rather than an acquisition mode) compared to the size of the photoemissive surface is to replace $\mathcal{R}$ by $\mathcal{R}^2$. With this approximation, the optical tracking problem is solved by the result in Proposition 1.

5. *Conclusion*

The solution has been given for a stochastic control problem involving observations of a time–space point process. The solution is in terms of a separated estimator–controller in which the estimator is nonlinear but closely related to a linear, discrete-time Kalman–Bucy filter, and the controller is the certainty equivalent linear controller. The estimation performance, $E[\tilde{x}(t)]$, and control performance corresponding to this have not been given and, indeed, appear extremely difficult to evaluate exactly. We have established lower and upper bounds on these performances which will be given in another paper [7].

References

[1] D. L. SNYDER and P. M. FISHMAN: How to track a swarm of fireflies by observing their flashes. *IEEE Trans. Inform. Theory* IT-21, 692–295, Nov. (1975).

[2] K. J. ÅSTRÖM: *Introduction to Stochastic Control Theory*, Section 8.7. Academic Press, New York (1970).

[3] P. M. FISHMAN and D. L. SNYDER: The statistical analysis of space–time point processes. *IEEE Trans. Inform. Theory* IT-22, May (1976).

[4] D. L. SNYDER: *Random Point Processes*, Chap. 7. Wiley, New York (1975).

[5] D. L. SNYDER and E. V. HOVERSTEN: Optical position sensing and tracking. *Proc. Int. Conf. on Communications*, San Francisco, June (1975).

[6] S. KARP, E. L. O'NEILL, and R. M. GAGLIARDI: Communication theory for the free–space optical channel. *Proc. IEEE* 58(10), 1611–1626 Oct. (1970).

[7] I. RHODES and D. SNYDER: Estimation and control performance for space-time point process observations *Proc. 14th Annual Allerton Conf. on Circuit and System Theory*, Univ. of ILL., Urbana, IL, Sept. (1976).

APPENDIX 2

Reprint of Paper:

"Estimation and Control Performance for Space-Time Point-Processes," I. B. Rhodes and D. L. Snyder, Proceedings of the Fourteenth Allerton Conference on Circuit and System Theory, University of Illinois, September 1976, pp. 38-51.

(Pages 85 - 100)

ESTIMATION AND CONTROL PERFORMANCE FOR SPACE-TIME POINT-PROCESS OBSERVATIONS*.

IAN B. RHODES** and DONALD L. SNYDER***

ABSTRACT

Estimation and control problems are examined for a class of models involving a linear system, a quadratic cost, and observations that include a space-time point process as well as the familiar "signal in additive Wiener process" measurements. Motivation for this class of models is given in terms of position sensing and tracking for quantum-limited optical communication problems. These models include as special cases several simpler ones considered previously. As in the simpler cases, the optimum estimator is finite-dimensional and nonlinear, and the optimum controller separates into the optimum estimator followed by the certainty-equivalent control law.

Although the optimum estimator and the optimum controller are finite-dimensional, the corresponding expected error covariance and optimum cost require infinite-dimensional calculations. This motivates the derivation of easily-computed upper and lower bounds on estimator and controller performance. The upper bounds are derived by evaluating exactly the performance of a parametrized family of suboptimum designs; one of these is identified as having smaller performance than any other, thus providing a minimal upper bound within this family. The lower bounds are obtained directly by calculations involving inequalities.

*This work was supported in part by the Office of Naval Research under Contract N00014-76-C-0667, in part by the National Science Foundation under Research Grants ENG74-07800 and ENG76-11565, and in part by the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources.

**Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

***Department of Electrical Engineering and Biomedical Computer Laboratory, Washington University, St. Louis, Missouri 63130.

I. INTRODUCTION

Snyder and Fishman [1] have considered the problem of estimating the Gaussian state of a linear stochastic system from observations of a point process in which each point has both a spatial and a temporal co-ordinate. The state of the system influences the *spatial* component of the intensity of the observed space-time point process: at any given time, the contours of constant spatial intensity are ellipsoids whose common centroid depends linearly on the current system state. The *temporal* component of the intensity is assumed in [1] to be deterministic. The conditional density of the system state at any time given the past of the observation process is shown to be Gaussian, and the conditional mean and the conditional covariance satisfy finite-dimensional nonlinear stochastic differential equations that are driven by the observed space-time point process.

This model has been generalized in [2], [3] to include causal feedback interactions between the observed point process and the state of the linear stochastic system. Although inclusion of a feedback (control) term destroys the Gaussian-ness of the system state process, it does not alter either the Gaussian form of the *conditional* density of the state given past observations or the finite-dimensionality of the stochastic differential equations for the conditional mean and the conditional covariance. These and related properties underly the derivation of a separation theorem for a stochastic optimal control problem involving these system and observation processes and a quadratic cost functional. Motivation for this stochastic control problem is given in [2], [3] in terms of position sensing and tracking for quantum-limited optical communication problems.

In this paper, we first generalize the model of [2], [3] in two ways. On the one hand, the space-time point process observations are supplemented by continuous observations of a linear function of the system state in an additive Wiener process. The optimum estimator for a restricted version of this problem is included in the dissertation [4] and a corresponding separation theorem is to be included in a forthcoming paper [5]. Here we remove the requirement in [4], [5] that the supplementary observations have the same dimensions as the spatial component of the space-time point process. On the other hand, we allow the *temporal* component of the intensity of the observed space-time point process to be itself a random process. Under appropriate independence assumptions, it is shown that the joint problem of estimating the state of the system and the temporal intensity reduces to two separate problems, one of which is that considered in [2], [3] while the other is a standard estimation problem for point process observations having no spatial component, as discussed, e.g., in [6]. All properties needed to extend the separation theorem for stochastic control problems are retained. These two generalizations are discussed later in terms of the optical position-sensing and tracking problem that motivated [2], [3].

Second, we examine estimation and control performance via upper and lower bounds. While in all cases the optimum estimator and the corresponding conditional error covariance satisfy finite-dimensional stochastic differential equations and thus can be computed on-line, both depend on the observed space-time point process and cannot be precomputed. Insofar as the conditional covariance is concerned, this contrasts with the precomputability that holds for the Kalman filter. One is therefore led to consider the expectation of the conditional covariance, both as a natural measure of estimation performance in its own right and because it happens to be the particular measure of estimation performance that determines the optimum cost in the stochastic control problems considered here and in [2], [3]. However, while the expectation of the conditional covariance is deterministic and in principle can be precalculated, this calculation turns out to be infinite-dimensional. With this in mind, we derive in Sections IV and V

easily-precalculable matrix-ordering upper and lower bounds on the expected conditional covariance. The upper bounds are obtained by determining the exact performance of each estimator in a parametrized family of suboptimal estimators whose structure is similar to that of the optimum estimator but for which the mean-square error is precomputable. From within this class, we identify a particular suboptimum estimator whose mean-square error lies at all times below that of any other in the matrix ordering sense. The lower bound is obtained directly using differential and other inequalities.

II. FORMULATION OF THE ESTIMATION AND CONTROL PROBLEMS

Consider the stochastic linear system

$$dx_t = F(t)x_t dt + G(t)u_t dt + V(t)dv_t \tag{1a}$$

$$dz_t = C(t)x_t dt + dw_t; \quad z_0 = 0 \tag{1b}$$

where the state x_t is an n-dimensional random vector, the control u_t is a k-dimensional vector whose measurability is defined later, v and w are independent (normalized) ℓ- and q-dimensional Wiener processes, the random initial state x_0 of (1a) is independent of v and w and is Gaussian with mean $\bar{x}_0$ and covariance Σ_0, and the deterministic uniformly bounded matrix-valued time functions $F(\cdot)$, $G(\cdot)$, $V(\cdot)$ and $C(\cdot)$ have the appropriate dimensions.

In addition to observations of the process z, there are also available observations of a space-time point process defined on $[0,\infty) \times R^m$ as follows. Each point occurrence is identified by a temporal co-ordinate $t \in [0,\infty)$ and a spatial co-ordinate $r \in R^m$. Let τ and A be Borel sets in $[0,\infty)$ and R^m, respectively, and denote by $N(\tau \times A)$ the number of points occurring in $\tau \times A$. We define $N_t = N([0,t) \times R^m)$ to be the number of points up to but not including time t regardless of their spatial location; $N_\cdot$ is taken to be a Poisson counting process with intensity μ, where μ is a stochastic process that is independent of x_0, v and w, and μ_t is almost-surely positive. Given that N has a jump at t (i.e. $N_{t-} \neq N_{t+}$), the spatial location r of the point is taken to be an m-dimensional Gaussian random vector with mean $H(t)x_t$ and known positive definite covariance R(t), where $H(\cdot)$ is a known $m \times n$-matrix valued time function. Given N_s and x_s for $s \geq 0$, the spatial locations are independent random vectors that are independent of all other random entities. Thus the space-time point process can be thought of as having an intensity

$$\lambda_t(r,x_t,\mu_t) = \mu_t \gamma_t(r,x_t) \tag{2}$$

that separates into the product of a temporal component μ_t that underlies the Poisson counting process N and a spatial component

$$\gamma_t(r,x_t) \sim N(H(t)x_t, R(t)) = (2\pi)^{-m/2}[\det R(t)]^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(r-H(t)x_t)' R^{-1}(t)(r-H(t)x_t)\}$$

that gives the density of the spatial location r of the point occurring at t.

Let (Ω,F,P) be the underlying probability space. We denote by Z_t the the sub-σ-algebra of F generated by the process z over the interval $[0,t)$, and by N_t the sub-σ-algebra generated by the space-time point process over $[0,t)$. Let $B_t = Z_t \vee N_t$, the σ-algebra generated by Z_t and N_t. It is assumed throughout that u_t is B_t-measurable and such that the solution to (1a) is well-defined; such controls will be henceforth called *admissible*.

The estimation problem to which we address ourselves is to find the conditional means

$$\hat{x}_t \triangleq E[x_t | B_t], \quad \hat{\mu}_t \triangleq E[\mu_t | B_t] \tag{3}$$

and the corresponding conditional covariances

$$\Sigma_t \triangleq \text{cov}[x_t|\mathcal{B}_t], \quad \Gamma_t \triangleq \text{cov}[\mu_t|\mathcal{B}_t]. \tag{4}$$

The control problem we examine is to find the admissible control $\{u_t: t\in[0,T]\}$ that minimizes the quadratic cost functional

$$J[u] = E\{\int_0^T [u_t'P(t)u_t + x_t'Q(t)x_t]dt + x_T'Sx_T\} \tag{5}$$

where the symmetric uniformly-bounded matrix-valued time functions have the appropriate dimensions with $Q(t)$ and S non-negative definite and $P(t)$ positive difinite.

Our notation is generally as follows: lower case letters denote vectors, upper case letters denote matrices, and script letters denote σ-algebras; v_t denotes a time-indexed random vector, in contrast to $v(t)$ which denotes a time-indexed deterministic vector; everything takes place on the fixed, finite time interval $[0,T]$; $y\sim N(q,Q)$ means that y is Gaussian with mean q and covariance Q; the inequality $P\leq Q$ between symmetric, non-negative definite matrices means that $Q-P$ is non-negative definite.

III. SOLUTION OF THE OPTIMAL ESTIMATION AND CONTROL PROBLEMS

Theorem 1. The conditional density of x_t given $\mathcal{B}_t$ is Gaussian and the conditional mean and the conditional covariance satisfy the finite-dimensional nonlinear stochastic differential equations:

$$d\hat{x}_t = F(t)\hat{x}_t dt + G(t)u_t dt + \Sigma_t C'(t)[dz_t - C(t)\hat{x}_t dt]$$

$$+ \int_{R^m} M_t[r - H(t)\hat{x}_t]N(dt\times dr); \quad \hat{x}_0 = E[x_0] \tag{6}$$

$$d\Sigma_t = F(t)\Sigma_t dt + \Sigma_t F'(t)dt + V(t)V'(t)dt - \Sigma_t C'(t)C(t)\Sigma_t dt$$

$$- M_t H(t)\Sigma_t dN_t; \quad \Sigma_0 = \text{cov}[x_0] \tag{7}$$

where

$$M_t = \Sigma_t H'(t)[H(t)\Sigma_t H'(t) + R(t)]^{-1} \tag{8}$$

If $\text{cov}[x_0]$ is positive definite then Σ_t is almost-surely positive definite and its inverse satisfies the finite-dimensional nonlinear stochastic differential equation

$$d\Sigma_t^{-1} = -\Sigma_t^{-1}F(t)dt - F'(t)\Sigma_t^{-1}dt - \Sigma_t^{-1}V(t)V'(t)\Sigma_t^{-1}dt + C'(t)C(t)dt$$

$$+ H'(t)R^{-1}(t)H(t)dN_t; \quad \Sigma_0^{-1} = (\text{cov}[x_0])^{-1} \tag{9}$$

The conditional density of μ_t given $\mathcal{B}_t$ coincides with the conditional density of μ_t given the σ-algebra $\mathcal{T}_t$ generated by the past of the process N_t, i.e. $f_t(\mu|\mathcal{B}_t) = f_t(\mu|\mathcal{T}_t)$, assuming the control u_t is satisfies a technical property specified in the proof and discussed immediately thereafter.

Proof. The derivation of (6) - (9) we give here parallels the proofs of Lemma 1 and Proposition 1 in [1] which establish the corresponding result for the special case where $u_t \equiv 0$, $\mu_t = \mu(t)$ is deterministic, and $C(t) \equiv 0$ (i.e. the observations z are not available). In outline, the modifications that are made to include each of these generalizations are as follows: the introduction of a $\mathcal{B}_t$-measurable u_t causes no difficulty since u_t is deterministic in all calculations which involve probability measures conditioned on $\mathcal{B}_t$; the presence of nonzero $C(\cdot)$ merely adds an additional term that is familiar for this "signal in Wiener process" observation model; the generalization to random μ_t is handled by temporarily conditioning everything on the σ-algebra M generated by μ over $[0,T]$ and subsequently finding that the stochastic differential equation for the conditional density of x_t given $\mathcal{B}_t$ and M turns out to be independent of M.

Letting $\phi_t = \exp[jy'x_t]$, where $y \in R^n$ is nonrandom, we find using the Ito rule that

$$d\phi_t = \phi_t \psi_t dt + \phi jy'V(t)dv_t$$

where

$$\psi_t = jy'[F(t)x_t + G(t)u_t] - \tfrac{1}{2}y'V(t)V'(t)y$$

Letting $\mathcal{F}_t = \mathcal{B}_t VM$ and defining for the moment $\hat{x}_t = E[x_t|\mathcal{F}_t]$ and $\hat{\lambda}_t(r) = E[\lambda_t(r,x_t,\mu_t)|\mathcal{F}_t]$, it follows from our standing assumptions that, for any Borel set $B \in R^m$, $dz_t - C(t)\hat{x}_t dt$ and $N(dt \times B) - \int_B \hat{\lambda}_t(r)dr\,dt$ are independent, independent-increment processes relative to $\mathcal{F}_t$. We then have, analogously to [1, Eq. 9], that the conditional characteristic function $\hat{M}_t(jy) = E[\phi_t|\mathcal{B}_t VM]$ of x_t given $\mathcal{B}_t VM$ satisfies

$$\begin{aligned} d\hat{M}_t(jy) &= E\{\phi_t\psi_t|\mathcal{F}_t\}dt + E\{\phi_t(x_t-\hat{x}_t)'|\mathcal{F}_t\}C'(t)[dz_t - C(t)\hat{x}_t dt] \\ &\quad + \int_{R^m} E\{\phi_t[\lambda_t(r,x_t,\mu_t) - \hat{\lambda}_t(r)]|\mathcal{F}_t\} \\ &\quad \times \hat{\lambda}_t^{-1}(r)[N(dt \times dr) - \hat{\lambda}_t(r)dr\,dt] \end{aligned}$$

Taking inverse Fourier transforms and simplifying then yields the following stochastic differential equation for the conditional density of x_t given $\mathcal{F}_t$ (c.f. [1, Eq. 5])

$$\begin{aligned} dp_t(X|\mathcal{B}_t VM) &= L[p_t(X|\mathcal{F}_t)]dt + p_t(X|\mathcal{F}_t)[X - \hat{x}_t]'C'(t)[dz_t - C(t)\hat{x}_t dt] \\ &\quad + p_t(X|\mathcal{F}_t)\int_{R^m}[\lambda_t(r,X,\mu_t) - \hat{\lambda}_t(r)]\hat{\lambda}_t^{-1}(r)N(dt \times dr) \end{aligned} \quad (10)$$

where

$$\begin{aligned} L[q] &= -\sum_{i=1}^{n} \partial[(F(t)X + G(t)u_t)q]_i/\partial X_i \\ &\quad + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^2[V(t)V(t)q]_{ij}/\partial X_i \partial X_j \end{aligned}$$

Recalling from (2) that $\lambda_t(r,x_t,\mu_t) = \mu_t \cdot \gamma_t(r,x_t)$, we see that the integrand of the last term in (10) can be rewritten as $[\gamma_t(r,X) - \hat{\gamma}_t(r)]\hat{\gamma}_t^{-1}(r)$, where $\hat{\gamma}_t(r) = E[\gamma_t(r,x_t)|\mathcal{F}_t]$. Noting that $\hat{\gamma}_t(r)$ and $\hat{x}_t$ can be written as integrals involving $p_t(X|\mathcal{F}_t)$, the evolution in time of (10) does not depend in

in any way on μ_t. Thus $p_t(X|\mathcal{B}_t \vee \mathcal{M}) = p_t(X|\mathcal{B}_t)$, and (10) can be rewritten as

$$dp_t(X|\mathcal{B}_t) = L[p_t(X|\mathcal{B}_t)]dt + p_t(X|\mathcal{B}_t)[X - \hat{x}_t]C'(t)[dz_t - C(t)\hat{x}_t dt]$$

$$+ p_t(X|\mathcal{B}_t)\int_{R^m}[\gamma_t(r,X) - \hat{\gamma}_t(r)]\hat{\gamma}_t^{-1}(r)N(dt \times dr) \qquad (11)$$

with the Gaussian distribution $N(\bar{x}_0, \bar{\Sigma}_0)$ as initial condition. Of course, $E[x_t|\mathcal{B}_t \vee \mathcal{M}] = E[x_t|\mathcal{B}_t]$, so the definition of x_t given in the statement of Theorem 1 coincides with the temporary definition introduced in the proof; similar remarks apply to $\hat{\gamma}_t$.

The proof that $p_t(X|\mathcal{B}_t)$ is Gaussian with mean $\hat{x}_t$ given by (6) and covariance Σ_t given by (7) or (9) then follows from a straightforward inductive proof similar to that of Proposition 1 in [1]: in the intervals between point occurrences of the space-time point process, $p_t(X|\mathcal{B}_t)$ evolves according to the first two terms on the right side of (11); this is simply Kushner's equation for linear system (1a) with linear observations (1b), and is known to yield a conditional density $p_t(X|\mathcal{B}_t)$ that is Gaussian with mean $\hat{x}_t$ and covariance Σ_t satisfying (6) and (7) or (9), respectively, with the last term on the right side of each deleted. At those instants when a point occurs in the space-time point process, a jump occurs in $p_t(X|\mathcal{B}_t)$ because of the last term on the right side of (11). However, it turns out that $p_t(X|\mathcal{B}_t)$ remains Gaussian after this jump because it was Gaussian before the jump and because the spatial intensity $\gamma_t(r,X)$ is Gaussian. As in [1], calculation of the last term on the right side of (11) shows that the jump in the conditional mean is given by the last term on the right side of (6), the jump in conditional covariance by the last term on the right side of (7), and the jump in the inverse of the conditional covariance by the last term on the right side of (9).

Finally, to prove that $f_t[\mu|\mathcal{B}_t] = f[\mu|T_t]$, let x^* satisfy $dx_t^* = F(t)x_t^*dt + G(t)u_t dt$; $x_0^* = E[x_0]$. Then $\tilde{x}_t \triangleq x_t - x_t^*$ and $\tilde{z}_t \triangleq z_t - C(t)x_t^*$ satisfy

$$d\tilde{x}_t = F(t)\tilde{x}_t dt + V(t)dv_t; \quad d\tilde{z}_t = C(t)\tilde{x}_t dt + dw_t \qquad (12a)$$

with $\tilde{x}_0 \sim N(0,\bar{\Sigma}_0)$ and $z_0=0$, while $\tilde{r} \triangleq r - Hx^*$ has spatial intensity

$$\tilde{\gamma}_t(\tilde{r},\tilde{x}_t) \sim N(H(t)\tilde{x}_t, R(t)). \qquad (12b)$$

Let $\tilde{\mathcal{Z}}_t$ be the σ-algebra generated by $\tilde{z}$ over $[0,t)$ and let $\tilde{N}_t$ be that generated by the space-time process that is obtained from the original one by leaving $N_\cdot$ unchanged and replacing r by $r-Hx^*$. Then, *under the assumption that* μ_t *is also* $\tilde{\mathcal{Z}}_t \vee \tilde{N}_t$*-measurable*, an argument that parallels the proof of Lemma 1 in [7] shows that $\mathcal{Z}_t \vee N_t = \tilde{\mathcal{Z}}_t \vee \tilde{N}_t$. This assumption is discussed shortly. Thus, it is equivalent to prove that $f_t[\mu|\tilde{\mathcal{Z}}_t \vee \tilde{N}_t] = f_t[\mu|T_t]$. Now, because μ and N are independent of x_0, v and w, so also are they independent of $\tilde{x}$ and $\tilde{z}$; thus, the joint density of μ and the event that $t_1,t_2,\ldots,t_k$ are the occurrence times of N over $[0,t)$ satisfies $f[\mu,t_1,\ldots,t_k|\tilde{\mathcal{Z}}_t] = f[\mu,t_1,..,t_k]$. Equivalently,

$$f[\mu|T_t \vee \tilde{\mathcal{Z}}_t]f[t_1,..,t_k|\tilde{\mathcal{Z}}_t] = f[\mu|T_t]f[t_1,..,t_k]$$

Because N and $\tilde{z}$ are independent, $f[t_1,..,t_k|\tilde{\mathcal{Z}}_t] = f[t_1,..t_k]$ and therefore

$$f[\mu|T_t \vee \tilde{\mathcal{Z}}_t] = f[\mu|T_t] \qquad (12c)$$

Also because the spatial components $\tilde{r}_1,\tilde{r}_2,\ldots\tilde{r}_{N_t}$ are independent of μ given $\tilde{X}_t$ and T_t, we have

$$f[\mu,\tilde{r}_1,\ldots,\tilde{r}_{N_t}|T_t \vee \tilde{X}_t] = f[\mu|T_t \vee \tilde{X}_t]f[\tilde{r}_1,\ldots,\tilde{r}_{N_t}|T_t \vee \tilde{X}_t]$$

where $\tilde{X}_t$ is the σ-algebra generated by $\tilde{x}$ over $[0,t)$. Replacing $\tilde{Z}_t$ by $\tilde{X}_t$ in the argument leading to (12c), we have $f[\mu|T_t \vee \tilde{X}_t] = f[\mu|T_t]$ for the first term on the right side, while the left side can be written $f[\mu|\tilde{r}_1,\ldots,\tilde{r}_{N_t}, T_t \vee \tilde{X}_t] f(r_1,\ldots,r_{N_t}|T_t \vee \tilde{X}_t]$. Cancelling the second term on each side, we are left with

$$(f[\mu|\tilde{N}_t \vee \tilde{X}_t] =) f[\mu|\tilde{r}_1,\ldots,\tilde{r}_{N_t}, T_t \vee \tilde{X}_t] = f[\mu|T_t],$$

which in combination with (12c) gives the required result.

Remark 1. The technical assumption that u_t is $\tilde{N}_t \vee \tilde{Z}_t$-measurable which is required for the proof that $f_t[\mu|\mathcal{B}_t] = f_t[\mu|T_t]$ also arises as a sufficient condition in [7]. A generalization of [7, Theorem 3] shows that this will be the case if u is generated from the past of z, N and r or $\tilde{z}$, N and $\tilde{r}$ using a suitably smooth control law. Specifically, it will be so if μ is generated as a Lipschitz function of the state of a suitably smooth finite-dimensional system; included here, in particular, is a control u_t so generated from $\hat{x}_t$ of (6), which is of interest because this is the case for the optimum control found later.

Remark 2. The stochastic differential equations for $\hat{x}_t$ and Σ_t given in Theorem 1 admit an intuitively simple interpretation. In the intervals between point occurrences of the space-time process, the problem reduces to one of estimating x_t from the observations z; the non-occurrence of further points in these intervals provides information about μ_t but none about x_t because of the separability (2) of λ and our standing assumptions concerning independence. Thus, during these intervals we are left with the standard Kalman filtering problem of estimating the state of the linear system (1a) from the observations (1b); if x_t is conditionally Gaussian at the beginning of each such interval, it remains so throughout with mean and covariance which evolve according to (6) and (8) or (9) with the last term deleted in each. (This, or course, is also reflected in the equation (11) for the conditional density reducing to Kushner's equation during these intervals.) We now observe that x_t is, in fact, conditionally Gaussian at the beginning of each such interval because it is at t=0 and because it remains Gaussian after each point occurrence: indeed, at each occurrence (t,r) of the space-time point process the spatial observation r is an independent observation on a Gaussian random variable with mean Hx_t and covariance R; this is equivalent to a discrete observation of the form

$$r = H(t)x_t + \xi$$

where $\xi \sim N(0,R)$ is independent of x and z. Thus, from standard estimation theory for Gaussian random variables [e.g. 10], the conditional density remains Gaussian and the change in conditional mean and covariance of x_t after accounting for this new observation are

$$d\hat{x}_t \triangleq \hat{x}_{t+} - \hat{x}_t = \Sigma_t H'[H\Sigma H' + R]^{-1}(r - H\hat{x}_t) = M_t(r - H\hat{x}_t)$$

$$d\Sigma_t \triangleq \Sigma_{t+} - \Sigma_t = -\Sigma_t H'[H\Sigma H' + R]^{-1} H\Sigma_t = -M_t H\Sigma_t$$

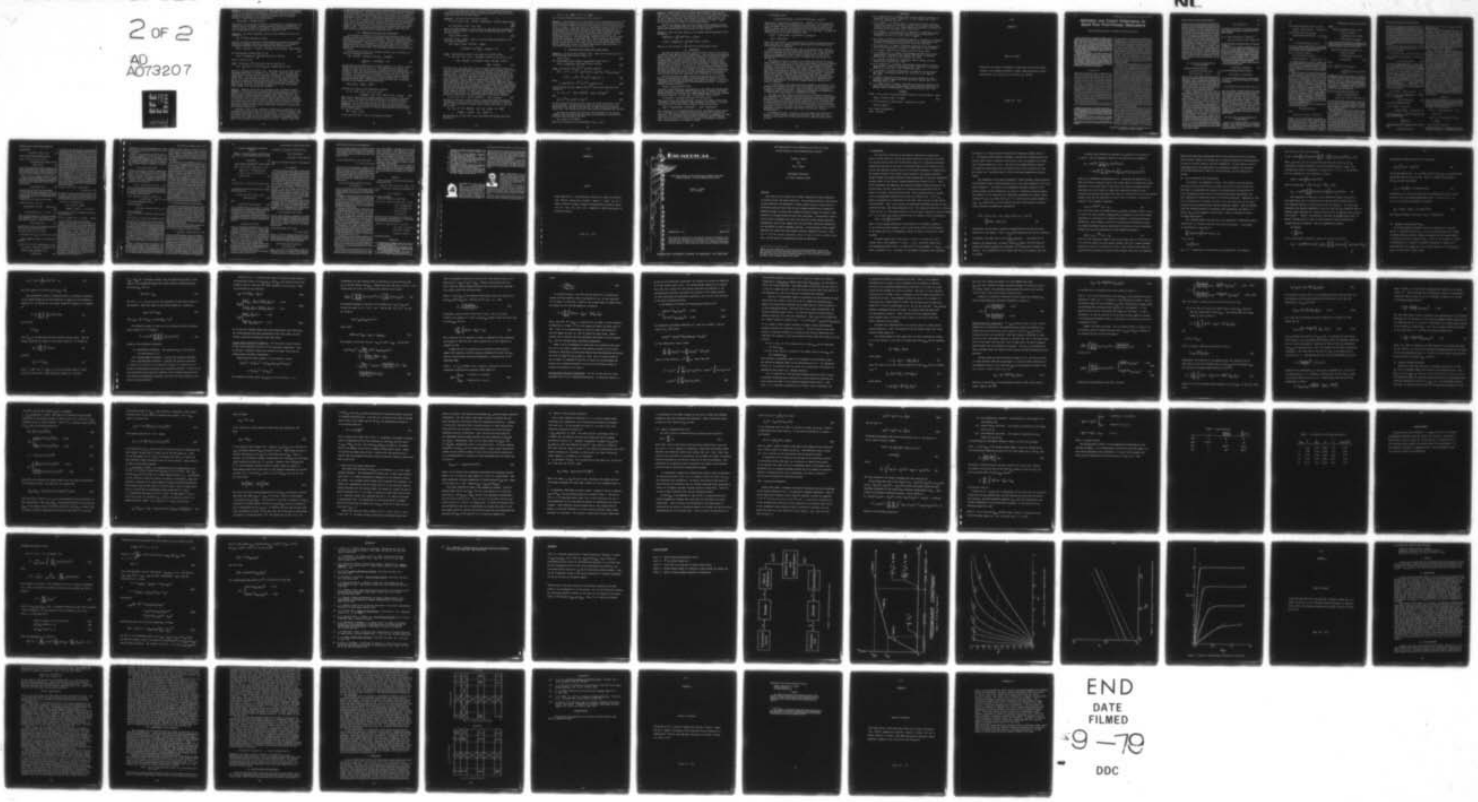
$$d\Sigma_t^{-1} \triangleq \Sigma_{t+}^{-1} - \Sigma_t^{-1} = H'R^{-1}H$$

Of course, this term is to be included only when an occurrence takes place at (t,r); multiplying each of these expressions by $N(dt \times dr)$ and integrating over R^m takes care of this, and constitues the last term in (6), (7) and (9), respectively.

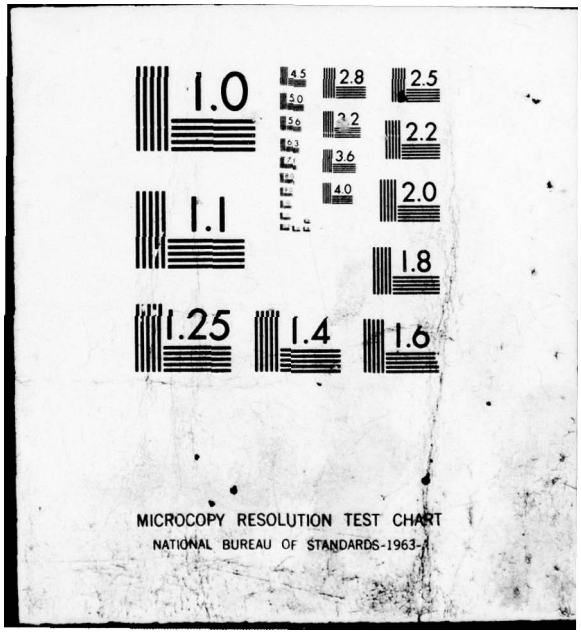
AD-A073 207 WASHINGTON UNIV ST LOUIS MO DEPT OF SYSTEMS SCIENCE --ETC F/G 12/2
DESIGN AND PERFORMANCE EVALUATION FOR SYSTEMS IN AN UNCERTAIN E--ETC(U)
AUG 79 I B RHODES N00014-76-C-0667

UNCLASSIFIED NL

2 OF 2
AD A073207



END
DATE FILMED
9-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Remark 3. It was observed in the proof that $p_t(x|\mathcal{B}_t VM)$ does not depend on M, and thus x_t and μ are conditionally independent given $\mathcal{B}_t$; in particular, x_t and μ_t are conditionally independent given $\mathcal{B}_t$, and thus the joint problem of estimating x_t and μ_t given $\mathcal{B}_t$ separates into two separate problems of estimating x_t and estimating μ_t:

$$f(x_t,\mu_t|\mathcal{B}_t) = f(x_t|\mathcal{B}_t)f(\mu_t|\mathcal{B}_t)$$

Furthermore, the final part of Theorem 1 establishes that $f(\mu_t|\mathcal{B}_t)=f(\mu_t|T_t)$ depends only on the Poisson counting process N and does not depend on z or the spatial locations of the points. This latter estimation problem with various models for μ is examined, for example, in [6].

Theorem 2. The unique admissible control u_t° that minimizes the cost (5) is given by

$$u_t^\circ = -P^{-1}(t)G'(t)K(t)\hat{x}_t \triangleq -L(t)\hat{x}_t \tag{13}$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ satisfies the finite-dimensional nonlinear stochastic differential equation (6) with Σ_t given by (7), and the $n \times n$ symmetric non-negative definite matrix $K(t)$ satisfies the Riccati equation

$$\dot{K}(t) = -K(t)F(t) - F'(t)K(t) + K(t)G(t)P^{-1}(t)G'(t)K(t) - Q(t);\ K(T)=S \tag{14}$$

The corresponding minimum value of J is

$$J[u^\circ] = E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'K\ E\{\Sigma_t\} + KVV']dt, \tag{15}$$

where tr denotes trace.

Proof. According to Åström [8], J[u] can be rewritten as

$$J[u] = \int_0^T E\{||u_t + L(t)x_t||^2_{P(t)}\}dt + [\text{Right side of (15)}]$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ and $||y||^2_P = y'Py$. The first term on the right side is non-negative, and zero if and only if $u_t = -L(t)\hat{x}_t$. Thus (13) gives the unique optimum control provided the right side of (15) is invariant under changes in u. The only way for (15) to be u-dependent is through $E\{\Sigma_t\}$, and (7) shows that the only possibility for Σ_t to vary with u is via N_t. But N_t is a Poisson counting process with rate μ_t, and both N_t and μ_t are specified at the outset as mappings on (Ω,F,P) without any reference to u. Hence Σ_t and, therefore, $E\{\Sigma_t\}$ and the right side of [15] are invariant under changes in u, and the proof is complete.

Remark 4. Theorem 2 shows that the solution to this stochastic control problem can be realized with a spearated estimator-controller in which the estimator is nonlinear, mean-square optimal, and finite-dimensional and the controller is the certainty-equivalent linear control law (i.e. the optimum linear control law for the deterministic problem in which $x_0=E[x_0]$, $v_t = E[v_t]\equiv 0$, and x_t is known exactly). This result, therefore, includes as special eases the familiar linear-quadratic-Gaussian "Separation Theorem" (where the space-time point process observations are absent) [e.g. 8] and the similar results in [1] - [5] for restricted versions of the space-time point process observations.

We observe that Σ_t, the conditional covariance of x_t given $\mathcal{B}_t$, is not precomputable because the last term on the right side of (7) depends on the particular realization of the counting process N_t. One is, therefore, led to consider $\bar{\Sigma}(t) \triangleq E[\Sigma_t]$, both as a natural measure of estimation performance in its own right and as the particular measure of estimation performance that determines the optimum control cost (15). However, while $\bar{\Sigma}(t)$

is deterministic and in principle can be precalculated, this calculation is infinite-dimensional. One way of seeing this is to observe that an attempt to calculate $\bar{\Sigma}(t)$ by taking expectations of both sides of (7) is complicated by the last term on the right side, which requires the calculation of $E\{\Sigma H'[H\Sigma H' + R]^{-1}H\Sigma\}$. While a differential equation for this can be written down, it, in turn, requires expectations of additional nonlinear functions of Σ_t, and so on *ad infinitum* in a mushrooming requirement for additional terms that is familiar from other nonlinear filtering situations. Accordingly, we turn our attention to deriving easily-computed upper and lower bounds on $\bar{\Sigma}(t)$. These estimation bounds then directly imply upper and lower bounds on the optimum control performance (15).

IV. SUBOPTIMUM ESTIMATORS AND UPPER BOUNDS

Our approach to finding easily-computed upper bounds on $E[\Sigma_t]$ is to examine a parametrized family of suboptimum estimators whose mean-square performance can easily be calculated exactly. For each suboptimum estimator the corresponding mean square-error is then trivially a matrix-ordering upper bound on $E[\Sigma_t]$. Furthermore, we show that there exists a suboptimum estimator in this family whose mean-square performance is at all times smaller than that of any other, thus providing a minimal upper bound within this class.

Motivated by the form of the optimum estimator (6), we consider the family of suboptimum estimators

$$dx_t^{\#} = F(t)x_t^{\#}dt + G(t)u_t dt + N(t)[dz_t - C(t)x_t^{\#}dt]$$

$$+\int_{R^m} M(t)[r - H(t)x_t^{\#}]N(dt \times dr) \tag{16}$$

parametrized by the deterministic uniformly-bounded $n \times m$- and $n \times q$-matrix valued time functions $N(\cdot)$ and $M(\cdot)$. This family does not include the optimum estimator (6) in which M_t is a random matrix which depends on N through Σ. Apart from the requirement that $M(\cdot)$, $N(\cdot)$ be deterministic, the sub-optimum estimator (16) and the optimum estimator (6) share the same structure. The nonrandomness of M enables us to write down an ordinary $n \times m$-matrix differential equation for the mean square error of the suboptimum estimator (16). Indeed, subtracting (16) from (1a), it follows directly by straightforward calculation that

$$S(t) \triangleq E[(x_t - x_t^{\#})(x_t - x_t^{\#})'] \tag{17}$$

satisfies the linear matrix differential equation

$$\dot{S} = (F - NC)S + S(F - NC)' + VV' + NN'$$

$$+ \bar{\mu}\{M[HSH' + R]M - MHS - SH'M'\};\ S(0) = \mathrm{cov}[x_0] \tag{18}$$

where we have suppressed the common argument t of all entries, and where $\bar{\mu}(t) = E[\mu_t]$. Because all coefficients in (18) are uniformly bounded, a unique solution to (18) exists for all $t \in [0,\infty)$. We thus have proved:

Theorem 3. For any uniformly bounded $M(\cdot)$ and $N(\cdot)$, the mean-square performance (17) of the estimator (16) satisfies the linear matrix differential equation (18), and this is a matrix ordering upper bound on $E[\Sigma_t]$, i.e. for all $t \in [0,\infty)$,

$$\bar{\Sigma}(t) = E[\Sigma_t] \leq S(t), \tag{19}$$

in the sense that $S(t) - \bar{\Sigma}(t)$ is non-negative definite.

We now show that there exists a choice of $M(\cdot)$ and $N(\cdot)$ in (16) for which the corresponding mean-square performance given by (18) lies at all times below that for any other choice.

Theorem 4. Let $S^*(\cdot)$, $M^*(\cdot)$ and $N^*(\cdot)$ satisfy

$$\dot{S}^* = FS^* + S^*F' + VV' - S^*C'CS^* - \bar{\mu}S^*H'[HS^*H' + R]^{-1}HS^*; \quad S^*(0)=\mathrm{cov}[x_0] \tag{20}$$

$$M^* = S^*H'[HS^*H' + R]^{-1}, \quad N^* = S^*C' \tag{21}$$

where the common argument t of all entries in (20) and (21) is suppressed. Let $S(\cdot)$ be the solution to (18) for some arbitrary $M(\cdot)$ and $N(\cdot)$. Then, for all $t\in[0,\infty)$,

$$E[\Sigma_t] \leq S^*(t) \leq S(t) \tag{22}$$

and $S^*(t) = E\{(x_t - x_t^*)(x_t - x_t^*)'\}$ is the mean-square performance of the bound-minimal estimator

$$dx_t^* = Fx_t^*dt + Gu_t dt + S^*C'[dz_t - Cx_t^*dt] + S^*H'[HS^*H' + R]^{-1}\int_{R^m}[r - Hx_t^*]N(dt \times dr) \tag{23}$$

Proof. Completing the square on the right side of (18) yields

$$\dot{S} = FS + SF' + VV' - SC'CS - \bar{\mu}SH'[HSH' + R]^{-1}HS + (N - SC')(N - SC')' + \bar{\mu}[M - SH'(HSH' + R)^{-1}](HSH' + R)[M - SH'(HSH' + R)^{-1}]',$$

$$S(0) = \mathrm{cov}[x_0] \tag{24}$$

For given $S(t)$, the right side of (23) is clearly minimized by making the last two non-negative definite terms 0, in which case the right side of (23) reduces to that of (20) while the minimizing choices of $M(t)$ and $N(t)$ are given by (21). It remains to show that this instantaneous ordering on the time derivative produces a permanent ordering of the solutions over $[0,T]$, i.e. that the solution to (20) lies at all times below that of (23) in the matrix ordering. This is readily accomplished by using Lemma 1 in (9), after appropriate modifications to reflect that initial conditions, rather than final conditions, are of interest here. This means that in [9, Lemma 1] the left sides of (*) and (**) should be replaced by $+\dot{X}$ and $+\dot{Y}$, respectively, and all time orderings $0\leq t\leq s\leq T$ replaced by $0\leq s\leq t\leq 1$. Then, letting (24) play the role of (*) and (20) the role of (**) and checking conditions 1) to 4) of [9, Lemma 1] we have: 1) and 2) are trivial under our standing assumptions; 3) holds because, by a subsidiary application of [9, Lemma], the solution to (20) lies at all times above that of the Riccati equation

$$\dot{\Gamma} = F\Gamma + \Gamma F' + VV' - \Gamma[C'C + \bar{\mu}H'R^{-1}H] \; ; \quad \Gamma(0) = \mathrm{cov}[x_0] \tag{25}$$

for which 3) is known to hold; finally, 4) holds for (18) and therefore (24) because if $S_1(t)$ and $S_2(t)$ are the respective solutions to (18) with initial conditions $S_1(0)$ and $S_2(0)$, $S_1(0) \geq S_2(0)$, then $S_1(t) \geq S_2(t)$ for all t, since

$$\dot{S}_1 - \dot{S}_2 = (F - NC - \bar{\mu}MH)(S_1 - S_2) + (S_1 - S_2)(F - NC - \bar{\mu}MH)' + \bar{\mu}MH(S_1 - S_2)H'M; \quad (S_1 - S_2)(0) = 0$$

and the solution to this lies at all times above the (identically zero) solution to

$$\dot{Y} = (F - NC - \bar{\mu}MH)Y + Y(F - NC - \bar{\mu}MH)'$$

by a further subsidiary application of [9, Lemma 1].

Remark 5. The evaluation of the performance of the suboptimum estimator (16) is a second-order analysis that uses only the means and covariances of the various random variables and processes and makes no use of the Gaussian-ness of x_0, v, w and r. Thus the results of this section remain valid if v and w are replaced by normalized uncorrelated-increment processes that are uncorrelated with each other and with x_0 and r, with x_0 having any distribution with mean $\bar{x}_0$ and covariance $\bar{\Sigma}_0$ and the spatial intensity $\gamma_t(r,x_t)$ being any distribution with mean Hx and covariance R such that r is uncorrelated with x_0. The bound-minimal estimator (23) can then be viewed as the best estimator in the family (16). These estimators are bilinear because of the product $r.N(dt \times dr)$ in the last term, though they might also be considered to be in a sense linear, to the extent that $N(dt \times dr)$ merely signals the arrival of a spatial observation r which, as with z, is utilized linearly in the production of $\hat{x}$.

V. ESTIMATION LOWER BOUNDS AND CONTROL BOUNDS

Theorem 5. Let S_* be the solution to (25). Then, for all $t \geq 0$, $S_*(t)$ is a matrix-ordering lower bound on $E[\Sigma_t]$, i.e.

$$S_*(t) \leq E[\Sigma_t] \tag{26}$$

The corresponding lower bound on the optimum control cost is

$$J[u^\circ] \geq E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'KS_* + KVV']dt \tag{27}$$

Proof. We have from (9) than $\overline{\Sigma^{-1}} \triangleq E[\Sigma_t^{-1}]$ satisfies

$$\dot{\overline{\Sigma^{-1}}} = -\overline{\Sigma^{-1}}F - F'\overline{\Sigma^{-1}} - E[\Sigma^{-1}VV'\Sigma^{-1}] + \bar{\mu}H'R^{-1}H; \quad \overline{\Sigma^{-1}}(0) = (\mathrm{cov}[x_0])^{-1}$$

$$= -\overline{\Sigma^{-1}}F - F'\overline{\Sigma^{-1}} - \overline{\Sigma^{-1}}VV'\overline{\Sigma^{-1}} + \bar{\mu}H'R^{-1}H - \Delta \tag{28}$$

where

$$\Delta = E[\Sigma^{-1}VV'\Sigma^{-1}] - \overline{\Sigma^{-1}}VV'\overline{\Sigma^{-1}} = \mathrm{cov}[\Sigma^{-1}V] \geq 0 \tag{29}$$

It then follows from [9, Lemma 1] that $\overline{\Sigma^{-1}}$ lies at all times below the solution to

$$\dot{\Xi} = -\Xi F - F'\Xi - \Xi VV'\Xi + \bar{\mu}H'R^{-1}H; \quad \Xi(0) = (\mathrm{cov}[x_0])^{-1} \tag{30}$$

Thus

$$\Xi(t) \geq \overline{\Sigma^{-1}}(t) \triangleq E[\Sigma_t^{-1}] \geq (E\Sigma_t)^{-1}, \tag{31}$$

the last inequality being a matrix version of Jensen's inequality proved in the Appendix. Taking inverses of (31) and noting that if S_* is the solution to (25) then S_*^{-1} is the solution to (30), we have the desired result (26). The control bound (27) then follows by combining (15) and (26).

We remark in passing that (25) gives the covariance of the optimum estimator when the space time point process observations are replaced by continuous observations of the form

$$dy_t = Hx_t dt + (\bar{\mu}^{-1}R)^{\frac{1}{2}}dn_t$$

where n_t is a Wiener process independent of x_0, v and w.

Remark 6. Comparing (20) for the minimal upper bound with (25) for the lower bound, we see that these two bounds will be close to each other and thus to $E[\Sigma_t]$ if HS^*H' is small compared with R (or, equivalently, if $H'R^{-1}H$ is small compared with S^*). Both bounds will also be close to each other and to the optimum performance if the mean intensity μ is small. These are discussed later in terms of our motivating example.

Once we have deduced upper and lower bounds on the estimation performance $E[\Sigma_t]$, corresponding bounds on the optimum control performance follow directly by substitution of these bounds for $E[\Sigma_t]$ in (15):

Theorem 6. Upper and lower bounds on the optimum control performance $J[u^\circ]$ of (15) are

$$E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'PS_* + KVV']dt$$

$$\leq J[u^\circ] \leq E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}[KGP^{-1}G'KS^* + KVV']dt$$

where S_* is the solution to (25) and S^* is the solution to (20).

VI. DISCUSSION

The above estimator-controller solution extends results in [2] and [3] to include a more general form of observation. Just as with the observation model in [2] and [3], this more general observation is motivated by communication systems that employ a narrow beam of light as a carrier, by star tracking systems, and by infra-red tracking systems, all of which have a requirement for position sensing and active tracking to maintain optical alignment in the presence of a variety of disturbances. We shall indicate how the models of [2, Sec. 4] and [3, Sec. 4] are usefully extended by this more general observation. The estimator-controller solution of Theorem 2 provides a possible tool for the design of an optical tracking system under the conditions indicated below, and the performance bounds of Sections IV and V provide the means for predicting the performance of such designs.

Let $I(t,\vec{r})$ denote the light intensity at time $t\in[0,\infty)$ and position $\vec{r}\in\mathcal{R}$ of an optical field incident on the photoemissive surface of a two-dimensional photodetector on boresight and without any motions. Here, $\mathcal{R}$ is a subregion of R^2 corresponding to the photoemissive surface. We assume a Gaussian intensity-profile

$$I(t,\vec{r}) = I_0(t)\exp\{-\tfrac{1}{2}\vec{r}'R^{-1}(t)\vec{r}\}.$$

Vibration, beam steering due to propagation of the light beam through atmospheric turbulence, and other effects cause the spot of light on the photoemissive surface to move about in a random fashion and to fluctuate randomly in optical intensity. In this case, the intensity profile becomes

$$I(t,\vec{r},y_m(t)) = I_0(t)\exp\{-\tfrac{1}{2}[\vec{r}-y_m(t)]'R^{-1}(t)[\vec{r}-y_m(t)]\},$$

where $y_m(t)$ models the random motions, and $I_0(t)$ is a random process (e.g., a lognormal process) that models random intensity fluctuations. We assume that $\{y_m(t);t\geq 0\}$ is derived from a Gaussian diffusion satisfying

$$dx_m(t) = F_m(t)x_m(t)dt + V_m(t)dv_m(t), \quad y_m(t) = H_m(t)x_m(t),$$

where $\{v_m(t);t\geq 0\}$ is a standard Wiener process. The fading process $\{I_0(t);t\geq 0\}$ is assumed to be independent of motion processes but is otherwise arbitrary. The purpose of the tracking controller is to compensate for these random motions and random fading in order to maintain optical alignment. Thus, in the presence of a controller to position telescopes, mirrors, or other pointing devices, the intensity becomes

$$I(t,\vec{r},y_m(t),y_p(t))$$

$$= I_0(t)\exp\{-\tfrac{1}{2}[\vec{r}-y_m(t) + y_p(t)]'R^{-1}(t)[\vec{r}-y_m(t) + y_p(t)]\}$$

where $y_p(t) - y_m(t)$ is the tracking error. Ideally, this error should be zero, but this cannot be accomplished for two reasons: the position error $y_m(t)$ is unknown and must be estimated from data available at the photodetector output, and the tracking devices will have some inertia so that $y_m(t)$ cannot be tracked instantaneously even if it were known. We model the tracking devices by a linear stochastic plant

$$dx_p(t) = F_p(t)x_p(t)dt + G_p(t)u(t)dt + V_p(t)dv_p(t)$$

$$y_p(t) = H_p(t)x_p(t),$$

where $u(t)$ is the input to the tracking devices from the tracking controller, and $\{v_p(t);t\geq 0\}$ is a standard Wiener process modeling local disturbances such as those due to vibration.

Photoelectron conversions take place in the photoemissive surface at a rate proportional to the incident light intenstiy [3]. Thus, the photoelectron conversion rate has the form of $\lambda_t(r,x_t,\mu_t)$ for $(t,\vec{r})\in[0,\infty)\times R$ with μ_t an appropriately scaled version of $I_0(t)$, and x is the vector obtained by adjoining x_m and x_p, and H is obtained from H_m and H_p in an obvious way.

The problem of optical tracking is to follow the position of maximum light intensity at time t in terms of both photoelectron conversions observed on $[0,t)\times R$ and observations of the plant state x_p obtained with sensors located at the tracking devices. These latter observations are modeled according to (1b) so as to account for sensor noise. Except for the finiteness of R, this problem is identical to control problem studied above when photoelectron conversions are identified as space-time points. An approximation that appears reasonable when the beam is small and the tracking errors are small (i.e. fine tracking mode rather than an acquisition mode) compared to the size of the photoemissive surface is to replace R by R^2. With this approximation, the optical tracking problem is solved by the result in Theorem 2.

It is important to note that according to Remark 3 and Theorem 2, the design of the tracking controller does not depend in any way upon the source or nature of randomness in $I_0(t)$. Thus, for example, the same design is obtained if $I_0(t)$ is random due to atmospheric turbulence or modulation by an information-bearing signal or a combination of these.

The upper and lower bounds of Sections IV and V provide a measure of the performance for the optical position-sensing and tracking system derived from Theorem 2. From Remark 6, the upper and lower bounds merge when HS*H' is small compared to the beam spread as measured by R. It is evident that the estimation and control lower bounds derived as above for observations of each photoelectron conversion are also lower bounds for both optimal and suboptimal trackers that employ observations obtained by temporal or spatial averaging as would be obtained using photon counting and a quadrant photomultiplier.

We mention also that A. Segall in [11] has applied the models of [1] and [3] to study computer communication networks. The upper and lower bounds on performance that we have derived can be applied in this context as well.

REFERENCES

1. D. L. Snyder and P. M. Fishman, "How to track a swarm of fireflies by observing their flashes", *IEEE Trans. on Info. Theory*, Vol. IT-21, No. 6, pp. 692-695, Nov. 1975.

2. D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, "An exact estimator-controller solution to a stochastic optimal control problem with point process observations", Proc. Sixth Symposium on Nonlinear Estimation and Its Applications, San Diego, Calif., Sept. 1975.

3. D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, "A separation theorem for stochastic control problems with point process observations", *Automatica*, Vol. 13, No. 1, Jan. 1977, to appear.

4. M. V. Vaca, "A measure transformation approach to estimation and decision for observations derived from martingales", D.Sc. dissertation, Sever Institute of Washington University, St. Louis, Mo., May 1975.

5. M. V. Vaca and D. L. Snyder, "Estimation and decision for observations derived from martingales", submitted to IEEE Trans. on Info. Theory.

6. D. L. Snyder, *Random Point Processes*, New York: Wiley, 1975.

7. I. B. Rhodes and A. S. Gilman, "Cone-bounded nonlinearities and mean-square bounds - estimation lower bound", *IEEE Trans. on Automatic Control*, Vol. AC-20, No. 5, pp. 632-642, Oct. 1975.

8. K. J. Åström, *Introduction to Stochastic Control Theory*, New York: Academic Press, Section 8.7, 1970.

9. A. S. Gilman and I. B. Rhodes, "Cone-bounded nonlinearities and mean-square bounds - quadratic regulation bounds", *IEEE Trans. on Automatic Control*, Vo. AC-21, No. 4, pp. 472-483, Aug. 1976.

10. I. B. Rhodes, "A tutorial introduction to estimation and filtering", *IEEE Trans. on Automatic Control*, Vol. AC-16, No. 6, pp. 688-706, Dec. 1971.

11. A. Segall, "Centralized and distributed control schemes for Gauss-Poisson processes", Report ESL-P-669, Electronic Systems Lab., M.I.T., Cambridge, Mass., 1976.

12. I. B. Rhodes and D. L. Snyder, "Estimation and control performance for space-time point-process observations", Report SSM 7610, Department of Systems Science and Mathematics, Washington University, Sept. 1976.

APPENDIX

Lemma. Let W_1 and W_2 be positive definite matrices, and let $\gamma \in [0,1]$. Then

$$[\gamma W_1 + (1-\gamma)W_2] \leq \gamma W_1^{-1} + (1-\gamma)W_2^{-1} \tag{A1}$$

i.e. W^{-1} is convex in a matrix sense. Furthermore, we have

$$E[W^{-1}] \geq (E[W])^{-1}$$

(c.f. Jensen's inequality)

Proof. See [12].

APPENDIX 3

Reprint of Paper:

"Estimation and Control Performance for Space-Time Point-Process Observations," Ian B. Rhodes and Donald L. Snyder, IEEE Transactions on Automatic Control, Vol. AC-22, No. 3, June 1977, pp. 338-346.

(Pages 101 - 112)

Estimation and Control Performance for Space-Time Point-Process Observations

IAN B. RHODES, MEMBER, IEEE, AND DONALD L. SNYDER, MEMBER, IEEE

Abstract—**Estimation and control problems are examined for a class of models involving a linear system, a quadratic cost, and observations that include a space-time point process as well as the familiar "signal in additive Wiener process" measurements. Motivation for this class of models is given in terms of position sensing and tracking for quantum-limited optical communication problems. These models include as special cases several simpler ones considered previously. As in the simpler cases, the optimum estimator is finite-dimensional and nonlinear, and the optimum controller separates into the optimum estimator followed by the certainty-equivalent control law.**

Although the optimum estimator and the optimum controller are finite-dimensional, the corresponding expected error covariance and optimum cost require infinite-dimensional calculations. This motivates the derivation of easily-computed upper and lower bounds on estimator and controller performance. The upper bounds are derived by evaluating exactly the performance of a parametrized family of suboptimum designs; one of these is identified as having smaller performance than any other, thus providing a minimal upper bound within this family. The lower bounds are obtained directly by calculations involving inequalities.

I. Introduction

SNYDER and Fishman [1] have considered the problem of estimating the Gaussian state of a linear stochastic system from observations of a point process in which each point has both a spatial and a temporal coordinate. The state of the system influences the *spatial* component of the intensity of the observed space-time point process: at any given time, the contours of constant spatial intensity are ellipsoids whose common centroid depends linearly on the current system state. The *temporal* component of the intensity is assumed in [1] to be deterministic. The conditional density of the system state at any time given the past of the observation process is shown to be Gaussian, and the conditional mean and the conditional covariance satisfy finite-dimensional nonlinear stochastic differential equations that are driven by the observed space-time point process.

This model has been generalized in [2] and [3] to include causal feedback interactions between the observed point process and the state of the linear stochastic system. Although inclusion of a feedback (control) term destroys the Gaussianness of the system state process, it does not alter either the Gaussian form of the *conditional* density of the state given past observations or the finite-dimensionality of the stochastic differential equations for the conditional mean and the conditional covariance. These and related properties underly the derivation of a separation theorem for a stochastic optimal control problem involving these system and observation processes and a quadratic cost functional. Motivation for this stochastic control problem is given in [2] and [3] in terms of position sensing and tracking for quantum-limited optical communication problems.

In this paper, we first generalize the model of [2] and [3] in two ways. On the one hand, the space-time point-process observations are supplemented by continuous observations of a linear function of the system state in an additive Wiener process. The optimum estimator for a restricted version of this problem is included in the dissertation of Vaca [4] and a corresponding separation theorem is to be included in a forthcoming paper [5]. Here we remove the requirement in [4] and [5] that the supplementary observations have the same dimensions as the spatial component of the space-time point process. On the other hand, we allow the *temporal* component of the intensity of the observed space-time point process to be itself a random process. Under appropriate independence assumptions, it is shown that the joint problem of estimating the state of the system and the temporal intensity reduces to two separate problems, one of which is that considered in [2] and [3] while the other is a standard estimation problem for point-process observations having no spatial component, as discussed, e.g., in [6]. All properties needed to extend the separation theorem for stochastic control problems are retained. These two generalizations are discussed later in terms of the optical position-sensing and tracking problem that motivated [2] and [3].

Second, we examine estimation and control performance via upper and lower bounds. While in all cases the optimum estimator and the corresponding conditional error covariance satisfy finite-dimensional stochastic differential equations and thus can be computed on-line, both depend on the observed space-time point process and cannot be precomputed. Insofar as the conditional covariance is concerned, this contrasts with the precomputability that holds for the Kalman filter. One is therefore led to consider the expectation of the conditional

Manuscript received October 8, 1976. Paper recommended by Y. Bar-Shalom, Chairman of the IEEE S-CS Stochastic Control Committee. This work was supported in part by the Office of Naval Research under Contract N00014-76-C-0667, the National Science Foundation under Research Grants ENG74-07800 and ENG76-11565, and the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources.

I. B. Rhodes is with the Department of Systems Science and Mathematics, Washington University, St. Louis, MO, on leave at the Department of Electrical Engineering, University of California, Berkeley, CA 94720.

D. L. Snyder is with the Department of Electrical Engineering and Boimedical Computer Laboratory, Washington University, St. Louis, MO 63130.

covariance, both as a natural measure of estimation performance in its own right and because it happens to be the particular measure of estimation performance that determines the optimum cost in the stochastic control problems considered here and in [2] and [3]. However, while the expectation of the conditional covariance is deterministic and in principle can be precalculated, this calculation turns out to be infinite-dimensional. With this in mind, we derive in Sections IV and V easily-precalculable matrix-ordering upper and lower bounds on the expected conditional covariance. The upper bounds are obtained by determining the exact performance of each estimator in a parametrized family of suboptimal estimators whose structure is similar to that of the optimum estimator but for which the mean-square error is precomputable. From within this class, we identify a particular suboptimum estimator whose mean-square error lies at all times below that of any other in the matrix ordering sense. The lower bound is obtained directly using differential and other inequalities.

II. Formulation of the Estimation and Control Problems

Consider the stochastic linear system

$$dx_t = F(t)x_t\,dt + G(t)u_t\,dt + V(t)\,dv_t \tag{1a}$$

$$dz_t = C(t)x_t\,dt + dw_t, \qquad z_0 = 0 \tag{1b}$$

where the state x_t is an n-dimensional random vector, the control u_t is a k-dimensional vector whose measurability is defined later, v and w are independent (normalized) l- and q-dimensional Wiener processes, the random initial state x_0 of (1a) is independent of v and w and is Gaussian with mean $\bar{x}_0$ and covariance $\bar{\Sigma}_0$, and the deterministic uniformly bounded matrix-valued time functions $F(\cdot)$, $G(\cdot)$, $V(\cdot)$, and $C(\cdot)$ have the appropriate dimensions.

In addition to observations of the process z, there are also available observations of a space-time point process defined on $[0,\infty)\times R^m$ as follows. Each point occurrence is identified by a temporal coordinate $t\in[0,\infty)$ and a spatial coordinate $r\in R^m$. Let τ and A be Borel sets in $[0,\infty)$ and R^m, respectively, and denote by $N(\tau\times A)$ the number of points occurring in $\tau\times A$. We define $N_t = N([0,t)\times R^m)$ to be the number of points up to but not including time t regardless of their spatial location; $N_\cdot$ is taken to be a doubly stochastic Poisson counting process with intensity μ, with μ and N stochastic processes that are independent of x_0, v, and w, and μ_t is almost-surely positive. Given that N has a jump at t (i.e., $N_{t-} \neq N_{t+}$), the spatial location r of the point is taken to be an m-dimensional Gaussian random vector with mean $H(t)x_t$ and known positive definite covariance $R(t)$, where $H(\cdot)$ is a known $m\times n$-matrix valued time function. Given N_s and x_s for $s>0$, the spatial locations are independent random vectors that are independent of all other random entities. Thus, the space-time point process can be thought of as having an intensity

$$\lambda_t(r,x_t,\mu_t) = \mu_t\gamma_t(r,x_t) \tag{2}$$

that separates into the product of a temporal component μ_t that underlies the Poisson counting process N and a spatial component

$$\gamma_t(r,x_t) \sim N(H(t)x_t, R(t)) = (2\pi)^{-m/2}\left[\det R(t)\right]^{-1/2} \cdot \exp\left\{-\tfrac{1}{2}(r-H(t)x_t)'R^{-1}(t)(r-H(t)x_t)\right\} \tag{3}$$

that gives the density of the spatial location r of the point occurring at t.

Let $(\Omega,\mathcal{F},P)$ be the underlying probability space. We denote by $\mathcal{Z}_t$ the sub-σ-algebra of $\mathcal{F}$ generated by the process z over the interval $[0,t)$, and by $\mathcal{N}_t$ the sub-σ-algebra generated by the space-time point process over $[0,t)$. Let $\mathcal{B}_t = \mathcal{Z}_t \vee \mathcal{N}_t$ be the smallest σ-algebra containing $\mathcal{Z}_t$ and $\mathcal{N}_t$. It is assumed throughout that u_t is $\mathcal{B}_t$-measurable and such that the solution to (1a) is well-defined; such controls will henceforth be called *admissible*.

The estimation problem to which we address ourselves is to find the conditional means

$$\hat{x}_t \triangleq E[x_t|\mathcal{B}_t], \quad \hat{\mu}_t \triangleq E[\mu_t|\mathcal{B}_t] \tag{4}$$

and the corresponding conditional covariances

$$\Sigma_t \triangleq \operatorname{cov}[x_t|\mathcal{B}_t], \quad \Gamma_t \triangleq \operatorname{cov}[\mu_t|\mathcal{B}_t].$$

The control problem we examine is to find the admissible control $\{u_t : t\in[0,T]\}$ that minimizes the quadratic cost functional

$$J[u] = E\left\{\int_0^T \left[u_t'P(t)u_t + x_t'Q(t)x_t\right]dt + x_T'Sx_T\right\} \tag{5}$$

where the symmetric uniformly-bounded matrix-valued time functions have the appropriate dimensions with $Q(t)$ and S nonnegative definite and $P(t)$ positive definite.

Our notation is generally as follows: lower case italic letters denote vectors, upper case italic letters denote matrices, and script letters denote σ-algebras; v_t denotes a time-indexed random vector, in contrast to $v(t)$ which denotes a time-indexed deterministic vector; everything takes place on the fixed, finite time interval $[0,T]$; $y\sim N(q,Q)$ means that y is Gaussian with mean q and covariance Q; the inequality $P\leqslant Q$ between symmetric, nonnegative definite matrices means that $Q-P$ is nonnegative definite.

III. Solution of the Optimal Estimation and Control Problems

Theorem 1: The conditional density of x_t given $\mathcal{B}_t$ is Gaussian and the conditional mean and the conditional covariance satisfy the finite-dimensional nonlinear stochastic differential equations

$$d\hat{x}_t = F(t)\hat{x}_t\,dt + G(t)u_t\,dt + \Sigma_t C'(t)\big[dz_t - C(t)\hat{x}_t\,dt\big] + \int_{R^m} M_t\big[r - H(t)\hat{x}_t\big]N(dt\times dr), \qquad \hat{x}_0 = E[x_0] \quad (6)$$

$$d\Sigma_t = F(t)\Sigma_t\,dt + \Sigma_t F'(t)\,dt + V(t)V'(t)\,dt - \Sigma_t C'(t)C(t)\Sigma_t\,dt - M_t H(t)\Sigma_t\,dN_t, \qquad \Sigma_0 = \operatorname{cov}[x_0] \quad (7)$$

where

$$M_t = \Sigma_t H'(t)\big[H(t)\Sigma_t H'(t) + R(t)\big]^{-1}. \quad (8)$$

If $\operatorname{cov}[x_0]$ is positive definite then Σ_t is almost-surely positive definite and its inverse satisfies the finite-dimensional nonlinear stochastic differential equation

$$d\Sigma_t^{-1} = -\Sigma_t^{-1}F(t)\,dt - F'(t)\Sigma_t^{-1}\,dt - \Sigma_t^{-1}V(t)V'(t)\Sigma_t^{-1}\,dt + C'(t)C(t)\,dt + H'(t)R^{-1}(t)H(t)\,dN_t, \qquad \Sigma_0^{-1} = (\operatorname{cov}[x_0])^{-1}. \quad (9)$$

The conditional density of μ_t given $\mathcal{B}$ coincides with the conditional density of μ_t given the σ-algebra $\mathcal{T}_t$ generated by the past of the temporal process N_t, i.e., $f_t(\mu|\mathcal{B}_t) = f_t(\mu|\mathcal{T}_t)$, assuming the control u_t satisfies a technical property specified in the proof and discussed immediately thereafter.

Proof: The derivation of (6)–(9) we give here parallels the proofs of [1, lemma 1 and proposition 1] which establish the corresponding result for the special case where $u_t \equiv 0$, $\mu_t = \mu(t)$ is deterministic, and $C(t) \equiv 0$ (i.e., the observations z are not available). In outline, the modifications that are made to include each of these generalizations are as follows: the introduction of a $\mathcal{B}_t$-measurable u_t causes no difficulty since u_t is deterministic in all calculations which involve probability measures conditioned on $\mathcal{B}_t$; the presence of nonzero $C(\cdot)$ merely adds an additional term that is familiar for this "signal in Wiener process" observation model; the generalization to random μ_t is handled by temporarily conditioning everything on the σ-algebra $\mathcal{M}$ generated by μ over $[0, T]$ and subsequently finding that the stochastic differential equation for the conditional density of x_t given $\mathcal{B}_t$ and $\mathcal{M}$ turns out to be independent of $\mathcal{M}$.

Letting $\phi_t = \exp[jy'x_t]$ where $y \in R^n$ is nonrandom, we find, using the Itô rule, that

$$d\phi_t = \phi_t\psi_t\,dt + \phi jy'V(t)\,dv_t$$

where

$$\psi_t = jy'\big[F(t)x_t + G(t)u_t\big] - \tfrac{1}{2}y'V(t)V'(t)y.$$

Letting $\mathcal{F}_t = \mathcal{B}_t \vee \mathcal{M}$ and defining for the moment $\hat{x}_t = E[x_t|\mathcal{F}_t]$ and $\hat{\lambda}_t(r) = E[\lambda_t(r, x_t, \mu_t)|\mathcal{F}_t]$, it follows from our standing assumptions that, for any Borel set $B \in R^m$, $dz_t - C(t)\hat{x}_t\,dt$ and $N(dt\times B) - \int_B \hat{\lambda}_t(r)\,dr\,dt$ are independent, independent-increment processes relative to $\mathcal{F}_t$. We then have, analogously to [1, eq. (9)], that the conditional characteristic function $\hat{M}_t(jy) = E[\phi_t|\mathcal{F}_t]$ of x_t given $\mathcal{F}_t$ satisfies

$$d\hat{M}_t(jy) = E\{\phi_t\psi_t|\mathcal{F}_t\}\,dt + E\{\phi_t(x_t - \hat{x}_t)'|\mathcal{F}_t\}\cdot C'(t)\big[dz_t - C(t)\hat{x}_t\,dt\big] + \int_{R^m} E\big\{\phi_t\big[\lambda_t(r, x_t, \mu_t) - \hat{\lambda}_t(r)\big]|\mathcal{F}_t\big\}\cdot\hat{\lambda}_t^{-1}(r)\big[N(dt\times dr) - \hat{\lambda}_t(r)\,dr\,dt\big].$$

Taking inverse Fourier transforms and simplifying then yields the following stochastic differential equation for the conditional density of x_t given $\mathcal{F}_t$ (cf. [1, eq. (5)]):

$$dp_t(X|\mathcal{F}_t) = \mathcal{L}\big[p_t(X|\mathcal{F}_t)\big]\,dt + p_t(X|\mathcal{F}_t)\big[X - \hat{x}_t\big]'\cdot C'(t)\big[dz_t - C(t)\hat{x}_t\,dt\big] + p_t(X|\mathcal{F}_t)\cdot\int_{R^m}\big[\lambda_t(r, X, \mu_t) - \hat{\lambda}_t(r)\big]\cdot\hat{\lambda}_t^{-1}(r)N(dt\times dr) \quad (10)$$

where

$$\mathcal{L}[q] = -\sum_{i=1}^{n}\partial\big[(F(t)X + G(t)u_t)q\big]_i/\partial X_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial^2\big[V(t)V'(t)q\big]_{ij}/\partial X_i\partial X_j.$$

Recalling from (2) that $\lambda_t(r, x_t, \mu_t) = \mu_t\gamma_t(r, x_t)$, we see that the integrand of the last term in (10) can be rewritten as $[\gamma_t(r, X) - \hat{\gamma}_t(r)]\hat{\gamma}_t^{-1}(r)$, where $\hat{\gamma}_t(r) = E[\gamma_t(r, x_t)|\mathcal{F}_t]$. Noting that $\hat{\gamma}_t(r)$ and $\hat{x}_t$ can be written as integrals involving $p_t(X|\mathcal{F}_t)$, the evolution in time of (10) does not depend in any way on μ_t. Furthermore, $p_0(X|\mathcal{F}_0 \sim N(\bar{x}_0, \bar{\Sigma}_0))$ is independent of μ by assumption. Thus, $p_t(X|\mathcal{B}_t \vee \mathcal{M}) = p_t(X|\mathcal{B}_t)$, and (10) can be rewritten as

$$dp_t(X|\mathcal{B}_t) = \mathcal{L}\big[p_t(X|\mathcal{B}_t)\big]\,dt + p_t(X|\mathcal{B}_t)\big[X - \hat{x}_t\big]'\cdot C'(t)\big[dz_t - C(t)\hat{x}_t\,dt\big] + p_t(X|\mathcal{B}_t)\cdot\int_{R^m}\big[\gamma_t(r, X) - \hat{\gamma}_t(r)\big]\hat{\gamma}_t^{-1}(r)N(dt\times dr) \quad (11)$$

with the Gaussian distribution $N(\bar{x}_0, \bar{\Sigma}_0)$ as initial condition. Of course, $E[x_t|\mathcal{B}_t \vee \mathcal{M}] = E[x_t|\mathcal{B}_t]$, so the definition of $\hat{x}_t$ given in the statement of Theorem 1 coincides with the temporary definition introduced in the proof; similar remarks apply to $\hat{\gamma}_t$.

The proof that $p_t(X|\mathcal{B}_t)$ is Gaussian with mean $\hat{x}_t$ given by (6) and covariance Σ_t given by (7) or (9) then follows from a straightforward inductive proof similar to that of [1, proposition 1]: in the intervals between point occurrences of the space-time point process, $p_t(X|\mathcal{B}_t)$ evolves according to the first two terms on the right side of (11); this is simply Kushner's equation for linear system (1a) with linear observations (1b), and is known to yield a conditional density $p_t(X|\mathcal{B}_t)$ that is Gaussian with mean $\hat{x}_t$ and covariance Σ_t satisfying (6) and (7) or (9), respectively, with the last term on the right side of each deleted. At those instants when a point occurs in the space-time

point process, a jump occurs in $p_t(X|\mathcal{B}_t)$ because of the last term on the right side of (11). However, it turns out that $p_t(X|\mathcal{B}_t)$ remains Gaussian after this jump because it was Gaussian before the jump and because the spatial intensity $\gamma_t(r,X)$ is Gaussian. As in [1], calculation of the last term on the right side of (11) shows that the jump in the conditional mean is given by the last term on the right side of (6), the jump in conditional covariance by the last term on the right side of (7), and the jump in the inverse of the conditional covariance by the last term on the right side of (9).

Finally, to prove that $f_t[\mu|\mathcal{B}_t]=f[\mu|\mathcal{T}_t]$, let x^* satisfy $dx_t^* = F(t)x_t^*\,dt + G(t)u_t\,dt$; $x_0^* = E[x_0]$. Then $\tilde{x}_t \triangleq x_t - x_t^*$ and $\tilde{z}_t \triangleq z_t - C(t)x_t^*$ satisfy

$$d\tilde{x}_t = F(t)\tilde{x}_t\,dt + V(t)\,dv_t, \quad d\tilde{z}_t = C(t)\tilde{x}_t\,dt + dw_t \tag{12a}$$

with $\tilde{x}_0 \sim N(0,\bar{\Sigma}_0)$ and $z_0=0$, while $\tilde{r} \triangleq r - Hx^*$ has spatial intensity

$$\tilde{\gamma}_t(\tilde{r},\tilde{x}_t) \sim N(H(t)\tilde{x}_t, R(t)). \tag{12b}$$

Let $\tilde{\mathcal{Z}}_t$ be the σ-algebra generated by $\tilde{z}$ over $[0,t)$ and let $\tilde{\mathcal{R}}_t$ be that generated by the space-time process that is obtained from the original one by leaving N unchanged and replacing r by $r-Hx^*$. Then, *under the assumption that u_t is also $\tilde{\mathcal{Z}}_t \vee \tilde{\mathcal{R}}_t$-measurable*, an argument that parallels the proof of Lemma 1 in [7] shows that $\mathcal{Z}_t \vee \mathcal{R}_t = \tilde{\mathcal{Z}}_t \vee \tilde{\mathcal{R}}_t$. This assumption is discussed shortly. Thus, it is equivalent to prove that $f_t[\mu|\tilde{\mathcal{Z}}_t \vee \tilde{\mathcal{R}}_t] = f_t[\mu|\mathcal{T}_t]$. Now, because μ and N are independent of x_0, v, and w, so also are they independent of $\tilde{x}$ and $\tilde{z}$; thus, the joint density of μ and the event that $t_1, t_2, \cdots, t_k$ are the occurrence times of N over $[0,t)$ satisfies $f[\mu, t_1, \cdots, t_k|\tilde{\mathcal{W}}_t] = f[\mu, t_1, \cdots, t_k]$, where $\tilde{\mathcal{W}}_t = \tilde{\mathcal{X}}_t \vee \tilde{\mathcal{Z}}_t$ and $\tilde{\mathcal{X}}_t$ is the σ-algebra generated by $\tilde{x}$ over $[0,t)$. Equivalently,

$$f[\mu|\mathcal{T}_t \vee \tilde{\mathcal{W}}_t]\, f[t_1, \cdots, t_k|\tilde{\mathcal{W}}_t] = f[\mu|\mathcal{T}_t]\, f[t_1, \cdots, t_k].$$

Because N, $\tilde{x}$ and $\tilde{z}$ are independent, $f[t_1, \cdots, t_k|\tilde{\mathcal{W}}_t] = f[t_1, \cdots, t_k]$ and therefore

$$f[\mu|\mathcal{T}_t \vee \tilde{\mathcal{W}}_t] = f[\mu|\mathcal{T}_t]. \tag{12c}$$

Also, we have from Bayes' rule that

$$f[\mu|\tilde{\mathcal{W}}_t \vee \tilde{\mathcal{R}}_t]\, f[\tilde{r}_1, \cdots, \tilde{r}_{N_t}|\tilde{\mathcal{W}}_t \vee \mathcal{T}_t]$$
$$= f[\mu|\tilde{\mathcal{W}}_t \vee \mathcal{T}_t]\, f[\tilde{r}_1, \cdots, \tilde{r}_{N_t}|\sigma(\mu) \vee \tilde{\mathcal{W}}_t \vee \mathcal{T}_t],$$

since both equal $f[\mu, \tilde{r}_1, \cdots, \tilde{r}_{N_t}|\tilde{\mathcal{W}}_t \vee \mathcal{T}_t]$. Now the second factor on each side equals

$$\prod_{i=1}^{N_t} \tilde{\gamma}_{t_i}(\tilde{r}_i, \tilde{X}_{t_i}),$$

and cancelling these and combining the result with (12c) yields

$$f[\mu|\tilde{\mathcal{W}}_t \vee \tilde{\mathcal{R}}_t] = f[\mu|\mathcal{T}_t].$$

Recalling that $\tilde{\mathcal{W}}_t = \tilde{\mathcal{X}}_t \vee \tilde{\mathcal{Z}}_t$, the desired result then follows immediately.

Remark 1: The technical assumption that u_t is $\tilde{\mathcal{R}}_t \vee \tilde{\mathcal{Z}}_t$-measurable which is required for the proof that $f_t[\mu|\mathcal{B}_t] = f_t[\mu|\mathcal{T}_t]$ also arises as a sufficient condition in [7]. A generalization of [7, theorem 3] shows that this will be the case if u is generated from the past of z, N, and r or $\tilde{z}$, N, and $\tilde{r}$ using a suitably smooth control law. Specifically, it will be so if u is generated as a Lipschitz function of the state of a suitably smooth finite-dimensional system; included here, in particular, is a control u_t so generated from $\hat{x}_t$ of (6), which is of interest because this is the case for the optimum control found later.

Remark 2: The stochastic differential equations for $\hat{x}_t$ and Σ_t given in Theorem 1 admit an intuitively simple interpretation. In the intervals between point occurrences of the space-time process, the problem reduces to one of estimating x_t from the observations z; the nonoccurrence of further points in these intervals provides information about μ_t but none about x_t because of the separability (2) of λ and our standing assumptions concerning independence. Thus, during these intervals we are left with the standard Kalman filtering problem of estimating the state of the linear system (1a) from the observations (1b); if x_t is conditionally Gaussian at the beginning of each such interval, it remains so throughout with mean and covariance which evolve according to (6) and (8) or (9) with the last term deleted in each. (This, of course, is also reflected in the equation (11) for the conditional density reducing to Kushner's equation during these intervals.) We now observe that x_t is, in fact, conditionally Gaussian at the beginning of each such interval because it is at $t=0$ and because it remains Gaussian after each point occurrence: indeed, at each occurrence (t,r) of the space-time point process the spatial observation r is an independent observation on a Gaussian random variable with mean Hx_t and covariance R; this is equivalent to a discrete observation of the form

$$r = H(t)x_t + \xi$$

where $\xi \sim N(0,R)$ is independent of x and z. Thus, from standard estimation theory for Gaussian random variables [e.g., 10], the conditional density remains Gaussian and the change in conditional mean and covariance of x_t after accounting for this new observation are

$$d\hat{x}_t \triangleq \hat{x}_{t+} - \hat{x}_t = \Sigma_t H'[H\Sigma H' + R]^{-1}(r - H\hat{x}_t)$$
$$= M_t(r - H\hat{x}_t)$$
$$d\Sigma_t \triangleq \Sigma_{t+} - \Sigma_t = -\Sigma_t H'[H\Sigma H' + R]^{-1} H\Sigma_t$$
$$= -M_t H\Sigma_t$$
$$d\Sigma_t^{-1} \triangleq \Sigma_{t+}^{-1} - \Sigma_t^{-1} = H'R^{-1}H.$$

Of course, this term is to be included only when an occurrence takes place at (t,r); multiplying each of these

calculation that

$$S(t) \triangleq E\left[(x_t - x_t^{\#})(x_t - x_t^{\#})'\right] \tag{17}$$

satisfies the linear matrix differential equation

$$\dot{S} = (F - LC)S + S(F - LC)' + VV' + LL' + \bar{\mu}\{M[HSH' + R]M - MHS - SH'M'\},$$
$$S(0) = \operatorname{cov}[x_0] \tag{18}$$

where we have suppressed the common argument t of all entries, and where $\bar{\mu}(t) = E[\mu_t]$. Because all coefficients in (18) are uniformly bounded, a unique solution to (18) exists for all $t \in [0, \infty)$. We thus have proved Theorem 3.

Theorem 3: For any uniformly bounded $M(\cdot)$ and $L(\cdot)$, the mean-square performance (17) of the estimator (16) satisfies the linear matrix differential equation (18), and this is a matrix ordering upper bound on $E[\Sigma_t]$, i.e., for all $t \in [0, \infty)$,

$$\bar{\Sigma}(t) = E[\Sigma_t] \leqslant S(t) \tag{19}$$

in the sense that $S(t) - \bar{\Sigma}(t)$ is nonnegative definite.

We now show that there exists a choice of $M(\cdot)$ and $L(\cdot)$ in (16) for which the corresponding mean-square performance given by (18) lies at all times below that for any other choice.

Theorem 4: Let $S^*(\cdot)$, $M^*(\cdot)$, and $L^*(\cdot)$ satisfy

$$\dot{S}^* = FS^* + S^*F' + VV' - S^*C'CS^* - \bar{\mu}S^*H'[HS^*H' + R]^{-1}HS^*, \qquad S^*(0) = \operatorname{cov}[x_0] \tag{20}$$

$$M^* = S^*H'[HS^*H' + R]^{-1}, \quad L^* = S^*C' \tag{21}$$

where the common argument t of all entries in (20) and (21) is suppressed. Let $S(\cdot)$ be the solution to (18) for some arbitrary $M(\cdot)$ and $L(\cdot)$. Then, for all $t \in [0, \infty)$,

$$E[\Sigma_t] \leqslant S^*(t) \leqslant S(t) \tag{22}$$

and $S^*(t) = E\{(x_t - x_t^*)(x_t - x_t^*)'\}$ is the mean-square performance of the bound-minimal estimator

$$dx_t^* = Fx_t^*\,dt + Gu_t\,dt + S^*C'[dz_t - Cx_t^*\,dt] + S^*H'[HS^*H' + R]^{-1}\int_{R^m}[r - Hx_t^*]N(dt \times dr). \tag{23}$$

Proof: Completing the square on the right side of (18) yields

$$\dot{S} = FS + SF' + VV' - SC'CS - \bar{\mu}SH'[HSH' + R]^{-1}HS + (L - SC')(L - SC')' + \bar{\mu}\left[M - SH'(HSH' + R)^{-1}\right] \cdot (HSH' + R)\left[M - SH'(HSH' + R)^{-1}\right]', \quad S(0) = \operatorname{cov}[x_0]. \tag{24}$$

For given $S(t)$, the right side of (24) is clearly minimized by making the last two nonnegative definite terms 0, in which case the right side of (24) reduces to that of (20) while the minimizing choices of $M(t)$ and $L(t)$ are given by (21). It remains to show that this instantaneous ordering on the time derivative produces a permanent ordering of the solutions over $[0, T]$, i.e., that the solution to (20) lies at all times below that of (24) in the matrix ordering. This is readily accomplished by using [9, lemma 1] after appropriate modifications to reflect that initial conditions, rather than final conditions, are of interest here. This means that in [9, lemma 1] the left sides of (*) and (**) should be replaced by $+\dot{X}$ and $+\dot{Y}$, respectively, and all time orderings $0 \leqslant t \leqslant s \leqslant T$ replaced by $0 \leqslant s \leqslant t \leqslant T$. Then, letting (24) play the role of (*) and (20) the role of (**) and checking conditions 1)–4) of [9, lemma 1] we have: 1) and 2) are trivial under our standing assumptions; 3) holds because, by a subsidiary application of [9, lemma 1], the solution to (20) lies at all times above that of the Riccati equation

$$\dot{\Gamma} = F\Gamma + \Gamma F' + VV' - \Gamma[C'C + \bar{\mu}H'R^{-1}H]\Gamma, \qquad \Gamma(0) = \operatorname{cov}[x_0] \tag{25}$$

for which 3) is known to hold; finally, 4) holds for (18) and therefore (24) because if $S_1(t)$ and $S_2(t)$ are the respective solutions to (18) with initial conditions $S_1(0)$ and $S_2(0)$, $S_1(0) \geqslant S_2(0)$, then $S_1(t) \geqslant S_2(t)$ for all t, since

$$\dot{S}_1 - \dot{S}_2 = (F - LC - \bar{\mu}MH)(S_1 - S_2) + (S_1 - S_2)(F - LC - \bar{\mu}MH)' + \bar{\mu}MH(S_1 - S_2)H'M, \qquad (S_1 - S_2)(0) \geqslant 0$$

and the solution to this lies at all times above the (identically zero) solution to

$$\dot{Y} = (F - LC - \bar{\mu}MH)Y + Y(F - LC - \bar{\mu}MH)'; \qquad Y(0) = 0$$

by a further subsidiary application of [9, lemma 1].

Remark 5: The evaluation of the performance of the suboptimum estimator (16) is a second-order analysis that uses only the means and covariances of the various random variables and processes and makes no use of the Gaussianness of x_0, v, w, and r. Thus, the results of this section remain valid if v and w are replaced by normalized uncorrelated-increment processes that are uncorrelated with each other and with x_0 and r, with x_0 having any distribution with mean $\bar{x}_0$ and covariance $\bar{\Sigma}_0$ and the spatial intensity $\gamma_t(r, x_t)$ being any distribution with mean Hx and covariance R such that r is uncorrelated with x_0. The bound-minimal estimator (23) can then be viewed as the best estimator in the family (16). These estimators are bilinear because of the product $rN(dt \times dr)$ in the last term, though they might also be considered to be in a sense linear, to the extent that $N(dt \times dr)$ merely signals the arrival of a spatial observation r which, as with z, is utilized linearly in the production of $\hat{x}$.

expressions by $N(dt \times dr)$ and integrating over R^m takes care of this, and constitutes the last term in (6), (7) and (9), respectively.

Remark 3: It was observed in the proof that $p_t(x|\mathcal{B}_t \vee \mathcal{M})$ does not depend on $\mathcal{M}$, and thus x_t and μ are conditionally independent given $\mathcal{B}_t$; in particular, x_t and μ_t are conditionally independent given $\mathcal{B}_t$, and thus the joint problem of estimating x_t and μ_t given $\mathcal{B}_t$ separates into two separate problems of estimating x_t and estimating μ_t:

$$f(x_t, \mu_t|\mathcal{B}_t) = f(x_t|\mathcal{B}_t) f(\mu_t|\mathcal{B}_t).$$

Furthermore, the final part of Theorem 1 establishes that $f(\mu_t|\mathcal{B}_t) = f(\mu_t|\mathcal{T}_t)$ depends only on the Poisson counting process N and does not depend on z or the spatial locations of the points. This latter estimation problem with various models for μ is examined, for example, in [6].

Theorem 2: The unique admissible control u_t^0 that minimizes the cost (5) is given by

$$u_t^0 = -P^{-1}(t)G'(t)K(t)\hat{x}_t \triangleq -L(t)\hat{x}_t \qquad (13)$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ satisfies the finite-dimensional nonlinear stochastic differential equation (6) with Σ_t given by (7), and the $n \times n$ symmetric nonnegative definite matrix $K(t)$ satisfies the Riccati equation

$$\dot{K}(t) = -K(t)F(t) - F'(t)K(t) + K(t)G(t)P^{-1}(t) \cdot G'(t)K(t) - Q(t), \qquad K(T) = S. \quad (14)$$

The corresponding minimum value of J is

$$J[u^0] = E\{x_0'K(0)x_0\} + \int_0^T \mathrm{tr}\left[KGP^{-1}G'KE\{\Sigma_t\} + KVV'\right]dt, \quad (15)$$

where tr denotes trace.

Proof: According to Åström [8], $J[u]$ can be rewritten as

$$J[u] = \int_0^T E\left\{\|u_t + L(t)\hat{x}_t\|^2_{P(t)}\right\}dt + [\text{right side of (15)}]$$

where $\hat{x}_t = E[x_t|\mathcal{B}_t]$ and $\|y\|_p^2 = y'Py$. The first term on the right side is nonnegative, and zero if and only if $u_t = -L(t)\hat{x}_t$. Thus, (13) gives the unique optimum control provided the right side of (15) is invariant under changes in u. The only way for (15) to be u-dependent is through $E\{\Sigma_t\}$, and (7) shows that the only possibility for Σ_t to vary with u is via N_t. But N_t is a Poisson counting process with rate μ_t, and both N_t and μ_t are specified at the outset as mappings on $(\Omega, \mathcal{F}, P)$ without any reference to u. Hence Σ_t and, therefore, $E\{\Sigma_t\}$ and the right side of (15) are invariant under changes in u, and the proof is complete.

Remark 4: Theorem 2 shows that the solution to this stochastic control problem can be realized with a separated estimator-controller in which the estimator is nonlinear, mean-square optimal, and finite-dimensional and the controller is the certainty-equivalent linear control law (i.e., the optimum linear control law for the deterministic problem in which $x_0 = E[x_0]$, $v_t = E[v_t] \equiv 0$, and x_t is known exactly). This result, therefore, includes as special cases the familiar linear-quadratic-Gaussian "separation theorem" (where the space-time point-process observations are absent) (e.g., [8]) and the similar results in [1]–[5] for restricted versions of the space-time point-process observations.

We observe that Σ_t, the conditional covariance of x_t given $\mathcal{B}_t$, is not precomputable because the last term on the right side of (7) depends on the particular realization of the counting process N_t. One is, therefore, led to consider $\bar{\Sigma}(t) \triangleq E[\Sigma_t]$, both as a natural measure of estimation performance in its own right and as the particular measure of estimation performance that determines the optimum control cost (15). However, while $\bar{\Sigma}(t)$ is deterministic and in principle can be precalculated, this calculation is infinite-dimensional. One way of seeing this is to observe that an attempt to calculate $\bar{\Sigma}(t)$ by taking expectations of both sides of (7) is complicated by the last term on the right side, which requires the calculation of $E\{\Sigma H'[H\Sigma H' + R]^{-1}H\Sigma\}$. While a differential equation for this can be written down, it, in turn, requires expectations of additional nonlinear functions of Σ_t, and so on *ad infinitum* in a mushrooming requirement for additional terms that is familiar from other nonlinear filtering situations. Accordingly, we turn our attention to deriving easily-computed upper and lower bounds on $\bar{\Sigma}(t)$. These estimation bounds then directly imply upper and lower bounds on the optimum control performance (15).

IV. Suboptimum Estimators and Upper Bounds

Our approach to finding easily-computed upper bounds on $E[\Sigma_t]$ is to examine a parametrized family of suboptimum estimators whose mean-square performance can easily be calculated exactly. For each suboptimum estimator the corresponding mean-square error is then trivially a matrix-ordering upper bound on $E[\Sigma_t]$. Furthermore, we show that there exists a suboptimum estimator in this family whose mean-square performance is at all times smaller than that of any other, thus providing a minimal upper bound within this class.

Motivated by the form of the optimum estimator (6), we consider the family of suboptimum estimators

$$dx_t^\# = F(t)x_t^\# dt + G(t)u_t dt + L(t)\left[dz_t - C(t)x_t^\# dt\right] + \int_{R^m} M(t)\left[r - H(t)x_t^\#\right]N(dt \times dr) \quad (16)$$

parametrized by the deterministic uniformly-bounded $n \times m$- and $n \times q$-matrix valued time functions $L(\cdot)$ and $M(\cdot)$. This family does not include the optimum estimator (6) in which M_t, $L_t = \Sigma_t C'$ are random matrices depending on N through Σ. Apart from the requirement that $M(\cdot), L(\cdot)$ be deterministic, the suboptimum estimator (16) and the optimum estimator (6) share the same structure. The nonrandomness of M, N enables us to write down an ordinary $n \times m$-matrix differential equation for the mean-square error of the suboptimum estimator (16). Indeed, subtracting (16) from (1a), it follows directly by straightforward

V. Estimation Lower Bounds and Control Bounds

Theorem 5: Let S_* be the solution to (25). Then for all $t \geq 0$, $S_*(t)$ is a matrix-ordering lower bound on $E[\Sigma_t]$, i.e.,

$$S_*(t) \leq E[\Sigma_t]. \tag{26}$$

Proof: We have from (9) that $\overline{\Sigma^{-1}} \triangleq E[\Sigma_t^{-1}]$ satisfies

$$\begin{aligned}\dot{\overline{\Sigma^{-1}}} &= -\overline{\Sigma^{-1}}F - F'\overline{\Sigma^{-1}} - E[\Sigma^{-1}VV'\Sigma^{-1}] + C'C \\ &\quad + \bar{\mu}H'R^{-1}H; \qquad \overline{\Sigma^{-1}}(0) = (\operatorname{cov}[x_0])^{-1} \\ &= -\overline{\Sigma^{-1}}F - F'\overline{\Sigma^{-1}} - \overline{\Sigma^{-1}}VV'\overline{\Sigma^{-1}} + C'C \\ &\quad + \bar{\mu}H'R^{-1}H - \Delta \end{aligned} \tag{27}$$

where

$$\Delta = E[\Sigma^{-1}VV'\Sigma^{-1}] - \overline{\Sigma^{-1}}VV'\overline{\Sigma^{-1}} = \operatorname{cov}[\Sigma^{-1}V] \geq 0. \tag{28}$$

It then follows from [9, lemma 1] that $\overline{\Sigma^{-1}}$ lies at all times below the solution to

$$\begin{aligned}\dot{\Xi} &= -\Xi F - F'\Xi - \Xi VV'\Xi + \bar{\mu}H'R^{-1}H + C'C \\ \Xi(0) &= (\operatorname{cov}[x_0])^{-1}. \end{aligned} \tag{29}$$

Thus,

$$\Xi(t) \geq \overline{\Sigma^{-1}}(t) \triangleq E[\Sigma_t^{-1}] \geq (E\Sigma_t)^{-1}, \tag{30}$$

the last inequality being a matrix version of Jensen's inequality proved in the Appendix. Taking inverses of (30) and noting that if S_* is the solution to (25), then S_*^{-1} is the solution to (29), we have the desired result (26).

We remark in passing that (25) gives the covariance of the optimum estimator when the space time point-process observations are replaced by continuous observations of the form

$$dy_t = Hx_t\,dt + (\bar{\mu}^{-1}R)^{1/2}\,dn_t$$

where n_t is a Wiener process independent of x_0, v, and w.

Remark 6: Comparing (20) for the minimal upper bound with (25) for the lower bound, we see that these two bounds will be close to each other and thus to $E[\Sigma_t]$ if HS^*H' is small compared with R (or, equivalently, if $H'R^{-1}H$ is small compared with S^*). Both bounds will also be close to each other and to the optimum performance if the mean intensity $\bar{\mu}$ is small. These are discussed later in terms of our motivating example.

Once we have deduced upper and lower bounds on the estimation performance $E[\Sigma_t]$, corresponding bounds on the optimum control performance follow directly by substitution of these bounds for $E[\Sigma_t]$ in (15).

Theorem 6: Upper and lower bounds on the optimum control performance $J[u^0]$ of (15) are

$$\begin{aligned} E\{x_0'K(0)x_0\} + \int_0^T \operatorname{tr}[KGP^{-1}G'PKS_* + KVV']\,dt \\ \leq J[u^0] \leq E\{x_0'K(0)x_0\} \\ + \int_0^T \operatorname{tr}[KGP^{-1}G'KS^* + KVV']\,dt \end{aligned}$$

where S_* is the solution to (25) and S^* is the solution to (20).

VI. Discussion

The above estimator-controller solution extends results in [2] and [3] to include a more general form of observation. Just as with the observation model in [2] and [3], this more general observation is motivated by communication systems that employ a narrow beam of light as a carrier, by star tracking systems, and by infrared tracking systems, all of which have a requirement for position sensing and active tracking to maintain optical alignment in the presence of a variety of disturbances. We shall indicate how the models of [2, sec. 4] and [3, sec. 4] are usefully extended by this more general observation. The estimator-controller solution of Theorem 2 provides a possible tool for the design of an optical tracking system under the conditions indicated below, and the performance bounds of Sections IV and V provide the means for predicting the performance of such designs.

Let $I(t,\vec{r})$ denote the light intensity at time $t \in [0,\infty)$ and position $\vec{r} \in \mathcal{R}$ of an optical field incident on the photoemissive surface of a two-dimensional photodetector on boresight and without any motions. Here, $\mathcal{R}$ is a subregion of R^2 corresponding to the photoemissive surface. We assume a Gaussian intensity-profile

$$I(t,\vec{r}) = I_0(t)\exp\{-\tfrac{1}{2}\vec{r}'R^{-1}(t)\vec{r}\}.$$

Vibration, beam steering due to propagation of the light beam through atmospheric turbulence, and other effects cause the spot of light on the photoemissive surface to move about in a random fashion and to fluctuate randomly in optical intensity. In this case, the intensity profile becomes

$$\begin{aligned} I(t,\vec{r},y_m(t)) = I_0(t)\exp\{-\tfrac{1}{2}[\vec{r}-y_m(t)]' \\ \cdot R^{-1}(t)[\vec{r}-y_m(t)]\} \end{aligned}$$

where $y_m(t)$ models the random motions, and $I_0(t)$ is a random process (e.g., a lognormal process) that models random intensity fluctuations. We assume that $\{y_m(t); t \geq 0\}$ is derived from a Gaussian diffusion satisfying

$$\begin{aligned} dx_m(t) &= F_m(t)x_m(t)\,dt + V_m(t)\,dv_m(t), \\ y_m(t) &= H_m(t)x_m(t) \end{aligned}$$

where $\{v_m(t); t \geq 0\}$ is a standard Wiener process. The fading process $\{I_0(t); t \geq 0\}$ is assumed to be independent of motion processes but is otherwise arbitrary. The pur-

pose of the tracking controller is to compensate for these random motions and random fading in order to maintain optical alignment. Thus, in the presence of a controller to position telescopes, mirrors, or other pointing devices, the intensity becomes

$$I(t,\vec{r},y_m(t),y_p(t)) = I_0(t)\exp\left\{-\tfrac{1}{2}\left[\vec{r}-y_m(t)+y_p(t)\right]' \cdot R^{-1}(t)\left[\vec{r}-y_m(t)+y_p(t)\right]\right\}$$

where $y_p(t)-y_m(t)$ is the tracking error. Ideally, this error should be zero, but this cannot be accomplished for two reasons: the position error $y_m(t)$ is unknown and must be estimated from data available at the photodetector output, and the tracking devices will have some inertia so that $y_m(t)$ cannot be tracked instantaneously even if it were known. We model the tracking devices by a linear stochastic plant

$$dx_p(t) = F_p(t)x_p(t)\,dt + G_p(t)u(t)\,dt + V_p(t)\,dv_p(t)$$
$$y_p(t) = H_p(t)x_p(t)$$

where $u(t)$ is the input to the tracking devices from the tracking controller, and $\{v_p(t); t \geq 0\}$ is a standard Wiener process modeling local disturbances such as those due to vibration.

Photoelectron conversions take place in the photoemissive surface at a rate proportional to the incident light intensity [3]. Thus, the photoelectron conversion rate has the form of $\lambda_t(r,x_t,\mu_t)$ for $(t,\vec{r})\in[0,\infty)\times\mathcal{R}$ with μ_t an appropriately scaled version of $I_0(t)$, and x is the vector obtained by adjoining x_m and x_p, and H is obtained from H_m and H_p in an obvious way.

The problem of optical tracking is to follow the position of maximum light intensity at time t in terms of both photoelectron conversions observed on $[0,t)\times\mathcal{R}$ and observations of the plant state x_p obtained with sensors located at the tracking devices. These latter observations are modeled according to (1b) so as to account for sensor noise. Except for the finiteness of $\mathcal{R}$, this problem is identical to control problem studied above when photoelectron conversions are identified as space-time points. An approximation that appears reasonable when the beam is small and the tracking errors are small (i.e., fine tracking mode rather than an acquisition mode) compared to the size of the photoemissive surface is to replace $\mathcal{R}$ by R^2. With this approximation, the optical tracking problem is solved by the result in Theorem 2.

It is important to note that according to Remark 3 and Theorem 2, the design of the tracking controller does not depend in any way upon the source or nature of randomness in $I_0(t)$. Thus, for example, the same design is obtained if $I_0(t)$ is random due to atmospheric turbulence or modulation by an information-bearing signal or a combination of these.

The upper and lower bounds of Sections IV and V provide a measure of the performance for the optical position-sensing and tracking system derived from Theorem 2. From Remark 6, the upper and lower bounds merge when HS^*H' is small compared to the beam spread as measured by R. It is evident that the estimation and control lower bounds derived as above for observations of each photoelectron conversion are also lower bounds for both optimal and suboptimal trackers that employ observations obtained by temporal or spatial averaging as would be obtained using photon counting and a quadrant photomultiplier.

We mention also that Segall in [11] has applied the models of [1] and [3] to study computer communication networks. The upper and lower bounds on performance that we have derived can be applied in this context as well.

Appendix

Lemma: Let W_1 and W_2 be positive definite matrices, and let $\gamma\in[0,1]$. Then

$$\left[\gamma W_1+(1-\gamma)W_2\right]^{-1} \leq \gamma W_1^{-1}+(1-\gamma)W_2^{-1}, \quad \text{(A1)}$$

i.e., W^{-1} is convex in a matrix sense. Furthermore, we have

$$E[W^{-1}] \geq (E[W])^{-1} \quad \text{(A2)}$$

(cf. Jensen's inequality).

Proof: Let $x=[\gamma W_1+(1-\gamma)W_2]y$. Then

$$\begin{aligned} x'\left[\gamma W_1^{-1}+(1-\gamma)W_2^{-1}\right]x &= \gamma^3 y'W_1y + 2\gamma^2(1-\gamma)y'W_2y \\ &\quad + \gamma(1-\gamma)^2 y'W_2W_1^{-1}W_2y \\ &\quad + (1-\gamma)\gamma^2 y'W_1W_2^{-1}W_1y \\ &\quad + 2\gamma(1-\gamma)^2 y'W_1y \\ &\quad + (1-\gamma)^3 y'W_2y. \end{aligned}$$

Now $W_2W_1^{-1} \geq 2W_2 - W_1$ because $(W_2-W_1)W_1^{-1}(W_2-W_1)\geq 0$; similarly, $W_1W_2^{-1}W_1 \geq 2W_1 - W_2$. Substituting these inequalities and simplifying yields

$$\begin{aligned} x'\left[\gamma W_1^{-1}+(1-\gamma)W_2^{-1}\right]x &\geq \gamma\left[(\gamma+1-\gamma)^2\right]y'W_1y \\ &\quad + (1-\gamma)\left[(1-\gamma+\gamma)^2\right]y'W_2y \\ &= \gamma y'W_1y + (1-\gamma)y'W_2y \\ &= x'\left[\gamma W_1+(1-\gamma)W_2\right]^{-1}x. \end{aligned}$$

From this (A2) follows.

References

[1] D. L. Snyder and P. M. Fishman, "How to track a swarm of fireflies by observing their flashes," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 692–695, Nov. 1975.

[2] D. L. Snyder, I. B. Rhodes, and E. V. Hoversten, "An exact estimator-controller solution to a stochastic optimal control problem with point process observations," in *Proc. 6th Symp. Nonlinear Estimation and Its Applications*, Sept. 1975.

[3] ——, "A separation theorem for stochastic control problems with point process observations," *Automatica*, vol. 13, Jan. 1977.

[4] M. V. Vaca, "A measure transformation approach to estimation

and decision for observations derived from Martingales," D.Sc. dissertation, Sever Institute of Washington University, St. Louis, MO, May 1975.
[5] M. V. Vaca and D. L. Snyder, "Estimation and decision for observations derived from Martingales: Part 2, Applications," *IEEE Trans. Inform. Theory*, to be published.
[6] D. L. Snyder, *Random Point Processes*. New York: Wiley, 1975.
[7] I. B. Rhodes and A. S. Gilman, "Cone-bounded nonlinearities and mean-square bounds—Estimation lower bound," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 632–642, Oct. 1975.
[8] K. J. Astrom, *Introduction to Stochastic Control Theory*. New York: Academic, 1970, sec. 8.7.
[9] A. S. Gilman and I. B. Rhodes, "Cone-bounded nonlinearities and mean-square bounds—Quadratic regulation bounds," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 472–483, Aug. 1976.
[10] I. B. Rhodes, "A tutorial introduction to estimation and filtering," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 688–706, Dec. 1971.
[11] A. Segall, "Centralized and distributed control schemes for Gauss–Poisson processes," Electron. Syst. Lab., Massachusetts Inst. Technol., Cambridge, Rep. ESL-P-669, 1976.

Ian B. Rhodes (M'67) received the B.E. and M.Eng. in electrical engineering from the University of Melbourne, Melbourne, Australia, in 1963 and 1965, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1968.

In January 1968 he was appointed Assistant Professor of Electrical Engineering at the Massachusetts Institute of Technology, Cambridge, and taught there until September 1970, when he joined the faculty of Washington University, St. Louis, MO, as an Associate Professor of Engineering and Applied Science with the Department of Systems Science and Mathematics. He spent the first six months of 1974 visiting the University of Newcastle, N.S.W., Australia, and is spending the first six months of 1977 at the University of California, Berkeley. His research interests lie in mathematical system theory and its applications.

Dr. Rhodes is a member of SIAM and Sigma Xi. He is an Associate Editor of the IFAC Journal *Automatica* and a member of the Administrative Committee of the IEEE Control Systems Society. He has served in the past as an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and Chairman of the Technical Committee on Large Systems and Differential Games of the IEEE Control Systems Society.

Donald L. Snyder (S'60–M'62) was born in Stockton, CA, on January 8, 1935. He received the B.S. degree in electrical engineering from the University of Southern California, Los Angeles, and the M.S. and Ph.D. degrees, both in electrical engineering, from the Massachusetts Institute of Technology, Cambridge, in 1961, 1963, and 1966, respectively.

From 1966 to 1969 he was on the faculty of the Massachusetts Institute of Technology. He served as Associate Director of the Biomedical Computer Laboratory, Washington University School of Medicine, St. Louis, MO, and is presently Professor of Electrical Engineering and Chairman of the Department of Electrical Engineering, Washington University. He is author of *Random Point Processes* (New York: Wiley, 1975).

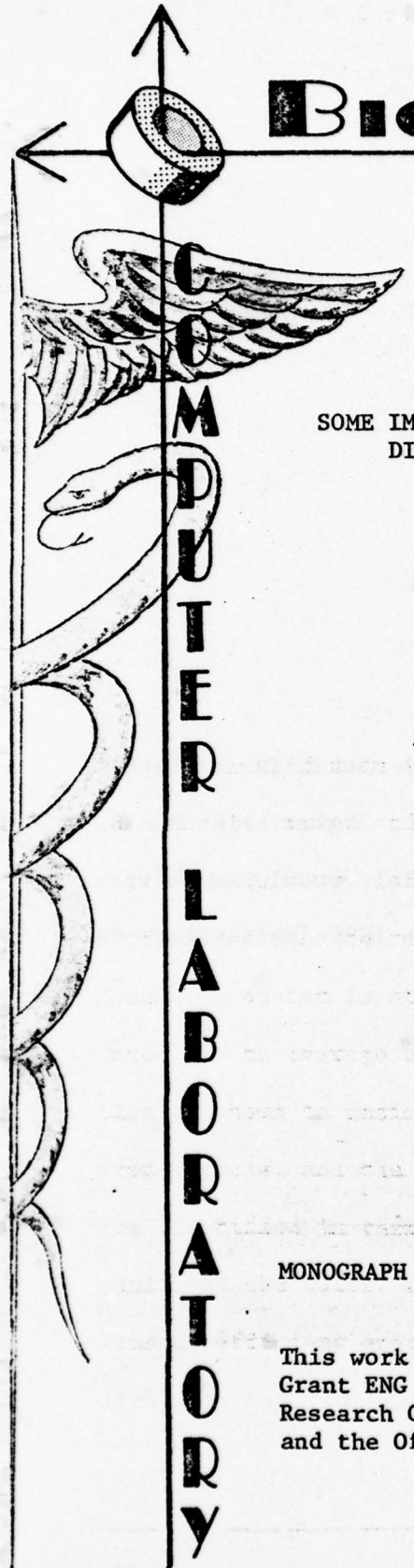
Dr. Snyder is presently Associate Editor for Stochastic Processes of the IEEE TRANSACTIONS ON INFORMATION THEORY and a member of the Administrative Committee of the IEEE Group on Information Theory.

APPENDIX 4

Report:

"Some Implications of the Cutoff-Rate Criterion for Coded, Direct-Detection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Biomedical Computer Laboratory Monograph 363, Washington University, St. Louis, MO, March 1979. Submitted to IEEE Transactions on Information Theory.

(Pages 113 - 162)



SOME IMPLICATIONS OF THE CUTOFF-RATE CRITERION FOR CODED, DIRECT-DETECTION, OPTICAL-COMMUNICATION SYSTEMS

Donald L. Snyder
Ian B. Rhodes

MONOGRAPH NO. 363 MARCH 1979

This work was supported by the National Science Foundation under Grant ENG 76-11565 and by the National Institutes of Health under Research Grant RR 00396 from the Division of Research Resources, and the Office of Naval Research under Contract N00014-76-C-0667.

WASHINGTON UNIVERSITY SCHOOL OF MEDICINE – ST. LOUIS, MO.

Some Implications of the Cutoff-Rate Criterion for Coded, Direct-Detection, Optical-Communication Systems†

Donald L. Snyder

and

Ian B. Rhodes

Washington University

St. Louis, Missouri 63130

ABSTRACT

The cutoff rate is derived for a digital communication system employing an optical carrier and direct detection. The coordinated design of the encoder, optical modulator, and demodulator is then studied using the cutoff rate as a performance measure rather than the more commonly employed error probability. Modulator design is studied when transmitted optical signals are subject simultaneously to average energy and peak value constraints. Pulse-position modulation is shown to maximize the cutoff rate when the average energy constraint predominates, and the best signals when the peak-value constraint predominates are identified in terms of Hadamard matrices. A time-sharing of these signals maximizes the cutoff rate when neither constraint dominates the other. Problems of efficient energy utilization, choice of input and output alphabet dimension, and the effect of random detector gain are addressed.

---

†This work was supported by the National Science Foundation under Grant ENG 76-11565 and by the National Institutes of Health under Research Grant RR 00396 from the Division of Research Resources, and the Office of Naval Research under Contract N00014-76-C-0667.

I. INTRODUCTION

Our concern in this paper is with digital communication systems that employ coherent light as a carrier and direct detection as the means to convert the received optical field into an electrical signal for subsequent processing. Communication systems of this type are discussed widely in the literature (see [1]-[5] and references therein) and are of increasing importance in applications. The optical portion of the overall system consists of the optical modulator, optical channel, and optical detector shown schematically in the basic information-theoretic model of the optical, digital-communication system of Fig. 1. Here, $E(t,\vec{r})$ represents the temporally and spatially dependent complex envelope of the optical field, and $N(t)$ represents the counting process associated with the output of an ideal, direct-detection device. This counting process is assumed to be an inhomogeneous Poisson process with rate function $\lambda(t) = s(t) + \lambda_0$, where λ_0 represents the contribution to the total count rate due to dark current in the detector. Also, λ_0 can account for background radiation when this is characterized by many, weak modal-components [2,3]. The assumption that $N(t)$ is a Poisson process is met to a close approximation on the free-space channel for coherent sources [3]. On our model, the signal count-rate $s(t)$ is related to $E(t,\vec{r})$ according to

$$s(t) = (\eta/h\nu) \int_A |E(t,\vec{r})|^2 d\vec{r}, \qquad (1)$$

where η is the quantum efficiency of the detector, h is Planck's constant, ν is the optical-carrier frequency, and A is the active surface of the detector; it is evident that $s(t)$ is nonnegative, which, of course, it must be as a rate function.

We shall suppose that a code letter x in Fig. 1 is drawn once each T seconds from a q-ary alphabet $X = \{X_1, X_2, \ldots, X_q\}$. We further suppose that each demodulator-output letter y is drawn from a q'-ary alphabet $Y = \{Y_1, Y_2, \ldots, Y_{q'}\}$, where in general $q' \geq q$. Initially, we investigate "infinitely soft" decisions

for which $q' = \infty$; then we study the penalty for choosing a smaller value of q'. The decoder output letters $\hat{u}$ supplied to the sink are reproductions of the encoder input letters u supplied by the source; these are presumed to be drawn from a binary alphabet $U = \{0,1\}$. The rate of the coding system in terms of the number of source digits for each channel letter will be denoted by R bits per channel use. This means that $R = R_s T$ if the source generates R_s bits per second.

The combination of the optical modulator, optical channel, optical detector, and demodulator forms a discrete channel with a q-ary input alphabet X and q'-ary output alphabet Y. By virtue of the independent-increments property of the Poisson process and the constancy of λ_0, this is a "constant, discrete, memoryless channel" in the sense that the conditional probability the channel output sequence is $b_1 b_2 \ldots b_n$, where each b_i is in Y, given that the input sequence is $a_1 a_2 \ldots a_n$, where each a_i is in X, factors into the n-fold product of the per-letter transition probabilities according to

$$\Pr(y_1 = b_1, y_2 = b_2, \ldots, y_n = b_n | x_1 = a_1, x_2 = a_2, \ldots, x_n = a_n)$$
$$= \prod_{i=1}^{n} \Pr[y_i = b_i | x_i = a_i]. \qquad (2)$$

Furthermore, the per-letter transition probabilities are the same for any T second use of the channel. Thus, if $p_{y|x}(Y|X)$ denotes the per letter transition probability, the right side of (2) is $\prod_{i=1}^{n} p_{y|x}(b_i|a_i)$. The design of the modulator and demodulator, of course, affects $p_{y|x}(Y|X)$. We shall study the design which makes $p_{y|x}(Y|X)$ most favorable for a given optical channel and detector. The coordination of this design with that of the encoder will also be studied.

A quantity that reflects the influence of $p_{y|x}(Y|X)$ on the quality of a constant discrete memoryless channel is the cutoff rate R_0 defined by

$$R_0 = -\log_2\{\min_Q \sum_{Y\in\mathcal{Y}}[\sum_{X\in\mathcal{X}}(p_{y|x}(Y|X))^{1/2}Q(X)]^2\}$$

$$= -\log_2\{\min_Q \sum_{i=1}^{q}\sum_{j=1}^{q}Q(X_i)Q(X_j)\sum_{k=1}^{q'}(p_{y|x}(Y_k|X_i)p_{y|x}(Y_k|X_j))^{1/2}\}, \tag{3}$$

where Q is a probability mass-function on $\mathcal{X}$. Wozencraft and Kennedy [6], in 1966, were first to argue in favor of the cutoff rate as a criterion for design because it is the upper limit of code rates R for which the average decoding computation per source digit is finite when sequential decoding is used. Wozencraft and Kennedy also showed that there is a block code of rate R and codeword length N such that the probability of error Pr(e) in decoding a sourceword of length K = NR is bounded according to

$$\Pr(e) \leq 2^{-N(R_0-R)} \text{ if } R < R_0. \tag{4}$$

Thus, for block codes, the single number R_0 provides a measure of both a range of rates R for which reliable communication is possible as well as the coding complexity, as reflected by N, required to guarantee a specified block-error probability. More recently, Viterbi [7] has shown for convolutional coding and maximum-likelihood sequence decoding on the constant, discrete memoryless channel that the error probability is upper bounded according to

$$\Pr(e) \leq C_R\, L\, 2^{-NR_0} \text{ if } R < R_0, \tag{5}$$

where N is the constraint length of the convolutional code, R is the code rate, L is the total number of source letters encoded, and C_R is a weakly dependent function of R and not a function of L and N. Thus, as with block codes, the single number R_0 provides a measure of both reliable rates and code complexity.

Massey [8,9] made these observations first and has used them to make an eloquent and persuasive argument for adopting R_0 as a modulator-demodulator design parameter in place of the more commonly used error probability. In what follows, we shall investigate some of the implications of attempting to maximize this parameter for modulator-demodulator design for direct-detection, optical communication systems.

II. R_0 for Infinitely Fine Quantization

In practice, the demodulator of Fig. 1 must quantize the point process observed on [0,T] in some fashion to produce one of the q' output letters in $\mathcal{Y}$. This might be accomplished, for example, by counting points in subintervals of [0,T], disregarding their times of occurrence within these subintervals, and then comparing the subinterval counts to prescribed thresholds. Regardless of what form of quantization is adopted, the finer it is, the larger will be the cutoff rate R_0 of the resulting constant discrete memoryless channel. Thus, we consider first the limiting situation of infinitely fine quantization, for which $q' = \infty$ and $R_0 \equiv R_{0,\infty}$ is not degraded by quantization. Then, we consider the effect of finite quantization.

For a Poisson process with rate $\lambda(t)$, the probability of observing n points during [0,T] in n disjoint intervals $[t_1,t_1+\Delta t_1),[t_2,t_2+\Delta t_2),\ldots,[t_n,t_n+\Delta t_n)$ is approximated to $o(\max_i \Delta t_i)$ by

$$\left(\prod_{\ell=1}^{n} \lambda(t_\ell)\right)\exp\left(-\int_0^T \lambda(t)dt\right)\ \Delta t_1 \Delta t_2 \ldots \Delta t_n$$

for $n \geq 1$ and by

$$\exp\left(-\int_0^T \lambda(t)dt\right)$$

for $n = 0$. Consequently, for infinitely fine quantization, the summation,

call it f(i,j), over k in (3) becomes

$$f(i,j) = \exp(-\tfrac{1}{2}\int_0^T (\lambda_i(t)+\lambda_j(t))\,dt)\,[1+\sum_{n=1}^{\infty}\iint\cdots\int\prod_{\ell=1}^{n}(\lambda_i(t_\ell)\lambda_j(t_\ell))^{\frac{1}{2}}dt_1 dt_2\cdots dt_n]$$

where $\lambda_i(t)$ and $\lambda_j(t)$ are the detection rates for code letters X_i and X_j, respectively, and the integration is over the region $0\leq t_1\leq t_2\leq\ldots\leq t_n\leq T$. By extending this range of integration to $0\leq t_i\leq T$ for i = 1, 2,...,n, and dividing by n! to compensate for this extension, we obtain

$$f(i,j) = \exp(-\tfrac{1}{2}\int_0^T (g_i(t)-g_j(t))^2 dt),$$

where we define $g_i(t) = \lambda_i^{\frac{1}{2}}(t)$ and $g_j(t) = \lambda_j^{\frac{1}{2}}(t)$. Thus,

$$R_{0,\infty} = -\log_2\{\min_Q \sum_{i=1}^{q}\sum_{j=1}^{q} Q(X_i)Q(X_j)\exp(-\tfrac{1}{2}\int_0^T (g_i(t)-g_j(t))^2 dt)\}. \qquad (6)$$

This expression is identical to that obtained by Massey [8, eq.(4)] if the signal $g_i(t)$ were to be observed in an additive white Gaussian noise of unit intensity when X_i is the code letter into the modulator. It is with this expression that Massey established for the first time the $R_{0,\infty}$-optimality under an average energy constraint of a simplex signal set for the additive white Gaussian noise channel. However, the additional constraint $g_i(t)\geq\lambda_0^{\frac{1}{2}}\geq 0$ obtains here, so Massey's argument does not hold for direct-detection optical-communication systems and must be modified. This is accomplished as follows.

By defining

$$S = \sum_{i=1}^{q} Q^2(X_i)$$

and by using Jensen's inequality, Massey [8] shows from (6) that

$$R_{0,\infty}\leq -\log_2\{\min_Q [S+(1-S)\exp\left(-\frac{1}{2(1-S)}\sum_{i=1}^{q}\sum_{j=1}^{q} Q(X_i)Q(X_j)\int_0^T (g_i(t)-g_j(t))^2 dt\right)]\}$$

(7)

with equality holding if and only if the quantity

$$d_{ij}^2 \triangleq \int_0^T (g_i(t) - g_j(t))^2 dt \tag{8}$$

is the same whenever $i \neq j$. It is evident from (7) that $R_{0,\infty}$ is a monotonically increasing function of d_{ij} for $i \neq j$. Thus, if d denotes the maximum of the d_{ij} for $i \neq j$, there holds

$$R_{0,\infty} \leq -\log_2\{\min_Q [S + (1-S)\exp(-\tfrac{1}{2} d^2)]\}. \tag{9}$$

Furthermore, it is easily verified that the minimizing code letter distribution in (9) is the uniform distribution $Q(X_i) = 1/q$ for $i=1,2,\ldots,q$. As $S = 1/q$ for this distribution, (9) becomes

$$R_{0,\infty} \leq \log_2 q - \log_2[1 + (q-1)\exp(-\tfrac{1}{2} d^2)], \tag{10}$$

with equality holding if and only if $d_{ij} = d$ whenever $i \neq j$.

III. Modulator Design Based On $R_{0,\infty}$

An optical modulator designed to produce a signal set $S = \{E_1(t,\vec{r}), E_2(t,\vec{r}),\ldots,E_q(t,\vec{r})\}$ such that $d_{ij} = d$ for $i \neq j$ and such that d is as large as possible produces the best overall performance for the digital optical communication system as measured by $R_{0,\infty}$. Thus, we are motivated to examine the maximization of d subject to suitable constraints on signals in S. Associated with each signal set S is a derived signal set $G = \{g_1(t), g_2(t),\ldots, g_q(t)\}$ in which $g_i(t) = \lambda_i^{\frac{1}{2}}(t)$, where

$$\lambda_i(t) = (\eta/h\nu) \int_A |E_i(t,\vec{r})|^2 \, d\vec{r} + \lambda_0. \tag{11}$$

Note that signals in G satisfy $g_i(t) \geq g_{min} = \lambda_0^{\frac{1}{2}}$.

This maximization problem is examined subject to additional constraints on the average energy and the peak amplitude of signals in the transmitted signal set S. We assume that the average energy $\bar{E}$ of signals in S, defined by

$$\bar{E} = \frac{1}{q} \sum_{i=1}^{q} \int_0^T \int_A |E_i(t,\vec{r})|^2 \, d\vec{r}dt, \tag{12}$$

must satisfy

$$\bar{E} \leq \bar{E}_{max}, \tag{13}$$

where $\bar{E}_{max}$ is a prespecified maximum allowable average energy. Then the average energy $\bar{E}_g$ for signals in the derived signal set G, defined by

$$\bar{E}_g = \frac{1}{q} \sum_{i=1}^{q} \int_0^T g_i^2(t)dt, \tag{14}$$

satisfies

$$\bar{E}_g - \bar{n} = \bar{s} \leq \bar{s}_{max}, \tag{15}$$

where $\bar{s} = \eta\bar{E}/h\nu$ and $\bar{n} = g_{min}^2 T = \lambda_0 T$ are the average number of signal counts and noise counts, respectively, per channel use, and where

$\bar{s}_{max} = \eta \bar{E}_{max}/h\nu$. We assume, further, that the amplitude $|E_i(t,\vec{r})|$ of each signal in the transmitted-signal set S cannot exceed a prespecified maximum value P_{max}; that is,

$$|E_i(t,\vec{r})| \leq P_{max} \tag{16}$$

for $i=1,2,\ldots,q$, $0 \leq t \leq T$, and for all locations $\vec{r}$ in the active surface of the detector. Then each signal in the derived signal set G satisfies

$$g_{min} \leq g_i(t) \leq g_{max}, \tag{17}$$

where $g_{min} = \lambda_0^{1/2}$ and $g_{max} = [(\eta A/h\nu)P_{max}^2 + \lambda_0]^{1/2}$.

For modulator design, we thus have the following optimization problem: select signals in G to maximize

$$d^2 = [q(q-1)]^{-1} \sum_{i=1}^{q} \sum_{j=1}^{q} \int_0^T [g_i(t)-g_j(t)]^2 dt \tag{18}$$

subject to the following constraints:

(i) *equidistance constraint:* the quantities d_{ij} in (8) should be the same whenever $i \neq j$.

(ii) *average-energy constraint:* equation (15) should be satisfied.

(iii) *peak-amplitude constraint:* equation (17) should be satisfied.

To simplify the development, we temporarily neglect the equidistance constraint in formulating and solving this optimization problem. It will be evident subsequently that among the solutions to the relaxed problem are ones satisfying the equidistance constraint, and these are then solutions to the fully constrained problem.

We find for $q' = \infty$ that the best choice of modulator design depends on the particular values of q, T, g_{min}, g_{max}, and $\bar{s}_{max}$, but whatever values these parameters may be, there are only three categories of best design. These are determined by the conditions

$$\bar{s}_{max} \in [0, \frac{1}{q}(g_{max}^2 - g_{min}^2)T] \tag{19a}$$

$$\bar{s}_{max} \in \begin{cases} (\frac{1}{q}(g_{max}^2 - g_{min}^2)T, \frac{1}{2}(g_{max}^2 - g_{min}^2)T), & q \text{ even} \\ (\frac{1}{q}(g_{max}^2 - g_{min}^2)T, \frac{q-1}{2q}(g_{max}^2 - g_{min}^2)T), & q \text{ odd} \end{cases} \tag{19b}$$

$$\bar{s}_{max} \in \begin{cases} [\frac{1}{2}(g_{max}^2 - g_{min}^2)T, \infty), & q \text{ even} \\ [\frac{q-1}{2q}(g_{max}^2 - g_{min}^2)T, \infty), & q \text{ odd.} \end{cases} \tag{19c}$$

We say that the "average energy constraint predominates" when (19a) holds, the "peak-amplitude constraint predominates" when (19c) holds, and that "neither constraint predominates" when (19b) holds.

Average Energy Constraint Predominates. We first give an upper bound on d^2 that holds regardless of which, if any, constraint predominates. Then we identify a modulator design that achieves this upper bound when the average-energy constraint predominates.

Suppressing the common argument t of all entities, we have

$$(g_i - g_j)^2 = (g_i - g_{min})^2 - 2(g_i - g_{min})(g_j - g_{min}) + (g_j - g_{min})^2$$

$$\leq (g_i - g_{min})^2 + (g_j - g_{min})^2,$$

the inequality holding because $(g_i - g_{min}) \geq 0$ for all $i \in \{1, 2, \ldots, q\}$.

Furthermore, for $i \neq j$, equality holds if and only if at most one of g_i and g_j is strictly greater than g_{min}. Summing over $i \neq j$, then over $j \in \{1,2,\ldots,q\}$, integrating over $[0,T]$, and dividing both sides by $q(q-1)$ yields

$$\frac{1}{q(q-1)} \int_0^T \sum_{i=1}^{q} \sum_{j=i}^{q} [g_i(t)-g_j(t)]^2 dt \leq \frac{2}{q} \int_0^T \sum_{i=1}^{q} [g_i(t)-g_{min}]^2 dt, \qquad (20)$$

with equality holding if and only if, for almost all t, $g_i(t) > g_{min}$ for at most one value of i in $\{1,2,\ldots,q\}$. Now for any $i \in \{1,2,\ldots,q\}$ and any $t \in [0,T]$,

$$[g_i(t)-g_{min}][g_{max}-g_i(t)] \geq 0,$$

which yields

$$g_i^2(t)-g_i(t)[g_{max} + g_{min}] \leq -g_{max}g_{min}, \qquad (21)$$

with equality if and only if $g_i(t) = g_{min}$ or $g_i(t) = g_{max}$. We then have

$$\begin{aligned}
[g_i(t)-g_{min}]^2 &= \frac{2g_{min}}{g_{max}+g_{min}} [g_i^2(t)-(g_{max}+g_{min})g_i(t)] \\
&\quad + \left[1 - \frac{2g_{min}}{g_{max}+g_{min}}\right] g_i^2(t) + g_{min}^2 \\
&\leq \frac{2g_{min}}{g_{max}+g_{min}} (-g_{max}g_{min}) + \left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right] g_i^2(t) + g_{min}^2 \\
&= \left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right] [g_i^2(t)-g_{min}^2], \qquad (22)
\end{aligned}$$

where the inequality follows by virtue of (21), and equality holds if and only if $g_i(t) = g_{max}$ or $g_i(t) = g_{min}$. Finally, substituting (22) into (20), using the average energy constraint (15), and noting the conditions for equality yields the following lemma.

Lemma 1. Given $\bar{s}_{max}$ as the maximum average signal counts per channel use, and $g_{min} \leq g_i(t) \leq g_{max}$ for $t \in [0,T]$ and $i \in 0,2,\ldots,q$, then

$$d^2 \leq 2\left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right]\bar{s}_{max}. \tag{23}$$

Furthermore, equality holds if and only if both: (a), at any time $t \in [0,T]$, all signals in G take on value g_{min} except at most one which takes on value g_{max}; and (b),

$$\frac{1}{q}\sum_{i=1}^{q}\int_0^T g_i^2(t)dt - g_{min}^2 T = \bar{s}_{max}. \tag{24}$$

Here, condition (a) for equality is simply a combination of the conditions for equality of (20) and (22), while condition (b) is the condition for equality in (15).

A signal set that is equidistant and achieves the upper bound in Lemma 1 with equality, and which therefore maximizes the cutoff rate $R_{0,\infty}$ when the average energy constraint predominates, is characterized in the following lemma.

Lemma 2. If $\bar{s}_{max}$ satisfies (19a), equality is achieved in (23) by the equidistant pulse-position modulation (PPM) signal set

$$g_i^*(t) = \begin{cases} g_{max}, & (i-1)T/q \leq t < (i-1+\varepsilon)T/q \\ g_{min}, & \text{otherwise for } 0 \leq t \leq T, \end{cases} \tag{25}$$

where

$$\varepsilon = \frac{q\,\bar{s}_{max}}{T(g_{max}^2 - g_{min}^2)} \,. \tag{26}$$

To establish Lemma 2, note that the PPM signal set is clearly equidistant and that condition (19a) is equivalent to $\varepsilon \leq 1$ so that condition (a) in Lemma 1 is satisfied. Moreover, the average number of signal counts per channel use for the PPM signal set is given by

$$\bar{s}^* = \frac{1}{q}\sum_{i=1}^{q}\int_0^T g_i^{*2}(t)dt - g_{min}^2 T = \frac{\varepsilon T}{q}(g_{max}^2 - g_{min}^2) \,.$$

Hence, from (26), $\bar{s}^* = \bar{s}_{max}$, so condition (b) of Lemma 1 is also satisfied. Consequently, by Lemma 1, d^2 for this signal set equals the upper bound in (23). Also, it is straightforward to verify by direct calculation for this PPM signal set and $\varepsilon \leq 1$ that d^2 equals the upper bound. Lemma 2 follows, and we conclude that the equidistant PPM signal set (25) maximizes $R_{0,\infty}$ when the average energy constraint predominates.

Lemma 2 can be strengthened by noting that the equidistant PPM signal set (25) is the unique signal set that achieves equality in (23) modulo shifting or splitting of pulses while keeping them nonoverlapping and keeping the total "on-time" of any g_i equal to $\varepsilon T/q$. This is because condition (a) of Lemma 1 is satisfied if and only if pulses are nonoverlapping and because $\varepsilon \leq 1$ is chosen precisely to use up all the available energy, as required by condition (b) of Lemma 1.

Peak-Amplitude Constraint Predominates. By this we mean that the energy constraint (15) is not a limiting consideration. We therefore neglect it,

as well as the equidistance constraint, and consider the problem of maximizing d^2 in (18) subject only to (17). The average energy required by the signals that solve this problem will then provide conditions for dominance of the peak-amplitude constraint, and among the solutions to the relaxed problem are ones satisfying the equidistance constraint, and these are then solutions to the fully constrained problem.

In the Appendix, we derive the following upper bounds on d^2:

$$d^2 \leq \begin{cases} \tfrac{1}{2}q(q-1)^{-1}(g_{max}-g_{min})^2T, & q \text{ even} \qquad (27a) \\ \tfrac{1}{2}(q+1)q^{-1}(g_{max}-g_{min})^2T, & q \text{ odd.} \qquad (27b) \end{cases}$$

An alternative and simpler derivation for q even is as follows. For any choice of g_s, there holds

$$(g_i-g_j)^2 = (g_i-g_s)^2-2(g_i-g_s)(g_j-g_s) + (g_j-g_s)^2,$$

so that summing over i and j yields

$$\sum_{i=1}^{q}\sum_{j=1}^{q}(g_i-g_j)^2 = 2q\sum_{i=1}^{q}(g_i-g_s)^2 - 2q^2(c-g_s)^2,$$

where c is the centroid $c = q^{-1}\sum_{i=1}^{q} g_i$. Thus, from (18)

$$d^2 = 2(q-1)^{-1}\int_0^T \sum_{i=1}^{q}[g_i(t)-g_s(t)]^2dt - 2q(q-1)^{-1}\int_0^T [c(t)-g_s(t)]^2dt$$

$$\leq 2(q-1)^{-1}\int_0^T \sum_{i=1}^{q}[g_i(t)-g_s(t)]^2dt, \qquad (28)$$

with equality holding if and only if $c(t) = g_s(t)$ for almost all $t \in [0,T]$. Taking $g_s(t) = \frac{1}{2}(g_{max}+g_{min})$ implies $|g_i(t)-g_s| \leq \frac{1}{2}(g_{max}-g_{min})$, and the bound in (27a) then follows from (28). This bound holds for both odd and even values of q, but it is tight only for q even, and the more precise bound (27b) derived in the Appendix for q odd is the one that is achieved with equality.

Any set of q equidistant signals G satisfying (17) and achieving the upper bound (27a) for q even or (27b) for q odd is a signal set maximizing $R_{0,\infty}$. Signal sets having these properties can be identified for certain values of q by the following procedure. Partition [0,T] into m equal subintervals, and define m functions $\rho_i(t)$, i=1,2,...,m, that are piecewise constant having a constant value of 1 or 0 over each subinterval. Then, $\rho_i(t)$ can be identified by a binary codeword of length m bits. If we write $g_i^*(t) = g_{min}+\rho_i(t)[g_{max}-g_{min}]$, it is enough to find q binary codewords of length m whose common Hamming distance satisfies the conditions in Table 1. The last column in this table reflects a necessary condition for optimality that follows immediately from conditions for equality in (A1) that yields the upper bound (27); namely, for all $t \in [0,T]$,

(i) for q even, q/2 of the signals take on value g_{max} and the remaining q/2 value g_{min},

(ii) for q odd, (q-1)/2 or (q+2)/2 of the signals take on value g_{max} and the remainder g_{min}.

This provides an additional check on the otimality of the following signal set and was an important aspect in our identification of it. For optimality, however, it is sufficient that the signal set be equidistant and achieve the appropriate upper bound (i.e., Hamming distance).

For q a multiple of 4 and such that a Hadamard matrix of order q exists, q codewords satisfying these conditions are easily obtained by deleting the first column (all ones) of the normalized Hadamard matrix [10,11]. From this, s = q-1 codewords satisfying row 3 of Table 1 with s replacing q can

be obtained by deleting the codeword of all ones. Also, $p = \frac{1}{2}q$ codewords satisfying row 2 of Table 1 with p replacing q can be obtained by deleting all rows of the normalized Hadamard matrix that have a 0 in (say) the second column and the deleting the first two columns. From this, $s = \frac{1}{2}q-1$ codewords satisfying row 4 of Table 1 with s replacing q can be obtained by deleting the codeword of all ones. Since Hadamard matrices for q=1,2, or a multiple of 4 are known up to q=200 except for q=188, this procedure gives an optimizing signal set G for all $q \leq 200$ except for q = 93, 94, 187, and 188. Also infinite families of Hadamard matrices are known, for example those for which $q=2^k$ for some positive integer k: these coincide with cyclic maximal-length shift register codes, and they are also a subset of the first-order Reed-Muller codewords of this length.

We remark that complementation of an optimum signal set yields another optimum signal set. Also, time sharing of any two optimum signal sets yields another optimum signal set.

The average energy of these signal sets is easily calculated as follows. For q even, because, at any time, q/2 signals have value g_{max} and the remainder g_{min},

$$\overline{E}_g = \tfrac{1}{2}(g_{max}^2 + g_{min}^2)T, \tag{29}$$

which implies

$$\overline{s} = \overline{E}_g - g_{min}^2 T = \tfrac{1}{2}(g_{max}^2 - g_{min}^2)T. \tag{30}$$

Also, for q odd, at any time, (q+1)/2 signals have value g_{min} and the remainder g_{max}, so

$$\overline{E}_g = \frac{1}{q}[\tfrac{1}{2}(q-1)g_{max}^2 + \tfrac{1}{2}(q+1)g_{min}^2]T, \tag{31}$$

which implies

$$\overline{s} = \overline{E}_g - g_{min}^2 T = \frac{q-1}{2q}(g_{max}^2 - g_{min}^2)T. \tag{32}$$

This uses less energy than taking $(q + 1)/2$ signals with value g_{max} and, thus, extends the range of average energies for which this choice is optimum; namely, the available average energy must exceed that required for $\bar{s}$ of (30) or (32), which yields condition (19c).

Finally, the distance d* achieved by these signals that maximize $R_{0,\infty}$ when the peak-amplitude constraint predominates is given by

$$d^{*2} = \begin{cases} \frac{q}{q-1}\left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right]\bar{s}, & q \text{ even} \\ \frac{q+1}{q-1}\left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right]\bar{s}, & q \text{ odd} \end{cases} \tag{33}$$

Neither Constraint Predominates. If $\bar{s}_{max}$ satisfies (19a) or (19c), the PPM signal set or, respectively, the Hadamard-derived signal set maximizes $R_{0,\infty}$. Unless q=2 or q=3, we are left with a range of values of $\bar{s}_{max}$ for which a solution has yet to be identified. This "gap" region is specified in (19b). For q=2 or q=3, this region collapses to the empty set, and at the common upper limit of the range (19a) and lower limit of range (19c), the PPM or Hadamard-derived signal sets are equivalent and optimum. For $q \geq 4$, we now demonstrate that an optimum signal set results by time sharing the PPM and Hadamard-derived solutions.

The gap region has strictly positive length if $q \geq 4$, and then any point in either interval (19b) can be expressed as a strictly convex combination of the endpoints; that is, for q even and $\bar{s}_{max}$ in the appropriate interval (19b), there exists a unique $\lambda \in (0,1)$ such that

$$\bar{s}_{max} = \left[\frac{\lambda}{q} + \frac{(1-\lambda)}{2}\right](g_{max}^2 - g_{min}^2)T, \tag{34a}$$

while for q odd and $\bar{s}_{max}$ in the appropriate interval (19b), there exists a unique $\lambda \in (0,1)$ such that

$$\bar{s}_{max} = [\frac{\lambda}{q} + \frac{(1-\lambda)(q-1)}{2q}](g_{max}^2 - g_{min}^2)T. \tag{34b}$$

An optimum choice of modulation can now be given in terms of λ.

Lemma 3. For q even (respectively, odd) and $\bar{s}_{max}$ in the appropriate interval specified in (19b), let λ be defined by (34a) (respectively, (34b)). Then an equidistant signal set that maximizes $R_{0,\infty}$ while satisfying the average energy and peak-amplitude constraints with equality is: for fraction λ of the interval [0,T], use the "full-width" PPM signal set (25) with $\varepsilon = 1$ and T replaced by λT, and for fraction $(1-\lambda)$ of [0,T], use the signal set defined by the appropriate Hadamard matrix, as discussed in the previous section with T replaced by $(1-\lambda)T$.

Lemma 3 is proven as follows. For an arbitrary choice of $\alpha \in [0,1]$ and an arbitrary choice of maximum average energy $\bar{s}_{1,max} \in [0, \bar{s}_{max}]$ allocated to the interval $[0, \alpha T]$, we have from Lemma 1

$$\frac{1}{q(q-1)} \int_0^{\alpha T} \sum_{i=1}^{q} \sum_{j=1}^{q} [g_i(t)-g_j(t)]^2 dt \leq 2 \left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right] \bar{s}_{1,max}, \tag{35}$$

and from (27)

$$\frac{1}{q(q-1)} \int_{\alpha T}^{T} \sum_{i=1}^{q} \sum_{j=1}^{q} [g_i(t)-g_j(t)]^2 dt \leq \begin{cases} \frac{(1-\alpha)Tq}{2(q-1)} (g_{max}-g_{min})^2, & q \text{ even} \quad (36a) \\ \frac{(1-\alpha)T(q+1)}{2q} (g_{max}-g_{min})^2, & q \text{ odd} \quad (36b) \end{cases}$$

Adding these expressions and using (18), we obtain

$$d^2 \leq \begin{cases} 2\left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right]\bar{s}_{1,max} + \frac{(1-\alpha)Tq}{2(q-1)}(g_{max}-g_{min})^2, & q \text{ even} \quad (37a) \\ 2\left[\frac{g_{max}-g_{min}}{g_{max}+g_{min}}\right]\bar{s}_{1,max} + \frac{(1-\alpha)T(q+1)}{2q}(g_{max}-g_{min})^2, & q \text{ odd} \quad (37b) \end{cases}$$

Now, from Lemma 1, equality holds in (35) if and only if both:

(i) at any time $t \in [0,\alpha T]$, all signals take on value g_{min}, except at most one, which takes on value g_{max}. This implies that the average energy $\bar{s}_1$ used on $[0,\alpha T]$ is

$$\bar{s}_1 = \frac{1}{q}\sum_{i=1}^{q}\int_0^{\alpha T} g_i^2(t)dt - g_{min}^2\alpha T \leq \frac{\alpha T}{q}(g_{max}^2-g_{min}^2)$$

and

(ii) $\bar{s}_1 = \bar{s}_{1,max}$.

Thus, a necessary condition for equality in (35) is

$$\bar{s}_{1,max} \leq \frac{\alpha T}{q}(g_{max}^2-g_{min}^2). \quad (38)$$

Furthermore, the derivation in the Appendix shows that equality holds in (36a) only if half of the signals take on value g_{min} and the remainder g_{max}. This involves an average energy usage $\bar{s}_2$ on $[\alpha T,T]$ of

$$\bar{s}_2 = \frac{1}{q}\sum_{i=1}^{q}\int_{\alpha T}^{T} g_i^2(t)dt - g_{min}^2(1-\alpha)T = \frac{(1-\alpha)T}{2}(g_{max}^2-g_{min}^2). \quad (39)$$

Because of the total average energy constraint, $\bar{s}_1 + \bar{s}_2 \leq \bar{s}_{max}$, we then have using (34a),

$$\bar{s}_1 \leq \bar{s}_{max} - \bar{s}_2 = \left(\frac{\lambda}{q} + \frac{\alpha-\lambda}{2}\right)(g_{max}^2 - g_{min}^2)T. \qquad (40)$$

For equality to hold in (37a), it is necessary that both (35) and (36a) hold with equality, and necessary conditions for these are in turn (38) and, combining $\bar{s}_1 = \bar{s}_{1,max}$ with (40),

$$\bar{s}_{1,max} \leq \left(\frac{\lambda}{q} + \frac{\alpha-\lambda}{2}\right)(g_{max}^2 - g_{min}^2)T, \qquad q \text{ even.} \qquad (41a)$$

For q odd, the corresponding necessary conditions for equality in (37b) become (38) and

$$\bar{s}_{1,max} \leq \left[\frac{\lambda}{q} + \frac{(\alpha-\lambda)(q-1)}{2q}\right] (g_{max}^2 - g_{min}^2)T, \qquad q \text{ odd.} \qquad (41b)$$

We now consider the selection of $\bar{s}_{1,max}$ and α to maximize the upper bound (37a) subject to the constraints (38) and (41a), which are necessary conditions for it to hold with equality. Because both (38) and (41a) are constraints on $\bar{s}_{1,max}$, we consider each in turn to be dominant in the sense of being more restrictive. The bound (38) is less than or equal to the bound (41a) if and only if $\alpha \geq \lambda$. Substituting (38) into (37a) and simplifying, we obtain

$$d^2 \leq \frac{(g_{max}-g_{min})^2 T}{2q(q-1)} \left[q^2 - \alpha(q-2)^2\right].$$

Because we are considering $q \geq 4$, this bound is maximized over $\alpha \in [\lambda, 1]$ by the unique choice $\alpha = \lambda$. Similarly, the bound (41a) is less than or equal to the bound (38) if and only if $\alpha \leq \lambda$. Substituting (41a) into (37a) and simplifying, we obtain

$$d^2 \leq (g_{max}-g_{min})^2 T \left[\frac{\alpha(q-2)}{2(q-1)} + \frac{q}{2(q-1)} + \frac{\lambda(2-q)}{q}\right].$$

Again because, $q \geq 4$, this bound is maximized over $\alpha \in [0,\lambda]$ by the unique choice $\alpha=\lambda$. Thus, the bound (37a) is maximized, subject to the necessary conditions (38) and (41a) for it to be achieved, by taking $\alpha = \lambda$, and the corresponding maximum bound is

$$d^2 \leq \left[\frac{2\lambda}{q} + \frac{q(1-\lambda)}{2(q-1)}\right] (g_{max}-g_{min})^2 T, \qquad q \text{ even.} \tag{42}$$

But this upper bound is readily achieved by the time-sharing of a "full-width" PPM signal set for fraction λ of $[0,T]$ and the Hadamard derived signal set for the remaining fraction $1-\lambda$ of $[0,T]$. Furthermore, the average energy required by this solution is exactly $\bar{s}_{max}$. For q odd, a similar analysis leads again to the unique choice $\alpha= \lambda$ and the corresponding maximum bound

$$d^2 \leq \left[\frac{2\lambda}{q} + \frac{(1-\lambda)(q+1)}{2q}\right](g_{max}-g_{min})^2 T, \qquad q \text{ odd.} \tag{43}$$

Again, this bound is achieved by the time sharing of a full-width PPM signal set for fraction λ of $[0,T]$ and the appropriate Hadamard-matrix derived signal set for the remaining fraction $1-\lambda$ of $[0,T]$, and the average energy required by this solution is exactly $\bar{s}_{max}$, as before.

IV. Efficient Energy Utilization

Denote by λ_s the count rate due to the signal alone when it is "on" for any of the optimal signal sets derived in the previous section. Then, $g_{max}^2 = \lambda_s + \lambda_0$ and $g_{min}^2 = \lambda_0$, where λ_0 is the count rate due to the noise alone. In considering designs for efficient energy utilization, we distinguish three situations depending on which of λ_s, $\bar{s}_{max}$, and ε are adjustable and which are fixed. We seek to identify values of the adjustable parameters so that

the cutoff rate per unit energy, $R_{0,\infty}/\bar{s}$, is greatest.

1. $\bar{s}_{max}$ adjustable, λ_s fixed. The value of d^2 achieved with the optimal signal sets of the previous section is shown in Fig. 2 as a function of $\bar{s}_{max}$ assuming that λ_s is a fixed constant. Here, d^2 is a piecewise linear function of $\bar{s}_{max}$ with the following parameters:

$$d_1^2 = \frac{2}{q}[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2\lambda_0 T \tag{44}$$

$$d_2^2 = \begin{cases} \frac{q}{2(q-1)}[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2\lambda_0 T, & q \text{ even} \\ \frac{q+1}{2(q-1)}[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2\lambda_0 T, & q \text{ odd} \end{cases} \tag{45}$$

$$s_1 = 2(\lambda_0/\lambda_s)[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2 \tag{46}$$

$$s_2 = \begin{cases} \frac{q-2}{q-1}(\lambda_0/\lambda_s)[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2, & q \text{ even} \\ \frac{q^2-3q+4}{q^2-4q+3}(\lambda_0/\lambda_s)[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2, & q \text{ odd.} \end{cases} \tag{47}$$

Using (10), with equality for optimal signal sets, and using the expressions for d^2 implied by Fig. 2 and (44)-(46), we conclude that

$$dR_{0,\infty}/d\bar{s}_{max} = 0.72(q-1)[(q-1)+\exp(\tfrac{1}{2}d^2)]^{-1}(\text{slope}), \tag{48}$$

where the factor "slope" is s_1, s_2, or zero depending on which constraint, if any, predominates. Thus, $dR_{0,\infty}/d\bar{s}_{max}$ decreases monotonically with increasing $\bar{s}_{max}$, so the signal energy is used most efficiently when $\bar{s}_{max}$ is small, where the energy constraint predominates and where the PPM signal set is optimal

and utilizes energy $\bar{s} = \bar{s}_{max}$. This situation is analogous to that studied by Massey [8,9] for the additive Gaussian-noise channel. For $\bar{s} = \bar{s}_{max}$ small, we conclude that

$$R_{0,\infty}/\bar{s} \simeq 1.44 \frac{q-1}{q}(\lambda_0/\lambda_s)[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2, \tag{49}$$

with equality achieved for $\bar{s} = 0$. Hence,

$$R_{0,\infty} \leq 1.44 \frac{q-1}{q} (\lambda_0/\lambda_s)[1-(1+\lambda_s/\lambda_0)^{-1/2}]^2\bar{s} \tag{50}$$

is an upper bound on $R_{0,\infty}$ for any choice of $\bar{s}$ and any choice of modulation with near equality holding when $\bar{s}$ is small and for the PPM signal set. Since for the PPM signal set, $\bar{s} = \varepsilon T\lambda_s/q$, this means that when λ_s is fixed, the most efficient energy utilization occurs for narrow pulses, ε being selected as small as practically feasible.

2. λ_s adjustable, $\bar{s}$ fixed. By a somewhat messy but straightforward calculation, it is readily verified that $dR_{0,\infty}/d\lambda_s \geq 0$ for $\bar{s}$ fixed. Thus, $R_{0,\infty}$, and hence $R_{0,\infty}/\bar{s}$ for $\bar{s}$ fixed, is a nondecreasing function of λ_s. Consequently, the most efficient energy utilization is achieved by selecting λ_s large and, therefore, operating in the region where the energy constraint predominates. This implies using the PPM signal set with as large a value of signal count-rate λ_s as practical and sufficiently narrow pulses that $\bar{s} = \varepsilon T\lambda_s/q$.

3. $\bar{s}$ adjustable, PPM signal set with ε fixed. The PPM signal set with a fixed pulse width $\varepsilon T/q$ maximizes $R_{0,\infty}$ provided the energy constraint predominates, which we assume. For ε fixed and $\bar{s} = \varepsilon T\lambda_s/q$, we find that

$$R_{0,\infty}(\bar{s},\bar{n}_{eff}) = \log_2 q - \log_2\{1+(q-1)\exp[-\frac{1}{q}\bar{n}_{eff}(1-\sqrt{1+q\alpha_{eff}})^2]\}\ , \tag{51}$$

where we define

$$\bar{n}_{eff} = \varepsilon\bar{n} = \varepsilon\lambda_0 T$$

as the "effective" average number of noise counts per channel use, and where

$$\alpha_{eff} = \bar{s}/\bar{n}_{eff} \tag{52}$$

is the signal-to-noise energy ratio. Graphs of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ as a function of α_{eff} for several values of $\bar{n}_{eff}$ are given in Fig. 3. These graphs are seen to increase monotonically with α_{eff} for each fixed value of $\bar{n}_{eff}$. Thus, as expected, the performance improves systematically for fixed $\bar{n}_{eff}$ as the average signal energy per channel use, $\bar{s}$, increases. However, while starting from $\bar{s} = 0$, the performance initially improves rapidly, there is a point of diminishing returns after which there is only marginal improvement for further increases in $\bar{s}$. For each $\bar{n}_{eff}$, there is an $\bar{s} = \bar{s}^*(\bar{n}_{eff})$ such that for all $\bar{s} \geq 0$ there holds

$$\frac{R_{0,\infty}(\bar{s},\bar{n}_{eff})}{\bar{s}} \leq \frac{R_{0,\infty}(\bar{s}^*,\bar{n}_{eff})}{\bar{s}^*} . \tag{53}$$

This value of $\bar{s}$ can be found graphically for each $\bar{n}_{eff}$ by pivotting a vertical line about the origin ($R_{0,\infty} = 0$, $\alpha_{eff} = 0$) in Fig. 3 until it lies tangent to the graph of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$. The abscissa of the point of tangency is $\bar{s}^*/\bar{n}_{eff}$. Inequality (53) holds because the graph of $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ lies on or below the line so constructed for all $\alpha_{eff} \geq 0$. It follows from (53) that the most efficient utilization of energy, in the sense that the cutoff rate per unit energy is greatest, is achieved when $\bar{s} = \bar{s}^*$. The dashed line in Fig. 3 is a fit

of $\bar{s}^*/\bar{n}_{eff}$ versus $\bar{n}_{eff}$ obtained graphically by connecting together the points of tangency described above. From this fit, we find for the range of average noise counts in the figure that $\bar{s}^*$ and $\bar{n}_{eff}$ are approximately related by the following power-law:

$$\bar{s}^* = 2.349\, \bar{n}_{eff}^{-0.452}. \tag{54}$$

This is shown as the solid line in Fig. 4. A measure of the range of energies nearly as efficient as $\bar{s}^*$ can be determined for each $\bar{n}_{eff}$ from the values of $\alpha_{eff} = \bar{s}/\bar{n}_{eff}$ in Fig. 3 for which $R_{0,\infty}(\bar{s},\bar{n}_{eff})$ is close to, say within 10% of, the ordinate of the line of tangency constructed as above. Values of $\bar{s}$ within the dashed lines in Fig. 4 satisfy this 10% condition; Fig. 4 implies that for maximally efficient energy utilization $\bar{s}$ should be kept within about $\pm$ 2db of $\bar{s}^*$.

V. Effect of Finite Output Quantization

The cutoff rate decreases from $R_{0,\infty}$ as the dimension, q', of the output alphabet decreases. This degradation is greatest for a binary input alphabet ($q=2$) when $a' = 2$, which corresponds to making bit by bit decisions without any coding. For a Gaussian model, Massey [8,9] concludes that choosing $q' = 2$ results in a quantization loss of 2.04db; that is, in the efficient range of energy utilization for the Gaussian model, the energy per channel use must be about 2db greater for $q' = 2$ in order to achieve the same cutoff rate as when $q' = \infty$. Moreover, Massey also concludes that for $q' = 8$, there is virtually no quantization loss. The degradation for the Poisson model is somewhat smaller than that found by Massey when $5 \leq \bar{n}_{eff} \leq 40$ and is of about the same order when $\bar{n}_{eff} = 1$.

Suppose the input and output alphabets are $X = \{0,1\}$ and $Y = \{0,1\}$, so that $q=q' = 2$. We adopt a binary pulse-position modulation format with

pulses of duration $\varepsilon T/2$ because this maximizes $R_{0,\infty}$ when the energy constraint predominates. For this choice, each symbol interval is divided into two equal subintervals, and output letters are generated according to: "produce 1 if $n[0,\varepsilon T/2] < n[T/2,(1+\varepsilon)T/2]$, otherwise produce 0," where $n[0,\varepsilon T/2]$ and $n[T/2,(1+\varepsilon)T/2]$ are the number of points observed in the first and second signalling interval, respectively. Here, $n[0,\varepsilon T/2]$ and $n[T/2,(1+\varepsilon)T/2]$ are independent Poisson random variables with mean parameters $\bar{s} + (\bar{n}_{eff}/2)$ and $\bar{n}_{eff}/2$, respectively, when 0 is the input letter and $\bar{n}_{eff}/2$ and $\bar{s} + (\bar{n}_{eff}/2)$, respectively, when 1 is the input letter. As in the previous discussion, $\bar{s}$ and $\bar{n}_{eff}$ are the average number of signal counts received per channel use and effective number of noise counts received per channel use. It is straightforward to conclude for these assumptions that the cutoff rate is given by

$$R_{0,q'=2} = 1 - \log_2\{1+2[p(1-p)]^{\frac{1}{2}}\} , \qquad (55)$$

where p is the binary error probability associated with producing an output symbol 1 (or a 0) when the input symbol is a 0 (or a 1, respectively). This error probability is given graphically for certain values of $\bar{n}_{eff}$ and a range of $\bar{s}$ by Pratt [4, p. 209: identify $\bar{n}_{eff} = 2\mu_{H,B}$ and $\bar{s}=\mu_{S,B}$].

The values tabulated in Table 2 were obtained as follows: (1), $\bar{s}^*$ is obtained from (54) for each $\bar{n}_{eff}$; (3), p^* is the value of p in (55) such that $0 \leq p^* \leq 1$ and $R_{0,q'=2} = R^*_{0,\infty}$; and (4), $\bar{s}$ is obtained by interpolation from the graph given by Pratt. Thus, $\bar{s}^*$ and $\bar{s}$ of the table yield the same cutoff rate for $q' = \infty$ and $q' = 2$, respectively. To within the accuracy that the interpolation step can be accomplished, we conclude that about 1.5 db more signal energy is required with hard decisions than with infinitely soft decisions for $\bar{n}_{eff}$ in the range of 5 to 40 counts per channel use.

VI. Effect of Input Alphabet Dimension

For an input alphabet of dimension q, $q' = \infty$, and an average energy constraint that predominates, q-ary pulse-position modulation maximizes the cutoff rate. We now consider the effect of q in each of the three situations identified in Section IV.

1. $\bar{s}_{max}$ adjustable, λ_s fixed. From (48) and (49), increasing q from 2 to ∞ implies that the greatest rate per unit energy that can be achieved increases by a factor of 2. Moreover, examination of graphs of $R_{0,\infty}/\bar{s}$ for $R_{0,\infty}$ given by (10) with equality and with $d^2 = s_1\bar{s}$, where s_1 is given in (46), shows that the range of values of $\bar{s}$ for which the approximation (49) holds closely increases as q increases; in other words, the range of efficient signal energies is extended as q is increased.

2. λ_s adjustable, $\bar{s}$ fixed. For λ_s large and the PPM signal set, we see from Fig. 2 and (46) that $d^2 \simeq 2\bar{s}$. Then,

$$R_{0,\infty} \simeq \log_2 q - \log_2[1+(q-1)e^{-\bar{s}}] \leq \frac{q-1}{q}\bar{s}. \tag{56}$$

Hence, for large λ_s, $R_{0,\infty}/\bar{s} \lesssim (q-1)/q$ and, therefore, the largest rate per unit energy increases by no more than a factor of 2 as q increases from 2 to ∞.

3. $\bar{s}$ adjustable, PPM signal set with ε fixed. A graph of (51) as a function α_{eff} for $\bar{n}_{eff} = 16$ and various values of q is shown in Fig. 5. For each q, there is a corresponding signal energy that is most efficient; this can be found graphically in the same manner as before, as indicated by the lines of tangency. These efficient energies depend upon q; very roughly from the graphs, we find that $(\bar{s}^*/16)q \simeq 1$, so that the most efficient signal energy decreases as q increases. This implies a significant potential improvement

in performance at low signal energies by the use of a large input-alphabet dimension q and q-ary pulse-position modulation. These observations appear to hold for other values of $\bar{n}_{eff}$ as well.

VII. Effect of Random Detector Gain

Let $\{M(t);\ t\geq 0\}$ be a compound Poisson counting process defined by

$$M(t) = \sum_{n=0}^{N(t)} u_n, \tag{57}$$

where $\{N(t);\ t\geq 0\}$ is the Poisson counting process defined above, $u_0=0$, and $\{u_n;\ n=1,2,\ldots\}$ is a sequence of independent, identically distributed random variables each having an integer value greater than zero. Here, $\{N(t);\ t\geq 0\}$ models primary photoelectron conversions, and u_n models the number of secondary electrons appearing at the detector output due to the n*th* conversion. This random gain is an important effect encountered, for example, with avalanche detectors used in optical-fiber communication systems.

In considering a digital-data communication system in which measurements are derived from $\{M(t);\ t\geq 0\}$, it is of interest to know the cutoff rate $R_{0,\infty}$ for infinitely fine quantization. As before, this quantity then places an upper limit on the performance for any receiver employing finite quantization, such as an "integrate and dump" receiver [12,13] in which M(nT) - M[(n-1)T] is used to make a decision about the n*th* transmitted symbol.

We find $R_{0,\infty}$ to be identical to that in (6), so random detector-gain neither degrades nor inhances the cutoff rate for infinitely fine output quantization. This is because the distribution of the random gains is unaffected by the choice of transmitted signal on our model and can be verified mathematically by the following steps. First, we write the summation over k

in (3) as $f(i,j) = E_j[\Lambda_{i,j}^{1/2}(y)]$, where

$$\Lambda_{i,j}(Y) = p_{y|x}(Y|X_i)/p_{y|x}(Y|X_j) \tag{58}$$

is the likelihood ratio for symbol X_i relative to symbol X_j and $E_j(\cdot)$ denotes a conditional expectation given X_j. As the output quantization is refined, this becomes

$$f(i,j) = E_j[\Lambda_{i,j}^{1/2}(M(t);\ 0\leq t\leq T)], \tag{59}$$

where $\Lambda_{i,j}(M(t);\ 0\leq t\leq T)$ is given by the ratio of the sample function densities [18] of $\{M(t);\ 0\leq t\leq T\}$ for symbols X_i and X_j. This likelihood ratio is found not to be a function of the random gains, and the assertion follows.

A consequence of this assertion is that many of the conclusions reached in preceeding sections also apply in the presence of random detector-gain. At the present time, there are too few published results on the binary error-probability for an integrate-and-dump receiver for us to examine the potential benefits of employing finer output quantization, but this is a matter of some practical interest for fiber-optic systems.

VIII. Polarization Modulation

Suppose that binary, orthogonal polarization modulation can also be employed in the optical modulator of Fig. 1 in addition to temporal modulation. Then the scalar field $E(t,r)$ becomes a vector $(E_1(t,\vec{r}),\ E_2(t,\vec{r}))$ in which one component is the 0^o field and the other one the 90^o field. A polarization decomposition of the received field followed by direct detection in each channel then results in two independent point processes, which we label $N_1(t)$ and $N_2(t)$, $0\leq t\leq T$. Assume that when the input codeletter is $X_i \in \{X_1, X_2, \ldots, X_q\}$ that the count rate for $N_1(t)$ is

$$\lambda_{1i}(t) = s_{1i}(t) + \lambda_0 = g_{1i}^2(t) \tag{60a}$$

and for $N_2(t)$ is

$$\lambda_{2i}(t) = s_{2i}(t) + \lambda_0 = g_{2i}^2(t). \tag{60b}$$

Following the procedure used in the last section, as $q' \to \infty$, the sum over k in (3), call it f(i,j), becomes

$$f(i,j) = E_j \{\Lambda_{i,j}^{\frac{1}{2}}[N_1(t), N_2(t); 0 \le t \le T]$$

$$= \exp(-\tfrac{1}{2}d_{ij}^2), \tag{61}$$

where

$$d_{ij}^2 = \int_0^T ([g_{1i}(t) - g_{1j}(t)]^2 + [g_{2i}(t) - g_{2j}(t)]^2)dt. \tag{62}$$

The steps leading to (10) remain unchanged with (62) replacing (8).

We now assume that each of the signals in $S_1 = \{E_{11}(t,\vec{r}), E_{12}(t,\vec{r}), \ldots, E_{1q}(t,\vec{r})\}$ and $S_2 = \{E_{21}(t,\vec{r}), E_{22}(t,\vec{r}), \ldots, E_{2q}(t,\vec{r})\}$ satsify the average energy and peak-amplitude constraints in the section about modulator design based on $R_{0,\infty}$. Then, we have the following optimization problem: select signals in $G_1 = \{g_{11}(t), g_{12}(t), \ldots, g_{1q}(t)\}$ and $G_2 = \{g_{21}(t), g_{22}(t), \ldots, q_{2q}(t)\}$ to maximize

$$d^2 = [q(q-1)]^{-1} \sum_{i=1}^{q} \sum_{j=1}^{q} \int_0^T ([g_{1i}(t)-g_{ij}(t)]^2 + [g_{2i}(t)-g_{2j}(t)]^2)dt \tag{63}$$

subject to the following constraints:

(i) *the equidistance constraint:* the quantities in (62) should be the same whenever $i \neq j$.

(ii) *average energy constraint:* (15) should be satisfied for both signal sets G_1 and G_2.

(iii) *peak-amplitude constraint:* (17) should be satisfied for both signal sets G_1 and G_2.

By paralleling the development leading to Lemma 1, we have the following.

Lemma 1: Given $\bar{s}_{max}$ as the maximum average signal counts per channel use in each polarization component and given (17) for both signal sets G_1 and G_2, then

$$d^2 \leq 4\left[\frac{g_{max} - g_{min}}{g_{max} + g_{min}}\right] \bar{s}_{max} \quad . \tag{64}$$

Furthermore, equality holds if and only if both: (a), at any time $t \in [0,T]$, all signals in G_1 and G_2 take on the value g_{min} except at most one in G_1 and one in G_2 which takes on value g_{max}; and (b),

$$\frac{1}{q} \sum_{i=1}^{q} \int_0^T g_{ki}^2(t)dt - g_{min}^2 T = \bar{s}_{max}$$

both both k=1 and k=2.

A signal set $G = G_1 \cup G_2$ that is equidistant, in the sense that the quantities in (62) are the same whenever $i \neq j$, and that achieves the upper bound in Lemma 1' with equality, and which therefore maximizes $R_{0,\infty}$ when the average energy constraint predominates in each polarization component, is characterized in the following lemma for q even.

Lemma 2: If q is even and $\bar{s}_{max}$ satisfies (19a), equality is achieved in (64) by the following signal set: for $1 \leq i \leq (q/2)$ and $j = i + (q/2)$,

$$g^*_{1i}(t) = g^*_{2j}(t) = \begin{cases} g_{max}, & (i-1)2T/q \leq t < (i-1+\epsilon)2T/q \\ g_{min}, & \text{otherwise for } 0 \leq t \leq T \end{cases}$$

$$g^*_{1j}(t) = g^*_{2i}(t) = g_{min}, \qquad 0 \leq t \leq T.$$

where ϵ is given in (26).

The verification of Lemma 2′ is straightforward paralleling the verification of Lemma 2. It is interesting to note for q=4 that this signal set, then called "quatranary pulse modulation," is used in the one gigabit per second optical communication system reported by M. Ross, et al. [15].

TABLE 1. Code Constraints

q	m	Hamming Distance	$\sum_{i=1}^{m} \rho_i(t)$
even	q-1	½q	½q
even	2(q-1)	q	½q
odd	q	½(q+1)	½(q-1)
odd	q	q+1	½(q-1)

TABLE 2. Degradation Due to Finite Quantization

$\bar{n}_{eff}$	$\bar{s}^*$	$R^*_{0,\infty}$	p^*	$\bar{s}$	$10\log(\bar{s}/\bar{s}^*)$
1	2.35	0.53	0.038	3.8	2.09
5	4.86	0.65	0.020	7.0	1.58
10	6.65	0.68	0.016	9.25	1.43
20	9.10	0.70	0.014	12.7	1.45
40	12.45	0.71	0.013	16.9	1.33

ACKNOWLEDGEMENT

We particularly thank Professor J. L. Massey of U.C.L.A. for several enlightening discussions about the cutoff rate criterion. Professors S. Mitter (M.I.T.) and M. Pursley (U. of Ill.) and Dr. M. Hurtado (I.B.M.) also provided useful insights into various parts of the development. We are indebted to Professor R. Kennedy of M.I.T. for identifying an error in an earlier version of the manuscript.

APPENDIX (derivation of (27))

Let $Q = \{1,2,\ldots,q\}$ and define J* by

$$J^* = \max_{g_k(\cdot),k\in Q} \int_0^T \sum_{i,j\in Q} [g_i(t)-g_j(t)]^2 dt. \tag{A1}$$

Then

$$J^* \leq T \max_{t\in[0,T]} \max_{g_k(t),k\in Q} \sum_{i,j\in Q} [g_i(t)-g_j(t)]^2, \tag{A2}$$

with equality if and only if the integrand in (A1) is a constant independent of t. Thus, we consider the problem of choosing q real numbers g_k, $k\in Q$ to maximize

$$I(\underset{\sim}{g}) = \sum_{i,j\in Q} (g_i-g_j)^2 \tag{A3}$$

subject to $g_{min} \leq g_i \leq g_{max}$, $i\in Q$. A necessary condition for g_i^*, $i\in Q$ to minimize -I(and so maximize I) is the existence of 2q real numbers $\nu_i \geq 0$, $\mu_i \geq 0$, $i=1,2,\ldots,q$, such that [16][1]:

$$L(\underset{\sim}{g}^*,\mu,\nu) \leq L(\underset{\sim}{g},\mu,\nu), \text{ for all } \underset{\sim}{g} \text{ in } R^q \tag{A4}$$

$$g_i^* < g_{max} \text{ implies } \mu_i = 0, \tag{A5}$$

$$g_i^* > g_{min} \text{ implies } \nu_i = 0, \tag{A6}$$

where the Lagrangian L is defined by

$$L(\underset{\sim}{g},\mu,\nu) = -\sum_{i,j\in Q} (g_i-g_j)^2 + \sum_{i\in Q} \mu_i (g_i-g_{max}) + \sum_{i\in Q} \nu_i (g_{min}-g_i). \tag{A7}$$

Equating to zero the derivative of L with respect to g_i, we obtain for $i \in Q$

$$-2q\, g_i^* + 2q\, c^* + \mu_i - \nu_i = 0, \tag{A8}$$

where $c^* = q^{-1} \sum_{i \in Q} g_i$. From (A5) and (A6), if $g_{min} < g_i^* < g_{max}$, then $\mu_i = \nu_i = 0$, and

$$g_i^* = c^*. \tag{A9}$$

Thus, each g_i^* takes on one of three values: g_{min}, g_{max}, or c^*. Let there be n_{min}, n_{max}, and $q - n_{min} - n_{max}$ of these, respectively. Then, from the definition of c^*. we have

$$q\, c^* = n_{min} g_{min} + n_{max} g_{max} + (q - n_{min} - n_{max}) c^*$$

or

$$c^* = (n_{min} g_{min} + n_{max} g_{max})/(n_{min} + n_{max}). \tag{A10}$$

Furthermore,

$$\begin{aligned} \sum_{i,j \in Q} (g_i^* - g_j^*)^2 &= 2 n_{min} n_{max} (g_{max} - g_{min})^2 \\ &+ 2 n_{max} (q - n_{min} - n_{max}) (g_{max} - c^*)^2 \\ &+ 2 n_{min} (q - n_{min} - n_{max}) (c^* - g_{min})^2 \end{aligned} \tag{A11}$$

Substituting (A10) into (A11) and simplifying, we obtain

$$I(\underset{\sim}{g^*}) = L(\underset{\sim}{g^*}, \mu, \nu) = 2q (g_{max} - g_{min})^2 \left(\frac{1}{n_{min}} + \frac{1}{n_{max}} \right)^{-1}, \tag{A12}$$

and this is to be minimized subject to $0 \leq n_{min} + n_{max} < q$, n_{min} and n_{max} being non-negative integers, which is the same as the minimization of $(1/n_{min}) + (1/n_{max})$ with the same constraints. The solution to this is: for q even, $n_{min} = n_{max} = q/2$;

and for q odd, either $n_{min} = (q+1)/2$ and $n_{max} = (q-1)/2$ or $n_{min} = (q-1)/2$ and $n_{max} = (q+1)/2$. Thus,[2] for q even, we have

$$I(\underset{\sim}{g}^*) = \tfrac{1}{2}\, q^2 (g_{max} - g_{min})^2 \tag{A13}$$

and, for q odd,

$$I(\underset{\sim}{g}^*) = \tfrac{1}{2}(q+1)(q-1)(g_{max} - g_{min})^2. \tag{A14}$$

The corresponding upper bounds on $d^{*2} = q^{-1}(q-1)J^*$ are, from (A1),

$$d^{*2} \leq \begin{cases} \tfrac{1}{2}q(q-1)(g_{max} - g_{min})^2 T, & q \text{ even} \\ \tfrac{1}{2}(q+1)q^{-1}(g_{max} - g_{min})^2 T, & q \text{ odd.} \end{cases} \tag{A15}$$

REFERENCES

1. S. Karp, E. L. O'Neill, and R. M. Gagliardi, "Communication Theory for the Free Space Optical Channel," Proc. IEEE, vol. 58, no. 10, pp. 1611-1626; October 1970.

2. E. V. Hoversten, R. O. Harger, and S. J. Halme, "Communication Theory for the Turbulent Atmosphere," Proc. IEEE, vol. 58, no. 10, pp. 1626-1650; October 1970.

3. E. V. Hoversten, "Optical Communication Theory," Chapter F8 in: Laser Handbook, F. T. Arecchi and E. O. Schulz-DuBois (editors), North Holland Publ. Co.; 1972.

4. W. K. Pratt, Laser Communication Systems, John Wiley and Sons, Inc., New York; 1969.

5. S. Karp and R. M. Gagliardi, Optical Communications, John Wiley and Sons, Inc., New York; 1976.

6. J. M. Wozencraft and R. S. Kennedy, "Modulation and Demodulation for Probabilistic Coding," IEEE Trans. on Info. Th., vol. IT-12, pp. 291-297; July 1966.

7. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," IEEE Trans. on Info. Th., vol. IT-13, pp. 260-269; April 1967.

8. J. L. Massey, "Coding and Modulation in Digital Communications," Proc. International Zurich Seminar on Digital Communications, Switzerland; March 12-15, 1974.

9. J. L. Massey, course notes for EE 453, Department of Electrical Engineering, University of Notre Dame, Indiana; 1976.

10. S. W. Golomb (Ed.), Digital Communications,, Prentice-Hall, Inc., Englewood Cliffs, N.J., p. 53; 1964.

11. W. W. Peterson and E. J. Weldon, Jr., Error-Correcting Codes, M.I.T. Press, Cambridge, Mass., Section 5.6; 1972.

12. S. D. Personick, P. Balaban, J. H. Bobsin, and P. R. Kumar, "A Detailed Comparison of Four Approaches to the Calculation of the Sensitivity of Optical Fiber System Receivers," IEEE Trans. on Comm., vol. COM-25, pp. 541-548; May 1977.

13. J. E. Mazo and J. Salz, "On Optical Data Communication via Direct Detection of Light Pulses," The Bell System Tech. J., vol. 55, pp. 347-369; March 1976.

14. D. L. Snyder, Random Point Processes, John Wiley and Sons, Inc., New York, p. 156; 1975.

15. M. Ross, P. Freedman, J. Abernathy, G. Matassov, J. Wolf, and J. D. Barry, "Space Optical Communications with the Nd:YAG Laser," Proc. IEEE, Vol. 66, Nl. 3, pp. 319-344; March, 1978.

16. D.G. Luenberger, Introduction to Linear and Nonlinear Programming, John Wiley and Sons, Inc., New York; 1969.

FOOTNOTES

[1]This is a necessary condition for a *regular* point $\underset{\sim}{g}^*$ to minimize -I subject to $g_{min} \leq g_i \leq g_{max}$, $i \in Q$. Since $g_i - g_{max} \leq 0$ and $g_{min} - g_i \leq 0$ cannot be simultaneously active (that is, satisfied with equality), it is evident that the set of gradient vectors of the active constraints can include $\underset{\sim}{e}_i$ (the i-*th* natural basis vector) or $-\underset{\sim}{e}_i$, but not both (and possibly neither). Thus, the set of gradient vectors of the active constraints is linearly independent for any $\underset{\sim}{g}$, and any $\underset{\sim}{g}$ is therefore regular.

[2]Because this is the only solution to the necessary conditions (A4)-(A6), either it is the maximum of I or none exists. But, by the Weierstrass theorem, the continuous function I defined by (A3) achieves its maximum on the compact subset of R^q defined by $g_{min} \leq g_i \leq g_{max}$. Thus, it is, indeed, the maximum.

FIGURE CAPTIONS

Figure 1. Optical Digital-Communication System

Figure 2. d^2 for Optimal Signal Sets

Figure 3. Cutoff Rate as a Function of Signal-to-Noise Ratio

Figure 4. Optimal Signal Energy as a Function of Noise Energy Per Channel Use

Figure 5. Effect of Input-Alphabet Dimension on Cutoff Rate

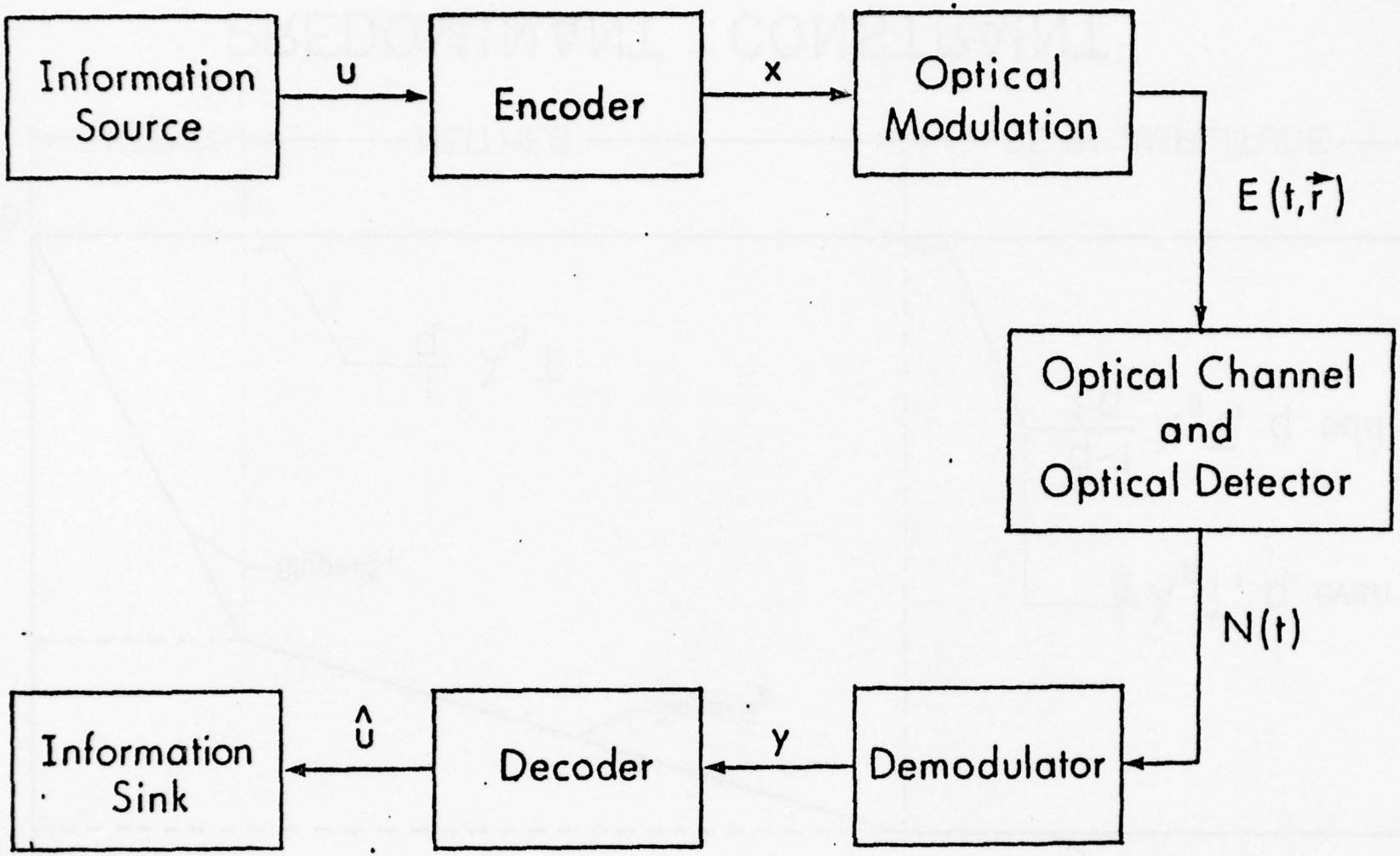


Figure 1. Optical Digital-Communication System

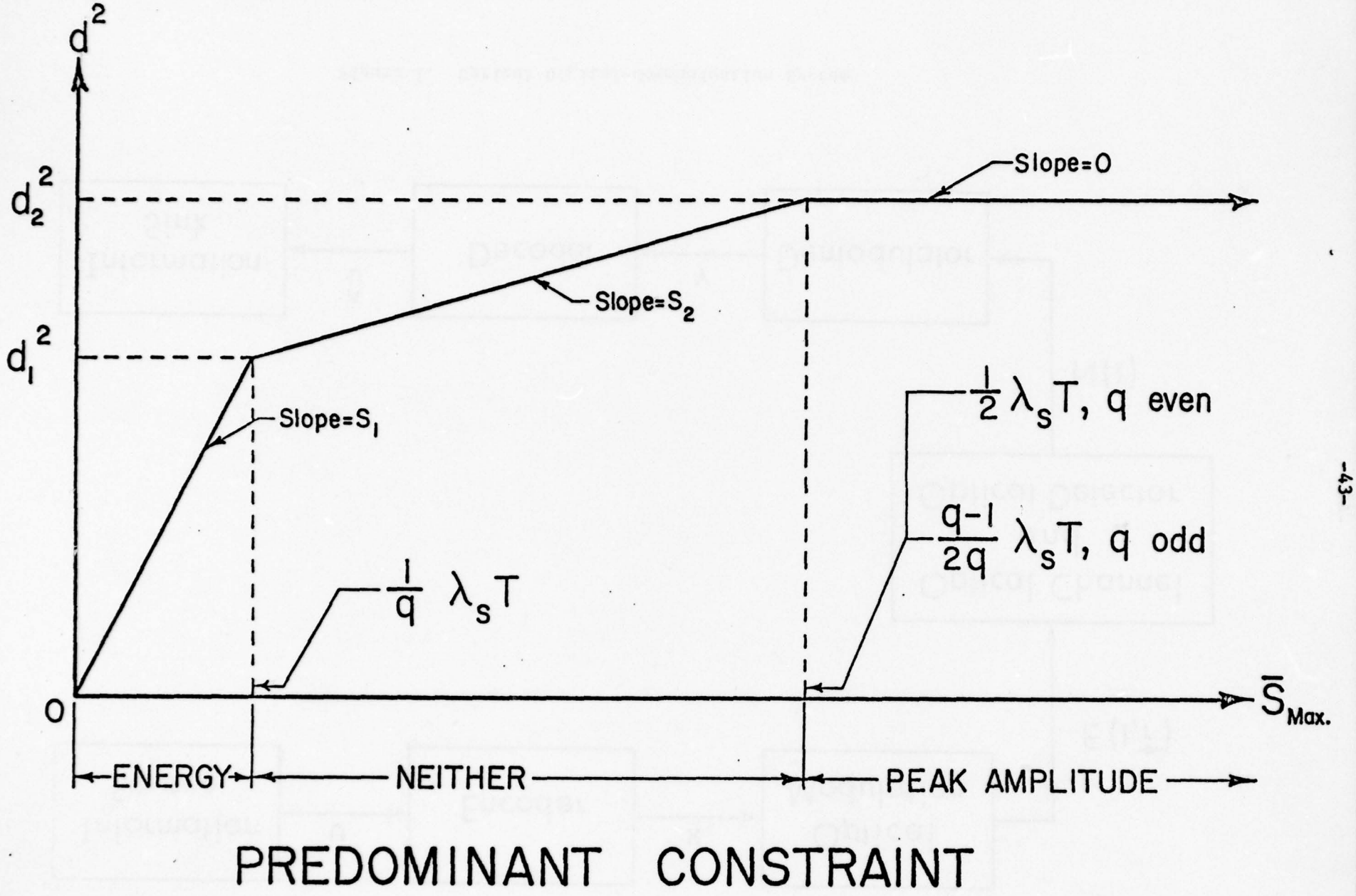


Figure 2. d^2 for Optimal Signal Sets

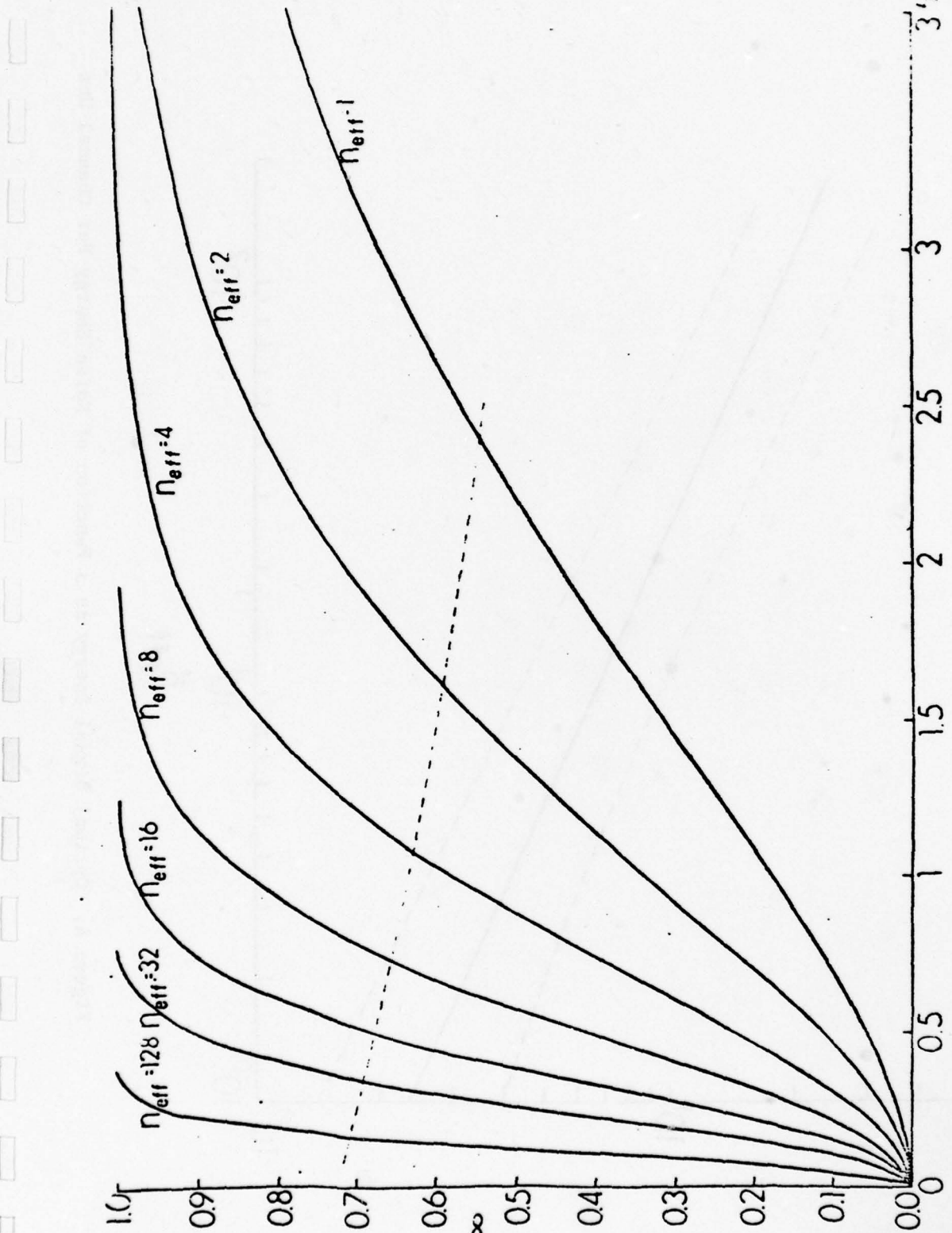


Figure 3. Cutoff Rate as a Function of Signal-to-Noise Ratio

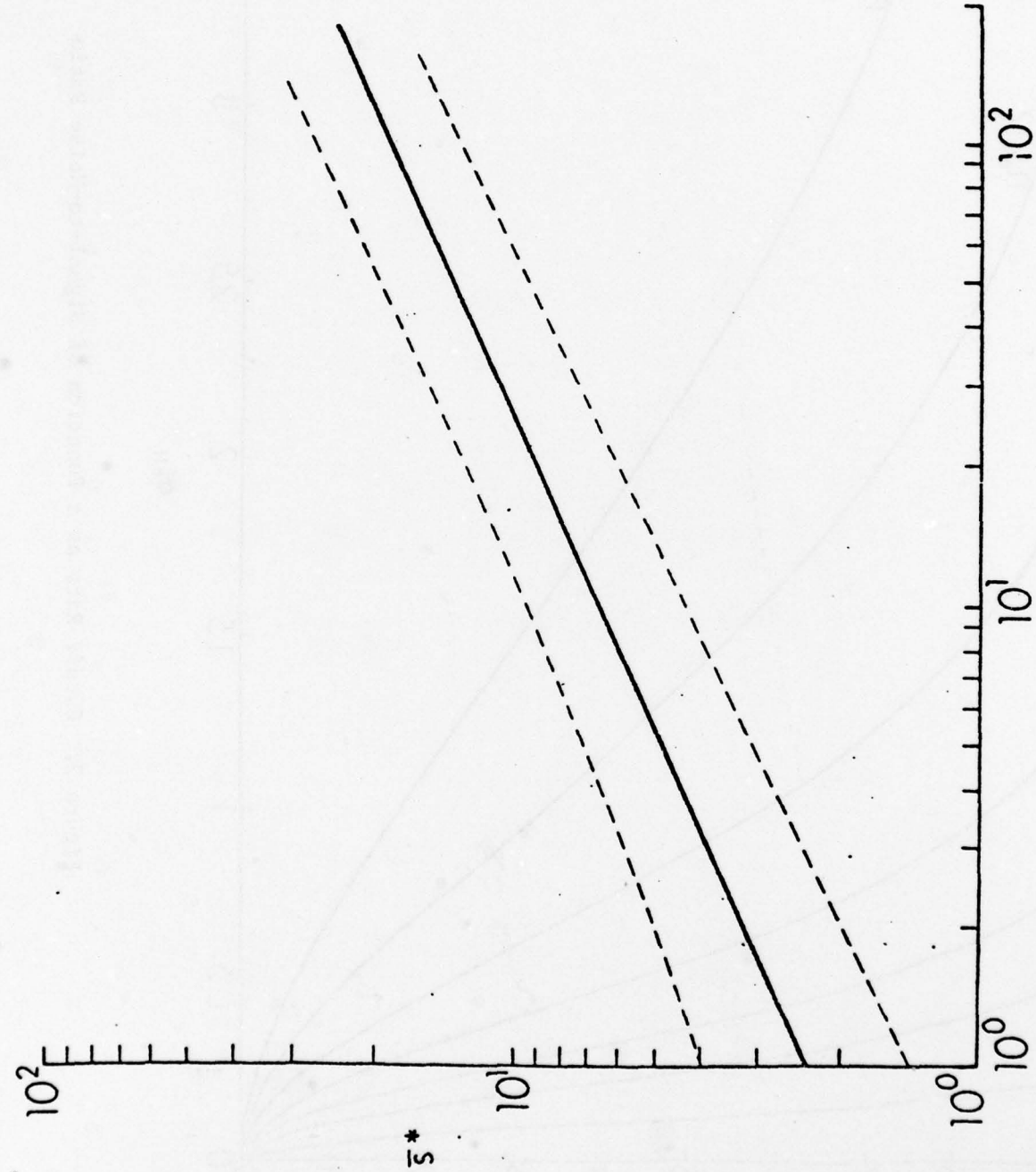


Figure 4. Optimal Signal Energy as a Function of Noise Energy Per Channel Use

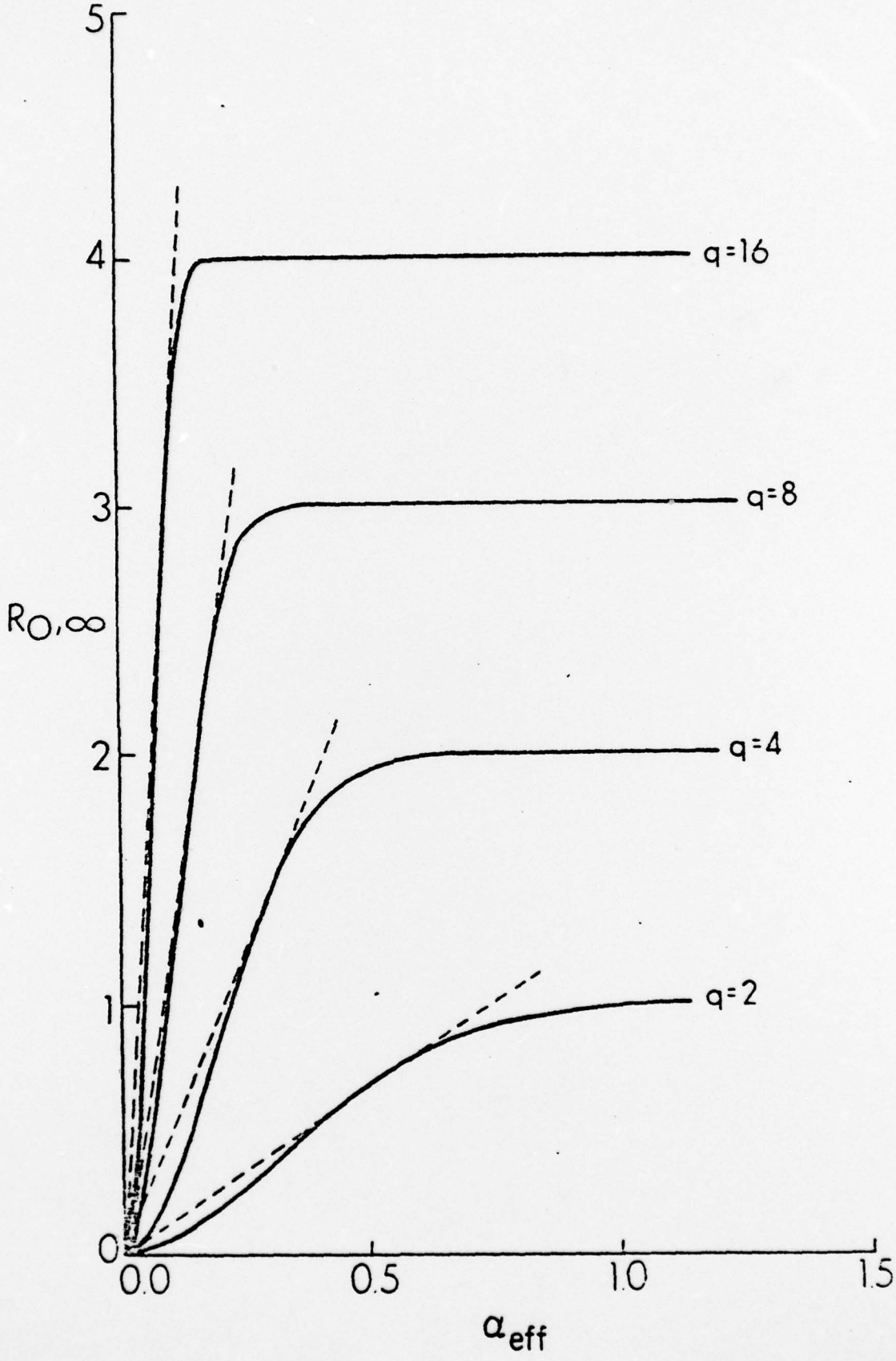


Figure 5. Effect of Input-Alphabet Dimension on Cutoff Rate

APPENDIX 5

Reprint of Paper:

"A Decentralized Shortest Path Algorithm," Jeffrey M. Abram and Ian B. Rhodes, Proceedins of the Sixteenth Allerton Conference on Communications, Control and Computing, University of Illinois, October 4-6, 1978, pp. 271-277.

(Pages 163 - 170)

A DECENTRALIZED SHORTEST PATH ALGORITHM

JEFFREY M. ABRAM and IAN B. RHODES
Department of Systems Science and Mathematics
Washington University, St. Louis, Missouri 63130

ABSTRACT

A decentralized algorithm for determining the shortest paths in a network is presented. Using information received only from neighboring nodes, a sequence of additions and comparisons is performed at each node in the network. Convergence to the optimal solution takes place in finite time.

I. INTRODUCTION

A common problem in graph theory is that of finding the shortest paths between all pairs of nodes in a network, and numerous algorithms exist for its solution, e.g. [1], [2], [3]. Nearly all of these algorithms are developed under the presumption that the computations will be performed by a decision maker with knowledge of the entire graph topology and of all branch lengths. The implementation of algorithms of this type can be thought of as requiring each node to transmit distance and topology information to a central controller, who is then responsible for solving the problem. After the shortest paths have been determined, the controller will send the appropriate routing information to each of the nodes. In a large network this could involve a significant amount of communication. Additionally, for some networks establishment of a central controller may be expensive, infeasible, or undesirable from a security or reliability viewpoint.

The purpose of this paper is to present a decentralized shortest path algorithm in which each node computes its shortest distance to each other node, while requiring communication only with its adjacent nodes, thus eliminating the need for a control center. This algorithm does not arise from any new concept; it is based primarily on a shortest path algorithm of Ford and Fulkerson [4]. In a similar spirit to the modifications made by Lau, Persiano and Varaiya [5] to a similar algorithm for the maximum flow problem, the Ford and Fulkerson algorithm is modified and reinterpreted in order to extract and emphasize its localized information requirements. Very little topological information is needed. Each node needs to know only which nodes are attached to the incoming branches, which are attached to the outgoing ones, and the lengths of the links to the outgoing nodes. A node sends information only to its incoming nodes. For each ultimate destination, a node calculates an estimate of the shortest path via each of its outgoing links; the smallest of these is taken to be its estimate of the shortest path to that destination. All of these approximate shortest distances will have become true shortest distances by the time the algorithm converges. Convergences is guaranteed in finite time, even if the algorithm is implemented by the nodes of the network in an asynchronous manner.

II. THE ALGORITHM

Consider a directed graph consisting of N nodes, denoted $\{1,2,\ldots N\}$, and a collection of arcs, $A = \{(i,j) : i,j \in N$ and there exists a direct link from i to $j\}$. To each arc $(i,j) \in A$ is associated a length $\ell(i,j)$. These lengths could represent physical distance, time, energy, money, or

any other quantity suitable for the network of interest. The lengths are unrestricted in sign, but the sum of the lengths in any closed loop of the network is assumed to be positive. Also, for every $i \in N$ define

$$I(i) = \{j : (j,i) \in A\}.$$
$$J(i) = \{j : (i,j) \in A\}.$$

We will refer to I(i) as the set of incoming nodes to i, and J(i) as the set of outgoing nodes from i. Each node maintains a matrix of shortest distances to each ultimate destination via each outgoing branch. In Fig. 1, d(i,j;k) represents the "current shortest distance" from i to j, given that k must be the next node along any path considered,

$$d(i,j) = \min_k d(i,j;k),$$

n(i,j) is the next node on the path that achieves the distance d(i,j). Row i is crossed out because it would represent distances from i to itself. For any $j \notin J(i)$, $j \neq i$, column j is crossed out because no direct link exists from i to j.

Initialization of the matrices requires only local topological information. Each node i begins by crossing out row i and all appropriate columns, as discussed previously. The diagonal elements represent direct link distances to other nodes. Thus, every diagonal element which is not crossed out (viz., those in the columns of the outgoing nodes) is assigned the length of the arc from i to the given outgoing node. Assume $\ell(i,j) = M$, a very large number, whenever $(i,j) \notin A$. Then in column i, $d(i,j) = \ell(i,j)$ and n(i,j) = j. In other words, since direct paths are the only ones known at this time, column i, which consists of the shortest paths to each destination based on information received to date, initially contains the lengths of the direct links to each node. Note that whenever d(i,j) = M, this indicates that no real path from i to j has yet been found. (M should be treated like ∞. M + d = M for any "real distance" d.)

Now, for purposes of analysis, imagine stacking the N X N distance matrices in order, one above the other, to form a distance cube with the matrix of node 1 at the top, and that of node N at the bottom. The basic idea behind the algorithm is the following. Suppose that node i makes a change (this includes the initialization step) in some d(i,j) component in column i. The only distances that are directly affected by this change are the distances to j, via i, for each of the incoming nodes to i. The distance cube is arranged in such a way that these affected elements are those that are not crossed out and lie along the vertical line that passes through d(i,j). That is, information transmission is purely vertical. Thus, distances to other nodes are received from the set of outgoing nodes, making it possible to calculate distances via these nodes.

More specifically, at each node i the algorithm is begun by initializing the distance matrix. Each distance in column i must then be transmitted up and down the corresponding vertical line. Node i now does nothing until a new distance, say to node k, is received from an outgoing neighbor j. To calculate the distance to k via j, the distance received from j must be added to the direct distance, $\ell(i,j)$, from i to j, which can always be found in the (j,j) element of matrix i. This sum is the new d(i,k;j) and replaces that stored in the (k,j) element of node i's matrix. If it is larger than d(i,k), node i does nothing because a better path has already been found. If it equals d(i,k), j can be included in n(i,k) because there is a tie for shortest path. If it is less than d(i,k), it becomes the "current shortest distance" from i to k, so in the (k,i) element of this distance matrix, node i replaces d(i,k) by d(i,k;j), and n(i,k) by j, and transmits this new distance along the vertical line through his (k,i)

element. The algorithm continues in this manner until no more changes can be made. At this point, each node will know the shortest distance to each destination (or that no path exists, which will be reflected as a shortest path length M), the next node in the path that achieves this distance, and the shortest distance via each alternative outgoing node.

Observe that in any given node's distance matrix the operations in any row are self-contained and independent of those in any other row: the initialization step and each subsequent addition and comparison operation involves only elements in a specific row. This means that the operations performed by a given node for one ultimate destination are independent of those for any other ultimate destination. It has also been noted that in the distance cube constructed by vertically stacking the individual distance matrices, the only communication takes place along vertical lines. This reflects the fact that the information transfer concerning one particular ultimate destination is independent of that for any other. Together, these observations mean that both information transfer and addition-comparisons are independent from one ultimate destination to the next. In the distance cube, this means that the vertical "slices" corresponding to each fixed ultimate destination are self-contained in so far as both communication and addition-comparison operations are concerned. This decomposition property is the basis for our proof of convergence in the next section. It should be emphasized, however, that the topological information required to construct and update each of these vertical slices is <u>not</u> localized. Indeed, interpretation of the algorithm for these vertical planes is closely related to the "centralized" Ford-Fulkerson algorithm. What is important for our purposes is that while the decentralized nature of the algorithm is exhibited by separating the distance cube into horizontal slices, convergence is most easily proven by thinking of the cube as being separated into vertical slices that are self-contained and for which convergence can be proven separately and individually. Clearly, the two are equivalent since they are simply alterative decompositions of the same distance cube.

III. CONVERGENCE OF THE ALGORITHM

The convergence of the algorithm will be proved, by induction, for an arbitrary vertical slice. Since the algorithm can be applied independently to each vertical matrix, convergence for a vertical slice implies convergence for the entire distance cube. Consider the vertical matrix composed of all of the row j's,i.e. corresponding to node j being the common ultimate destination. This matrix will take the form given in Fig. 2. Row j is crossed out because node j is not interested in distances to itself. As an example of the fact that arcs may not exist between all pairs of nodes, the (1,2) element has been crossed out, indicating that in this case the graph contains no direct link from node 1 to node 2. Each diagonal element contains the current shortest distance to destination j and the next node in the corresponding shortest path. Distance changes in any diagonal element will be communicated throughout the corresponding column.

Suppose j is an isolated node, so that no node has a path into j. Then each diagonal element will initially have $d(i,j) = M$. After these M's have all been transmitted, all distances in the vertical matrix will have the value M, no further changes can be made and the algorithm stops. Thus, the final matrix does indeed indicate that no node can find a path to j. Now suppose that at least one node has a path to j. Define

$$S(m) = \{i : \exists \text{ a shortest path from } i \text{ to } j \text{ containing exactly } m \text{ arcs}\}$$

Since for some node at least one path exists to j, an optimal path exists. By the Principle of Optimality, the last link in this path, say (k,j), must

be a shortest path from k to j. Thus $k \in S(1)$ and S(1) is not empty. The diagonal element corresponding to each node in S(1) will initially contain the direct distance to j. By the definition of S(1), these distances are optimal. At some time, T(1), each of these distances will have been transmitted throughout its column, and each of the columns corresponding to the nodes in S(1) will now be optimal. In general, define T(m) to be the time at which each node in S(m) has transmitted its true shortest distance to j. Now assume that at time T(m), the columns associated with the nodes in $S(1) \cup S(2) \cup \ldots \cup S(m)$ have been optimized. By the Principle of Optimality, for each node in S(m+1), the shortest paths to j with m+1 arcs must all involve going first through a node in S(m). Therefore, at time T(m), after each of the nodes in S(m) has transmitted its shortest distance, each row corresponding to a node in S(m+1) will have the shortest distance to j in one of its S(m) columns. Since the distances in each of these rows are actual distances, they must be greater than or equal to the shortest distances to j. Thus, after comparisons are performed, the diagonal elements in the S(m+1) rows will contain true shortest distances to j. These distances will be transmitted, and at time T(m+1), the columns corresponding to the $S(1) \cup S(2) \cup \ldots \cup S(m+1)$ nodes will all be optimal. The only fact still to be proven, is that the algorithm stops. Let m' be the smallest integer such that $S(1) \cup S(2) \cup \ldots \cup S(m')$ contains all of the nodes in the graph. The existence of such an integer is guaranteed by the fact that m' cannot exceed N-1, which is the maximum number of links possible in any shortest path. Obviously, at time T(m'), every column in the matrix will consist entirely of optimal distances. At this point, the algorithm must stop.

In order to calculate an upper bound on the number of operations required, the algorithm is modified to operate on a synchronous basis. The event-driven nature of the original algorithm, i.e. sending new distances as soon as they are calculated, is quite convenient for the nodes using the algorithm. However, this characteristic makes it difficult, if not impossible, to bound the number of operations. So we assume instead that nodes are forced to take turns, being allowed to transmit once every N units of time. Define a triple operation to be the sequence of performing an addition, a comparison, and a replacement if necessary. Again, consider an arbitrary vertical slice of the cube. During the first cycle of transmissions, the distances in each of the N-1 diagonal elements will be propagated, and each will cause at most N-2 triple operations. At the conclusion of this cycle, at least one column is optimal, and does not enter into future calculations. During the second cycle, at most N-2 diagonal elements will transmit changes, so a maximum of (N-2)(N-2) triple operations are performed during this cycle. The upper bound for total operations for one vertical matrix is given by

$$(N-2)(N-1) + (N-2)(N-2) + \ldots + (N-2) = \frac{1}{2} N(N-1)(N-2)$$

Therefore, the entire distance cube will be optimal after at most $\frac{1}{2} N^2(N-1)(N-2)$ triple operations. This is an order of magnitude larger than the algorithm in Hu[1], but there are other considerations. The calculations are shared by all of the nodes of the network, and each node does approximately the same amount of work as a central controller would. The principal advantage of the algorithm is that each node computes its own shortest distance matrix.

IV. MODIFICATIONS AND EXTENSIONS

Practical implementation of the algorithm will obviously differ somewhat from the form given above, which has been chosen with ease of presentation in mind. Construction and maintenance of the entire N X N distance

matrix by each node would be inefficient. Storage space should be allocated only for the columns corresponding to the set of outgoing nodes, as well as the column containing the "current shortest distances" and "next nodes". The distance cube cannot be constructed by any single node because that would require knowledge of the full topology of the network, so information transfer must be handled differently. In the distance cube model distances are propagated vertically, but these distances are ignored whenever the vertical line passes through a crossed-out element. Those elements along the line which are not crossed out are precisely the set of incoming nodes. Thus, when a node i has a new distance to be transmitted, he sends it only to the nodes $j \in I(i)$. The message sent must include the identity of the sender, the name of the destination to whom the distance has changed, and the new distance to that destination. The addition-comparision operation remains unchanged.

An additional topic of interest is topological changes in the network. A problem arises when a topological change causes one or more shortest paths to get longer. This can occur when an arc length increases or when there is a breakdown in a node or link. It is crucial in the convergence proof that when a $d(i,j;k)$ assumes the value of the true shortest path from i to j, it must be the smallest distance in row j and $d(i,j)$ will be assigned this optimal value. It is the propagation of these true shortest distances which guarantees the convergence of the algorithm. Fix an ultimate destination j and consider the corresponding vertical matrix. Suppose a topological change occurs, and some shortest distances increase. For the new topology, we can construct the sets $S'(m)$, which may differ from the sets $S(m)$. However, the $d(i,j;k)$'s in some rows may no longer be valid; they could be smaller than real distances to j if they correspond to paths affected by the topological change. Thus, in the convergence proof, there is no guarantee that when a true shortest distance enters a row, it will be smaller than the other distances in the row. It is possible that the algorithm will converge in some such cases, but a result along these lines has not yet been proved.

On the other hand, if a topological change decreases some shortest paths, while none increase, convergence can be proven. Suppose the change occurs, and consider vertical matrix j. Define the sets $S'(m)$ for the new topology. Since the nodes directly affected by a topological change will know about this change immediately, the distances in column j, i.e. the lengths of the direct links into j, will be correct immediately after the topological change. In particular, these direct distances will be in the rows of the $S'(1)$ nodes. The other distances in each of these rows will either be actual distances, or will be larger than actual distances. Therefore, the distance in column j will be the minimum in each $S'(1)$ row, and will be assigned to the diagonal element of the row. Then at time $T'(1)$, the $S'(1)$ columns will be optimal. The same argument is valid in the inductive step of the convergence proof.

V. CONCLUSION

A decentralized algorithm for finding shortest paths between all pairs of nodes in a graph has been presented. Each node requires only local topological information. All communication in the algorithm is between adjacent nodes. The algorithm can be implemented asynchronously, an advantage in networks where the individual nodes have different processing capabilities. The algorithm may require more operations than a centralized scheme, but the calculations are divided up among all of the nodes in the graph. There is also the unanswered question of convergence of the algorithm after certain topological changes. However, most importantly, the algorithm finds all optimal paths in the network, while giving the users the advantages of decentralized topological and information requirements.

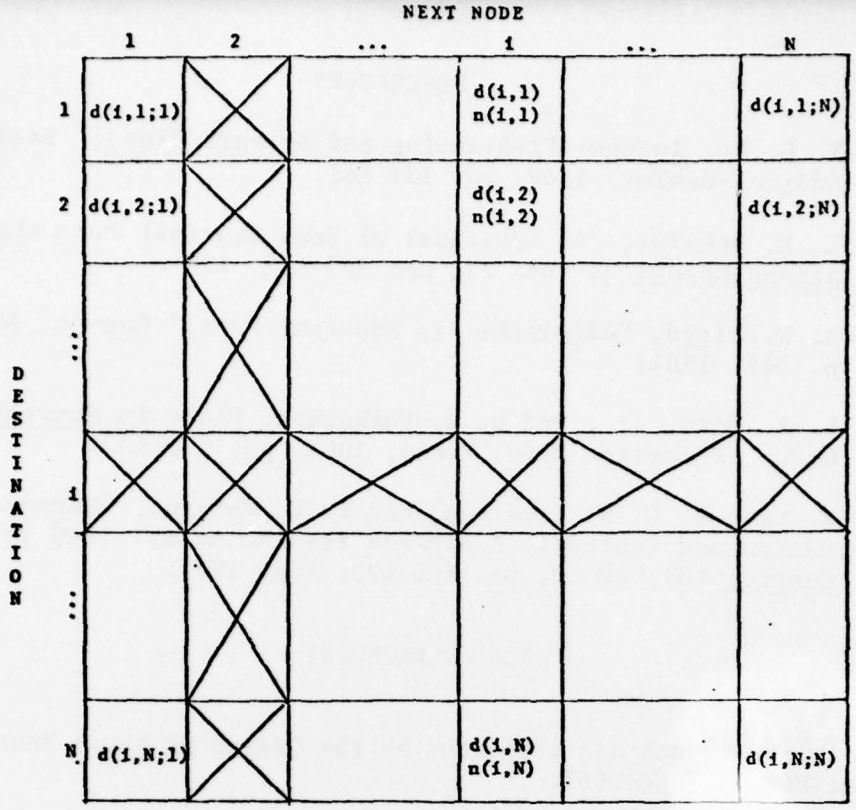


Fig. 1. Distance matrix for node i.

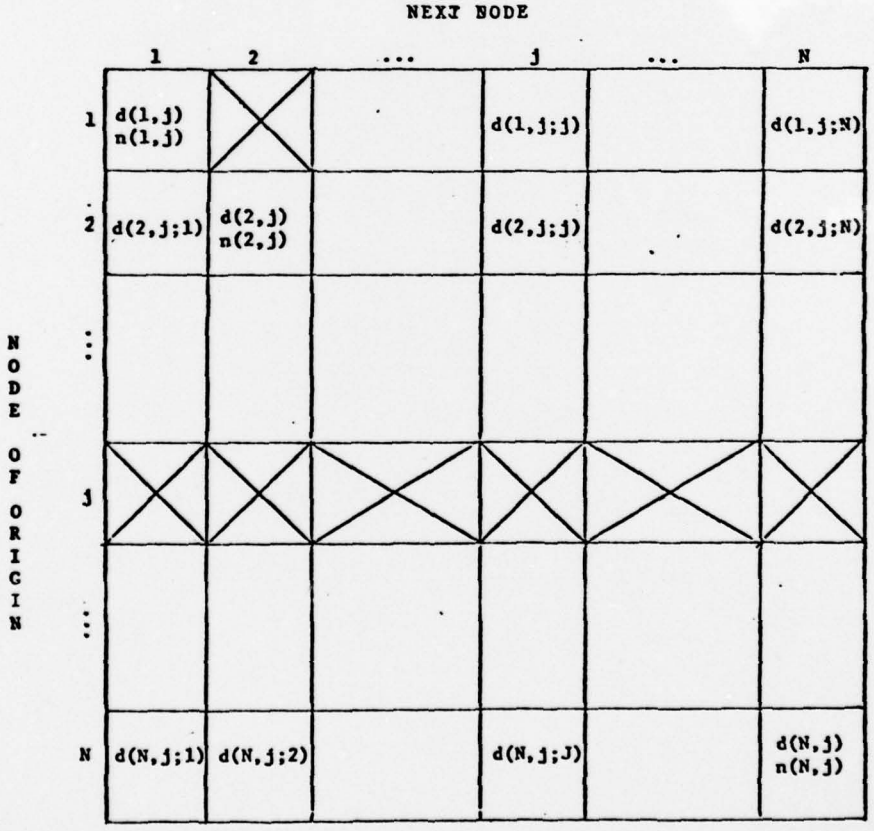


Fig. 2. Vertical "slice" j.

REFERENCES

[1] T. C. Hu, Integer Programming and Network Flows. Reading, Mass.: Addison-Wesley, 1969, pp. 151-161.

[2] S. E. Dreyfus, "An Appraisal of Some Shortest Path Algorithms," Operations Research, Vol. 17, pp. 395-412, 1969.

[3] R. W. Floyd, "Algorithm 97, Shortest Path," Commun. ACM, Vol. 5, p. 345, 1962.

[4] L. R. Ford, Jr., and D. R. Fulkerson, Flows in Networks. Princeton, N.J.: Princeton Univ. Press, 1962, pp. 130-133.

[5] R. Lau, R. C. M. Persiano, and P. P. Varaiya, "Decentralized Information and Control: A Network Flow Example," IEEE Trans. Automat. Contr., Vol. AC-17, pp. 466-473, Aug. 1972.

ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research under Contract N00014-76-C-0667.

APPENDIX 6

Reprint of Abstract:

"Quantization Loss in Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Proceedings of the Sixteenth Allerton Conference on Communication, Control and Computing, University of Illinois, October 4-6, 1978, p. 344.

QUANTIZATION LOSS IN OPTICAL COMMUNICATION SYSTEMS*

DONALD L. SNYDER and IAN B. RHODES
Washington University
St. Louis, Missouri 63130

ABSTRACT

In a digital communication system employing an optical carrier and direct detection and having five to forty dark-current counts per channel use, about 1.5db more signal energy is required with hard decisions to achieve the same cutoff rate as with infinitely soft decisions.

This research was supported by the National Science Foundation under Grant ENG 76-11565, by the National Institutes of Health under Research Grant RR00396 from the Division of Research Resources, and by the Office of Naval Research under Contract N00014-76-C-0667.

APPENDIX 7

Reprint of Abstract:

"Some Implications of the Cutoff Rate Criterion for Coded, Direct-Detection, Optical Communication Systems," Donald L. Snyder and Ian B. Rhodes, Abstracts of Papers, 1979 IEEE International Information Theory Symposium, Grignano, Italy, June 25-29, 1979, Session F2.

(Pages 173 - 174)

SOME IMPLICATIONS OF THE CUTOFF RATE CRITERION FOR CODED, DIRECT-DETECTION, OPTICAL COMMUNICATION SYSTEMS, Donald L. Snyder and Ian B. Rhodes (Washington University, St. Louis, Missouri 63130). The cutoff rate is derived for a digital communication system employing an optical carrier and direct detection. The coordinated design of the optical modulator and demodulator is then studied using the cutoff rate as a performance measure rather than the more commonly employed error probability. The best choice of optical modulation is identified for various relationships between peak amplitude and average energy constraints on the transmitted optical field. When the average energy constraint is predominant, pulse position modulation is shown to maximize the cutoff rate. When the peak amplitude constraint is predominant, Hadamard matrices can be used to define an optimum choice of modulation. Problems of efficient energy utilization, choice of input and output alphabet size, and the effect of random detector gain and addressed.

This work was supported by the National Science Foundation under Grant ENG 76-11565 and by the National Institutes of Health under Research Grant RR 00396 from the Division of Research Resources, and the Office of Naval Research under Contract N00014-76-C-0667.