71-27,100

HWANG, Myung Kyu, 1943-
ESTIMATION AND CONTROL OF STOCHASTIC CHEMICAL SYSTEMS.

California Institute of Technology, Ph.D., 1971
Engineering, chemical

University Microfilms, A XEROX Company, Ann Arbor, Michigan

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ESTIMATION AND CONTROL OF STOCHASTIC CHEMICAL SYSTEMS

Thesis by

Myung Kyu Hwang

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1971

(Submitted May 10, 1971)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

PLEASE NOTE:

Some pages have small
and indistinct type.
Filmed as received.

University Microfilms

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ACKNOWLEDGMENTS

I wish to express my hearty gratitude to my advisor, Professor Seinfeld, who suggested the topics, provided timely encouragement, and guided me with endless patience throughout my slow progress, both in research and the English language. I appreciate many invaluable discussions and suggestions from my other advisor, Professor Gavalas. Because of my two advisors, my graduate study at the California Institute of Technology has been enriched and fruitful. Also, I would like to mention many devotions and sacrifices by my parents through one of the most difficult periods in the Korean history.

During my graduate study I have been supported financially by the California Institute of Technology and National Science Foundation in the form of graduate teaching assistantships and graduate research assistantships. I am also indebted to Mrs. Ruth Stratton for her excellent typing.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

Chapter II

The control of nonlinear lumped-parameter systems is considered with unknown random inputs and measurement noise. A scheme is developed whereby a nonlinear filter is included in the control loop to improve system performance. Pure time delays in the control loop are also examined. A computational example is presented for the proportional control on temperature of a CSTR subject to random disturbances, applying a nonlinear least square filter.

Chapter III

Least square filtering and interpolation algorithms are derived for states and parameters in nonlinear distributed systems with unknown additive volume, boundary and observation noises, and with volume and boundary dynamical inputs governed by stochastic ordinary differential equations. Observations are assumed to be made continuously in time at continuous or discrete spatial locations. Two methods are presented for derivation of the filter. One is the limiting procedure of the finite dimensional description of partial differential equation systems along the spatial axis, applying known filter equations in ordinary differential equation systems. The other is to define a least square estimation criterion and convert the estimation problem into an optimal control problem, using extended invariant imbedding technique in partial differential equations. As an example, the derived filter is used to estimate the state and parameter in a nonlinear hyperbolic system describing a tubular plug flow chemical reactor. Also a heat

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

conduction problem is studied with the filtering and interpolation algorithms.

Chapter IV

New necessary and sufficient conditions are presented for the observability of systems described by nonlinear ordinary differential equations with nonlinear observations. The conditions are based on extension of the necessary and sufficient conditions for observability of time-varying linear systems to the linearized trajectory of the nonlinear system. The result is that the local observability of any initial condition can be readily determined, and the observability of the entire initial domain can be computed. The observability of constant parameters appearing in the differential equations is also considered.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table of Contents

	Abstract	iii
I.	Introduction	1
	1. Stochastic Estimation	2
	2. Stochastic Control	4
	3. Observability	6
	4. Summary of Contents	8
II.	Control of Nonlinear Stochastic Systems	10
	1. Introduction	10
	2. General Formulation	11
	3. Notation	15
	Appendix II-A	16
	Appendix II-B	22
III.	Optimal Least Square Filtering and Interpolation in Distributed Parameter Systems	25
	1. Introduction	25
	2. Problem Statement	26
	3. Optimal Least Square Filtering	29
	4. Optimal Least Square Interpolation	44
	5. Examples	55
	6. Remarks	60
	7. Figures	61
	Appendix III-A	65
	8. Notation	76
IV.	Observability of Nonlinear Systems	79
	1. Introduction	79
	2. Review of Linear Observability Theory	80
	3. Local Observability of Nonlinear Systems	84
	4. Examples	91
	5. Notation	94

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

V. Conclusions and Remarks 96

References 98

Propositions 104

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter I

INTRODUCTION

All descriptions of physical systems contain some degree of inherent uncertainties due to idealization of real processes in modeling and to whimsical environmental effects. Once the mathematical model of a dynamic system and observation process is given, the uncertainties are lumped as random interactions between the system and its surroundings. The random interactions are usually denoted by dynamical noise and observation errors. Therefore, the realization of the system is given as the model and observed data. With this realization the function of the given system is designed and controlled.

Conversely, fundamental questions can be posed whether the state of the system can be determined uniquely from given measurements and the process model, and whether the dynamic response of the given system can be controlled to achieve the prescribed performance specifications. The former problem is called the inverse or observability problem and the latter the controllability problem. With these observability and controllability assumptions the system can be analyzed and controlled with or without consideration of the randomness, i.e., stochastic or deterministic approach. However, the stochastic approach would be the only alternative if the process uncertainties become significant.

Consequently, the two important classes of engineering problems are: (1) how to estimate the state of the system, and (2) how to control the system with given noisy observations. The given measurements may

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be smaller dimensional quantities than the desired state of the system. The former is called stochastic estimation and the latter stochastic control.

1. Stochastic Estimation

Stochastic estimation of the state and parameters of a dynamic system has significant applications in modeling and adaptive control[59]. For example, estimating concentration, temperature, reaction constants, and catalyst activities in chemical reactors[18]; determining pressure history of oil reservoirs; and guiding and tracking of satellites[4,8,27] represent some applications of stochastic estimation. Also, stochastic estimation techniques can be applied to control stochastic dynamic systems.

Since Kalman initiated the filtering (sequential) estimation theory[30,31], exhaustive studies have been performed for lumped parameter systems, applying either probabilistic approaches[26,48] or optimization techniques[43]. Yet, exact solutions of nonlinear filtering problems have not been obtained, although many approximate nonlinear filters have been suggested[57]. The research activities on the topic can be summarized as the derivation of filtering equations[26, 43,48], error analysis of the filtering estimations[6,22,50,62], and duality study of the filtering theory and the optimal control theory[24, 30,31,41,61]. The excellent compilation for ordinary differential equation (O.D.E.) systems can be referred to Meditch[48], Jazwinski[26], and Lee[43].

For distributed parameter systems the similar approaches to lumped systems have been carried out only recently because of the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

mathematical difficulties involved. To make the literature survey compact, the following classification of available methods for the derivation of filtering equations is made:

1) Direct Method
   (a) Statistical approach
   (b) Optimal control approach
2) Indirect Method

In this classification "Direct Method" indicates the system equation is handled directly, and "Indirect Method" means that a finite differential difference approximation to the system equation is applied. The statistical approach requires known noise characteristics like zero mean Gaussian white noise assumption[51] to evaluate probability density functions for the system state. Thau[64] solved a linear case with one spatial measurement point, using minimum variance technique. Kushner[39] generalized Thau's case with continuous volume and boundary measurements, extending the Ito stochastic differential calculus[26] to linear parabolic systems. Kushner[39] also solved the case where boundary conditions contain stochastic inputs described by linear stochastic O.D.E.'s. In both Thau and Kushner's cases boundary conditions are linear and do not contain any additive boundary noise. Tzafestas and Nightingale[68] solved the linear case with additive boundary noise, applying the orthogonal projection lemma. The optimal control approach is the extension of a technique suggested by Detchmendy and Sridhar[11] for lumped systems. This method does not require noise specification. Balakrishnan and Lions[5] applied the least square method to solve the initial state estimation for a linear

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

deterministic system with Gaussian white measurement noise. Meditch[47] solved the filtering and smoothing problem for the similar case with a special form of boundary conditions. Lamont and Kumar[40] solved a nonlinear case with deterministic boundary conditions, using the invariant imbedding technique.

As an indirect method, Pell and Aris[52] applied finite discretization along the spatial axis to utilize the Kalmar-Bucy filter[31] for a linear system with deterministic boundary conditions. Tzafestas and Nightingale[67] used discretization along the time axis to apply the maximum likelihood method combined with differential dynamic programming[65] to obtain the nonlinear filter equations. Also they solved the smoothing and prediction estimation problems.

However, the above results cannot be applied to general cases such as parameter estimation problems or nonlinear systems with additive boundary noise and stochastic inputs in the volume and/or boundary which can be described by nonlinear stochastic O.D.E.'s. Also the fixed-point smoothing (interpolation) has not been solved for the above general cases.

2. Stochastic Control

To improve the performance of noisy dynamic systems, stochastic control theory has become an important area of optimal control theory after deterministic control theory was established. In a deterministic system the true state and output of the system can be predicted exactly if the initial condition and the control law are given. But in a stochastic system the state and the output are random variables, and so only the expectation values of the system variables can be found.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Hence probability density function of the system state updated with noisy observations, should necessarily be introduced. It is denoted by posteriori probability density function, which can be evaluated with the assumption of the Markovian process[3,51]. With the evolution of a posteriori probability density function represented by a set of integro-differential equations[17,38], the optimal feedback control of a lumped parameter system can be reduced to the solution of a functional differential equation with Gaussian white noise assumption[37]. Here the feedback control means that noisy observations have been used to generate the control law, otherwise it is the open loop control. If the system is linear, the structure of the optimal feedback control is a Kalman filter followed by a deterministic optimal controller, which is known as the certainty equivalence principle[3,26]. For a nonlinear case the functional differential equations[69] combined with probability density evolution equations are almost impossible to solve. Therefore, as a suboptimal feedback control of a nonlinear system, open loop control law which can be approximated by a finite dimensional O.D.E. was suggested[1,13,14]. However, this open loop approximation eliminates the advantages of using feedback schemes, especially when the noise level is significantly high. For distributed parameter systems the deterministic control theory is still in a developing stage, even though the corresponding maximum principle and the variational principle have been obtained for some cases[9,34,49]. For a linear stochastic partial differential equation (P.D.E.) system with a quadratic cost functional, Tzafestas and Nightingale[66] obtained optimal feedback control law, the certainty equivalence principle, and the Kalman's

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

duality principle. For general nonlinear P.D.E. systems, the corresponding probability density evolution equation or extended stochastic differential equation should be developed first for future study.

3. Observability

The fundamental assumption underlying the above discussion on stochastic estimation is observability, i.e., the convergence of the filtering estimates and their uniqueness. Since Kalman[32,33] introduced the notion, it has been studied in various fields such as system theory, identification and estimation theory, and optimal control theory. The observability study provides answers to the convergence of the state and parameter estimation schemes and the possible choice of observation processes for a given dynamic system.

The principal question is under what conditions we can have a one-to-one correspondence between the state and observation spaces. For lumped systems the problem can be stated as under what conditions we can find a unique initial condition of the system equation on the basis of given measurements. In deterministic linear O.D.E. cases, the explicit transition matrix representation of the solution enables us to determine observability conditions completely in terms of observability matrix which is independent of the initial condition. Also the duality relationship in a linear control system relates the observability and the controllability[28,61]. For linear stochastic systems it is expressed as Kalman's duality principle between the optimal control theory and the filtering theory. Furthermore, partial observability conditions have been studied with the structural theory of linear O.D.E. systems[71].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

As for deterministic nonlinear O.D.E. and P.D.E. systems, the research on observability is still in its initial stage although many engineering problems heavily depend on the topic. Regarding nonlinear O.D.E. cases, Lee and Markus[42] studied local observability around the equilibrium point for the first time, applying the results for linear time invariant systems. Kostyukovskii[35,36], and Griffith and and Kumar[23] investigated the one-to-one mapping conditions between the state and observation spaces, assuming necessary high order differentiability of both the system equations and observations with regard to their arguments and time. Roitenberg[55] considered the construction of a Lyapunov function[44] for the linearized system to study observability. For P.D.E. systems Wang[70] extended Kalman's approach, using semi-group operators (generalization of the Green's function) to obtain the conditions under which the initial condition of the system can be determined from measurements uniquely. Wang assumed the boundary conditions are known. Goodson and Klein[21] considered the observability of the first order P.D.E. systems from the viewpoint of solution uniqueness, given observation over a subset of the space-time domain. Their definition of observability is whether or not measurement data are sufficient to evaluate the unique solution to a P.D.E. in the absence of initial conditions and possibly boundary conditions. Nevertheless, its applicability is quite limited to simple linear cases because of mathematical difficulty of solution construction for given observations.

For linear stochastic O.D.E. systems[2], Sorenson[63] showed the connection between the nonsingularity of the covariance matrix of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the discrete Cox filter[10] and the corresponding deterministic observability. For continuous systems Bryson and Ho[7] and Meditch[48] presented a similar result. The parallel approach can be used to show the similar connection for distributed systems with Wang's result[70]. Figueiredo and Dyer[15] studied observability approximately for nonlinear stochastic O.D.E. cases by means of the convergence of the covariance equation.

4. Summary of Contents

The objectives of the present work are: to derive a most general nonlinear filter for distributed parameter systems which can be applied to nonlinear chemical reactor systems; to formulate a feedback scheme by which a nonlinear stochastic system can be controlled; and to investigate the nonlinear observability problem. In Chapter II the control of nonlinear stochastic systems is considered. A control scheme which includes a nonlinear filter to improve system performance is developed. The proposed scheme is applied to the control of a continuous stirred tank reactor (CSTR). Also various effects of noise on the dynamic response of the CSTR system are investigated. Two different derivations of a nonlinear filter for P.D.E. systems are presented in Chapter III. One is by the indirect method as an extension of Pell and Aris' result. The other is by the direct method using the least square method with invariant imbedding. Also nonlinear interpolation equations are obtained. The derived equations are applied to the state and parameter estimation of a heat conduction problem and of a plug flow tubular reactor system. In Chapter IV nonlinear observability of lumped

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

parameter systems is investigated. Necessary and sufficient conditions for local observability are obtained as an extension of the results for time varying linear systems.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter II

CONTROL OF NONLINEAR STOCHASTIC SYSTEMS

1. Introduction

The stochastic optimal feedback control law for nonlinear O.D.E. systems with Gaussian white noise has been reduced to the solution of a nonlinear functional differential equation with probability density evolution equations[17,38]. The solution is almost impossible to obtain either analytically or numerically. For the linear system with a quadratic performance index, the optimal feedback control law can be separated into a minimum variance filter followed by the corresponding optimal controller for the deterministic system. For nonlinear cases, usage of the optimal open loop control has been suggested with its O.D.E. approximation. The important control problems in chemical processes involve nonlinear stochastic systems with unknown noise characteristics. Often there are delays in control loop such as transportation lags.

The purpose of this study is as follows: to formulate an on-line feedback scheme by which a nonlinear stochastic system can be controlled; to extend the scheme to the case of time delays in the control loop and possibly to distributed systems.

The proposed scheme is a closed control loop which contains a nonlinear filter, i.e., a process control computer which integrates the filter equations. Hence the equivalent dynamic system consists of the actual system, observation process, and the filter. The control law is generated on the basis of the output of the equivalent system, i.e.,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the estimation of the system states. This scheme is applied to the proportional control of a CSTR with and without time delays. In addition, various noise effects are examined with the example system compared to the performance of the deterministic case.

Since this work has been published, this chapter presents only the general formulation of the problem. The detailed results are given in Appendix II-A. In Appendix II-B, the measurement noise effect on the control gain, i.e., the proportional constant, is examined by applying a describing function idea[20].

2. General Formulation

Let us consider a noisy dynamic system governed by

$$\dot{x}(t) = f(t,x(t),u(t)) + \xi(t) \tag{2.1}$$

Observations are related to the state of the system by

$$y(t) = h(t,x(t)) + \eta(t) , \quad 0 \leq t \leq T \tag{2.2}$$

where the state of the system is represented by the n-vector $x(t)$, the observations by the m-vector $y(t)$, the random inputs to the process by the n-vector $\xi(t)$, and the observation errors by the m-vector $\eta(t)$. The controller output is given by the r-vector $u(t)$. The possible delays in the control loop are the transport lag in the observation process of magnitude α_1 and a pure time delay of magnitude α_2 in the control action. The corresponding $u(t)$ and $h(t,x)$ will be $u(t-\alpha_2)$ and $h(t,x(t-\alpha_1))$ respectively. In this section the time delay case is not considered, but will be considered in Appendix II-A.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The performance index becomes

$$J = E\left\{ \int_0^T F(t,x,u)\, dt \right\} \tag{2.3}$$

The control problem is to choose $u(t)$ to minimize J, where $E\{\cdot\}$ represents the expectation operation. The open loop control depends only on the initial state of the system, while the feedback control depends on the state of the system at each moment. For a deterministic process with a given initial condition, the optimal open loop and the optimal feedback control solutions are equivalent. For stochastic system, the open loop control cannot compensate noise effects and is thus inapplicable.

To suppress dynamic and observation noises, the proposed scheme requires the estimate of the system state on the basis of noisy observations. For a system described by

$$\dot{x}(t) = w(t,x(t)) + \xi(t) \tag{2.4}$$

$$y(t) = h(t,x(t)) + \eta(t) \tag{2.5}$$

the corresponding filter equations, i.e., the differential equations which generate the estimate, minimizing the squared error functional, are[11]

$$\dot{\hat{x}}(t) = w(t,\hat{x}) + Ph_x^T Q(y - h(t,\hat{x})) \tag{2.6}$$

$$\dot{P}(t) = \hat{w}_x P + P\hat{w}_x^T + P[h_x^T Q(y - h(t,\hat{x}))]_x P + R^{-1} \tag{2.7}$$

where $\hat{x}$ is the estimate of the noisy system state, P is an $n \times n$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

symmetric matrix which is the covariance matrix of the estimate error in the linear case. Q is an $m \times m$ symmetric matrix which, in a linear case, is the inverse of the covariance matrix of $\eta(t)$. $R^{-1}(t)$ is an $n \times n$ symmetric matrix which becomes the covariance of $\xi(t)$ for the linear case. h_x and w_x are the appropriate Jacobian matrices. The initial conditions $\hat{x}(o)$ and $P(o)$ are taken as the expected initial state of the system and the covariance of its estimate respectively.

If the control law is represented by $u(t) = g(y(t))$ without the filter, it becomes $u(t) = g(h(\hat{x}(t)))$ with the filter. In other words, the control problem is changed with the filter so as to choose $u(t)$ in order to minimize

$$J = \int_0^T F(t,\hat{x},u)\, dt \tag{2.8}$$

subject to the constraints,

$$\dot{\hat{x}} = f(t,\hat{x},u) + Ph_x^T\, Q(y - h(t,\hat{x})) \tag{2.9}$$

$$\dot{P} = \hat{f}_x P + P\hat{f}_x^T + P[h_x^T\, Q(y - h(t,\hat{x})]_x\, P + R^{-1} \tag{2.10}$$

where $\hat{f}_x$ indicates $(\partial f/\partial x)_{\hat{x}}$. As shown in the above constraints, application of the filter increases the dimension of the system equations from n to $n(n+3)/2$. Owing to the increased dimension and the complexity of the feedback law, the proposed scheme in general is not feasible as an on-line scheme. This scheme is applicable only to the case where a simply prescribed controller function such as..

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$u(t) = g(t,x(t))$ can be chosen to achieve the performance specification. More precisely, the present scheme can be applied either with the choice of a simple control mode such as the proportional controller or with the choice of a simple instantaneous performance index. The latter case requires a linear structure of the control input to result the bang-bang control[42,53]. In the present study the former is considered, and the latter is examined by Seinfeld[58].

The proportional control mode combined with the proposed scheme is applied to regulate the output of a CSTR with an exothermic first-order reaction. The control action is to manipulate the heat transfer coefficient by means of heat removal with a coil or jacket. The detailed study is shown in Appendix II-A.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3. Notation

$E\{\cdot\}$ = expectation operator

f = n-dimensional vector function

g = r-dimensional vector function

h = m-dimensional vector function

J = performance index

$N(a,b)$ = normal distribution with mean a and covariance b

P = $n \times n$ covariance matrix

Q = $m \times m$ weighting matrix

R = $n \times n$ weighting matrix

t,T = time

u = r-dimensional control vector

w = n-dimensional vector function

x = n-dimensional state vector

y = m-dimensional measurement vector

GREEK SYMBOLS

$\alpha, \alpha_1, \alpha_2$ = time lags

η = m-dimensional noise vector

ξ = n-dimensional noise vector

SUPERSCRIPTS

$\hat{}$ = estimated value

$\cdot$ = time derivative

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix II-A

Reprinted from I&EC FUNDAMENTALS, Vol. 8, Page 257, May 1969
Copyright 1969 by the American Chemical Society and reprinted by permission of the copyright owner

CONTROL OF NONLINEAR STOCHASTIC SYSTEMS

JOHN H. SEINFELD, GEORGE R. GAVALAS, AND MYUNG HWANG

Chemical Engineering Laboratory, California Institute of Technology, Pasadena, Calif. 91109

The control of nonlinear lumped-parameter dynamical systems subject to random inputs and measurement errors is considered. A scheme is developed whereby a nonlinear filter is included in the control loop to improve system performance. The case of pure time delays occurring in the control loop is also treated. Computations are presented for the proportional control on temperature of a CSTR subject to random disturbances.

ALL real systems which one desires to control are subject to some degree of uncertainty. Even when the fundamental physical phenomena are known, the mathematical model may contain parameters whose values are unknown, or the actual system may be subject to unknown random disturbances. In designing a control system the easiest approach is to neglect the randomness associated with inputs, assign certain nominal values to parameters, and base the design on classical deterministic theory. However, it is obvious that a design based on deterministic control theory becomes inadequate when the process uncertainties become significant. The alternative is to consider the problem as one of control of a stochastic system.

The control of stochastic systems is of significant theoretical and practical importance. A large and elegant theory exists for the analysis of linear control systems subject to corrupting noise (Aris and Amundson, 1958b; Newton *et al.*, 1957; Soludovnikov, 1960). Recently, solutions have been obtained for the optimal control of linear systems with white noise forcing and quadratic performance criteria (Aoki, 1967; Kushner, 1965; Meditch, 1968; Sworder, 1967). The structure of the optimal feedback control in this case is a minimum variance (Kalman) filter followed by the optimal controller for the deterministic system. The optimal control of nonlinear systems with white noise inputs can be reduced to the solution of a set of nonlinear, integro-partial differential equations, which, as one might suspect, are almost impossible to solve. The key problems in chemical process control involve nonlinear systems with noisy inputs, the statistical properties of which are usually unknown. In addition, there are almost always delays in the control loop because of noninstantaneous control action and/or the time necessary for the analysis of measurements. A feasible way of handling such systems represents a challenging problem in chemical process control.

The objectives of this paper are: to formulate a scheme by which a nonlinear system with unknown random inputs can be controlled; to extend this scheme to the case of time delays in the control loop; and to apply the scheme to the proportional control of a continuous stirred-tank reactor

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(CSTR) and compare the performance of deterministic and stochastic control when the reactor is subjected to random disturbances.

The first alternative is to neglect the stochastic nature of the process entirely and rely upon the inherent property of feedback control to decrease the sensitivity of the entire loop to disturbances. It is expected that in the presence of substantial disturbances the controller would experience difficulty in regulating the system. The next alternative is to filter the system output in some manner before the output signal is sent to the controller. A simple R-C filter could be used to smooth the output signal; however, such a filter incorporates no information on the nature of the system. What is desired is not merely to smooth the output signal but to use this signal to estimate the state of the system. As noted above, the optimal feedback control of a linear plant with white noise disturbances and a quadratic performance index can be segmented into a Kalman filter followed by the optimal deterministic controller.

Consider a nonlinear system with random inputs and measurement errors for which we desire to estimate the actual state of the system. Optimal least square state estimates for such a system can be obtained from the solution of a set of nonlinear differential equations, termed the nonlinear filtering or sequential estimation equations (Athans *et al.*, 1968; Detchmendy and Sridhar, 1966; Gavalas *et al.*, 1969). The estimation equations use the process observations as input, their solution providing continuous estimates of the actual state of the process. When the process and output are linear, these equations are called the Kalman filter (Kalman and Bucy, 1961).

In this paper the scheme in which a filter—i.e., a process control computer which integrates the estimation equations—is incorporated prior to the controller in the control loop is examined. The filter provides an estimate of the actual state of the process at each instant, which, in terms of control, is the quantity of most interest. First, the proposed schemes, including the case of pure time delays in the control loop, are presented with the aid of block diagrams. The appropriate equations for the filter are derived in each case. Then the proportional control of a CSTR subject to random input disturbances and measurement errors is considered. The response of the CSTR with constant gain and no filter is compared to that with a filter in the control loop.

General Problem

Figure 1 shows the customary feedback control of a dynamical process. The state of the system is represented by the n-vector $\mathbf{x}(t)$, the observations or output by the m-vector

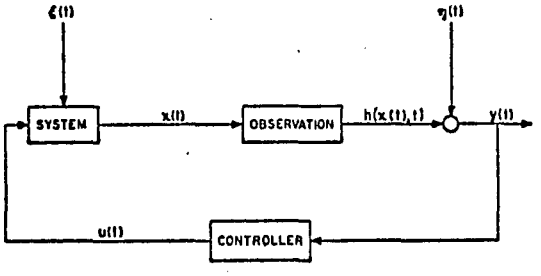

Figure 1. Feedback control of a system subject to disturbances

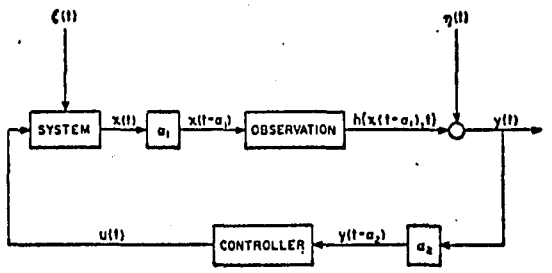

Figure 2. Feedback control of a system subject to disturbances and pure time delays

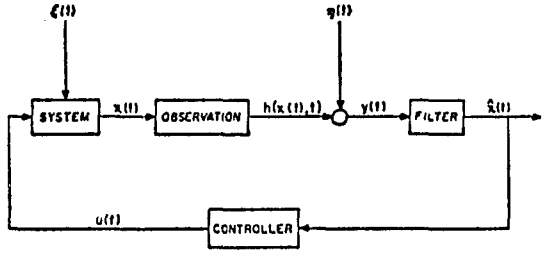

Figure 3. Feedback control of a system with filtering

No time delays

$\mathbf{y}(t)$, the random inputs to the process by the n-vector $\xi(t)$, and the measurement errors by the m-vector $\mathbf{n}(t)$. The controller output is represented by the r-vector $\mathbf{u}(t)$. If a pure time delay—e.g., transportation lag—of magnitude α_1 exists in the observation and a pure time delay of magnitude α_2 exists in the control action, the situation is depicted in Figure 2.

The control objective is to maintain the system at a desired state in spite of changes in the input. There are basically two types of input disturbances which affect the system: infrequent disturbances due to changes in the feed conditions or flow rate, and high frequency disturbances, the characteristic time of which is much smaller than the characteristic time of the system. Both types of disturbances normally occur in practice; however, if the amplitude of the high frequency noise is small, the conventional control schemes in Figures 1 and 2 should be successful in regulating the system. When the amplitude of the high frequency noise approaches the same order of magnitude as the amplitude of the low frequency upsets, the performance of the conventional system may be poor. In addition to noisy inputs, the output measurements invariably contain random errors, which arise typically as a result of inaccuracies in the measuring instruments. We will study the effect of both types of random disturbances on the performance of the controlled system. Figures 3 and 4 correspond to Figures 1 and 2 but include a filter after the measuring element. It is the comparative performance of Figures 1 and 3 and Figures 2 and 4 that we wish to consider.

Let us now formulate mathematically the situations depicted in Figures 1 to 4. For the scheme in Figure 1 the system is governed by

$$\dot{\mathbf{x}}(t) = \mathbf{f}[t, \mathbf{x}(t), \mathbf{u}(t)] + \xi(t) \qquad (1)$$

and the output is

$$\mathbf{y}(t) = \mathbf{h}[t, \mathbf{x}(t)] + \mathbf{n}(t) \qquad (2)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

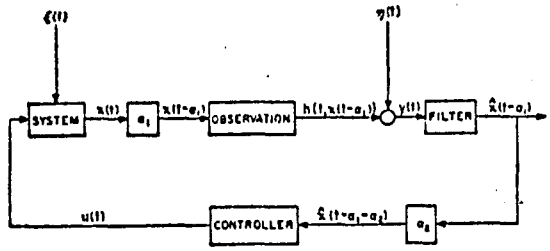

Figure 4. Feedback control of a system with filtering

Time delays in observation and control

It is assumed that $E[\xi(t)] = E[\mathbf{n}(t)] = 0$. Equations 1 and 2 are valid if the noisy inputs are additive or of low amplitude so that a linearization about expected values can be carried out. The state $\mathbf{x}(t)$, the output $\mathbf{y}(t)$, and the control $\mathbf{u}(t)$ are all random variables because $\xi(t)$ and $\mathbf{n}(t)$ are random variables. The control is a prescribed function of the output, $\mathbf{u}(t) = \mathbf{g}[\mathbf{y}(t)]$. For the case depicted in Figure 2 the system is governed by Equation 1 with the output

$$\mathbf{y}(t) = \mathbf{h}[t, \mathbf{x}(t - \alpha_1)] + \mathbf{n}(t) \tag{3}$$

while the *control action* is $\mathbf{u}(t) = \mathbf{g}[\mathbf{y}(t - \alpha_2)]$.

The schemes of Figures 3 and 4 have a filter after the observation element. For a system described by

$$\dot{\mathbf{x}}(t) = \mathbf{w}[t, \mathbf{x}(t)] + \xi(t) \tag{4}$$

$$\mathbf{y}(t) = \mathbf{h}[t, \mathbf{x}(t)] + \mathbf{n}(t) \tag{5}$$

the differential equations which constitute the least square filter are (Detchmendy and Sridhar, 1966)

$$\dot{\hat{\mathbf{x}}} = \mathbf{w}(t, \hat{\mathbf{x}}) + \mathbf{P}\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}[\mathbf{y} - \mathbf{h}(t, \hat{\mathbf{x}})] \tag{6}$$

$$\dot{\mathbf{P}} = \mathbf{w}_{\mathbf{x}}\mathbf{P} + \mathbf{P}\mathbf{w}_{\mathbf{x}}^T + \mathbf{P}\{\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}[\mathbf{y} - \mathbf{h}(t, \hat{\mathbf{x}})]\}_{\mathbf{x}}\mathbf{P} + \mathbf{R}^{-1} \tag{7}$$

where $\hat{\mathbf{x}}$ is the estimate of the actual system state $\mathbf{x}$, $\mathbf{P}$ is an $n \times n$ symmetric matrix which in the linear case is the covariance matrix of the estimate error, $\mathbf{Q}$ is an $m \times m$ symmetric matrix which in the linear case is the covariance matrix of $\mathbf{n}$, and $\mathbf{R}$ is an $n \times n$ symmetric matrix which in the linear case is the covariance matrix of ξ. $\mathbf{h}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{x}}$ are the appropriate Jacobian matrices e.g., $(\partial h_i/\partial x_j)_{\hat{\mathbf{x}}}$. The *initial conditions for Equations 6 and 7* are $\mathbf{x}(0)$ and $\mathbf{P}(0)$. These quantities are taken as the expected initial state of the system and the covariance of this estimate, respectively. If *no a priori information* is known, these values are chosen arbitrarily.

For the scheme of Figure 3 the system and observations are given by Equations 1 and 2. Since the filter has been inserted, $\mathbf{u}(t) = \mathbf{g}[\hat{\mathbf{x}}(t)]$ i.e., the control now depends on the filter output, the current state estimate $\hat{\mathbf{x}}(t)$. Thus, Equation 1 becomes

$$\dot{\mathbf{x}}(t) = \mathbf{f}\{t, \mathbf{x}(t), \mathbf{g}[\hat{\mathbf{x}}(t)]\} + \xi(t) \tag{8}$$

The filter equations are

$$\dot{\hat{\mathbf{x}}} = \mathbf{f}[t, \hat{\mathbf{x}}, \mathbf{g}(\hat{\mathbf{x}})] + \mathbf{P}\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}[\mathbf{y} - \mathbf{h}(t, \hat{\mathbf{x}})] \tag{9}$$

$$\dot{\mathbf{P}} = \mathbf{f}_{\mathbf{x}}\mathbf{P} + \mathbf{P}\mathbf{f}_{\mathbf{x}}^T + \mathbf{P}\{\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}[\mathbf{y} - \mathbf{h}(t, \hat{\mathbf{x}})]\}_{\mathbf{x}}\mathbf{P} + \mathbf{R}^{-1} \tag{10}$$

where $\mathbf{f}_{\mathbf{x}}$ indicates $(\partial f_i/\partial x_j)_{\hat{\mathbf{x}}}$.

We consider the delays α_1 and α_2 separately, as later we wish to consider the individual effect of each. When the sole delay is that in the observation with magnitude α_1, the state is governed by Equation 1 with the output given by Equation 3. Since the output at time t is related to the state at time $t - \alpha_1$, the filter produces an estimate of the state of the system at $t - \alpha_1$, $\hat{\mathbf{x}}(t - \alpha_1)$. The control action at time t, $\mathbf{u}(t)$, depends on the filter output at time t, or $\hat{\mathbf{x}}(t - \alpha_1)$. Thus, $\mathbf{u}(t) = \mathbf{g}[\hat{\mathbf{x}}(t - \alpha_1)]$. Then Equation 1 becomes in this case

$$\dot{\mathbf{x}}(t) = \mathbf{f}\{t, \mathbf{x}(t), \mathbf{g}[\hat{\mathbf{x}}(t - \alpha_1)]\} + \xi(t) \tag{11}$$

The filter equations are, correspondingly,

$$\dot{\hat{\mathbf{x}}}(\tau) = \mathbf{f}\{\tau, \hat{\mathbf{x}}(\tau), \mathbf{g}[\hat{\mathbf{x}}(\tau - \alpha_1)]\} + \mathbf{P}\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}\{\mathbf{y}(t) - \mathbf{h}]_{\tau}, \hat{\mathbf{x}}(\tau)]\} \tag{12}$$

$$\dot{\mathbf{P}}(\tau) = \mathbf{f}_{\mathbf{x}}\mathbf{P} + \mathbf{P}\mathbf{f}_{\mathbf{x}}^T + \mathbf{P}(\mathbf{h}_{\mathbf{x}}^T\mathbf{Q}\{\mathbf{y} - \mathbf{h}[\tau, \hat{\mathbf{x}}(\tau)]\})_{\mathbf{x}}\mathbf{P} + \mathbf{R}^{-1} \tag{13}$$

where $\tau = t - \alpha_1$. Although Equations 12 and 13 are integrated in real time with input as the current observation $\mathbf{y}(t)$, the result is the state estimate at $t - \alpha_1$.

For a delay α_2 in the controller, $\mathbf{u}(t) = \mathbf{g}[\hat{\mathbf{x}}(t - \alpha_2)]$. The system is governed by

$$\dot{\mathbf{x}}(t) = \mathbf{f}\{t, \mathbf{x}(t), \mathbf{g}[\hat{\mathbf{x}}(t - \alpha_2)]\} + \xi(t) \tag{14}$$

with output given by Equation 2. Since $\mathbf{y}(t)$ is related to $\mathbf{x}(t)$, the filter output is $\hat{\mathbf{x}}(t)$. The filter is described by Equations 9 and 10 with $\mathbf{f}[t, \hat{\mathbf{x}}, \mathbf{g}(\hat{\mathbf{x}})]$ replaced by $\mathbf{f}\{t, \hat{\mathbf{x}}, \mathbf{g}[\hat{\mathbf{x}}(t - \alpha_2)]\}$. Two papers have appeared on the subject of filtering systems with time delays. Kwakernaak (1967) extended the Kalman filter to linear systems with multiple time delays. Koivo and Stoller (1968) derived filter equations for a filter placed outside a control loop involving a pure time delay.

Since we have assumed that the control output is precisely known—i.e., $\mathbf{g}(\mathbf{x})$—this function can be directly used in place of $\mathbf{u}(t)$ in the *filter equations*. In practice, if the controller action cannot be represented precisely, the resulting $\mathbf{u}(t)$ can be measured and sent directly to the filter simply as a known function of time.

Control of a CSTR Subject to Random Inputs

We wish to consider the performance of a CSTR with proportional control on temperature in each of the schemes of Figures 1 to 4. Consider the dynamical equations of a CSTR with an exothermic first-order reaction and heat removal by a coil or jacket.

$$V(dc/ds) = q(c_o - c) - Vk_o e^{-E/RT} c \tag{15}$$

$$\rho V C_p(dT/ds) = qC_p(T_o - T) + (-\Delta H)Vk_o e^{-E/RT} c - D(T - T_c) \tag{16}$$

Defining the dimensionless variables,

$$\begin{aligned} t &= qs/V & \beta &= \ln(Vk_o/q) \\ \phi &= c/c_o & \gamma &= E\rho C_p/(-\Delta H)c_o R \\ \psi &= \rho C_p T/(-\Delta H)c_o & \theta &= D/\rho q C_p \end{aligned} \tag{17}$$

Equations 15 and 16 become

$$d\phi/dt = 1 - \phi - \exp[\beta - (\gamma/\psi)]\phi \tag{18}$$

$$d\psi/dt = \psi_o - \psi + \exp[\beta - (\gamma/\psi)]\phi - \theta(\psi - \psi_c) \tag{19}$$

If feedback proportional control on temperature is used to manipulate the flow rate of coolant, the dimensionless heat transfer coefficient, θ, can be expressed as (Aris and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Amundson, 1955a)

$$\theta = \begin{cases} \theta_s + k\psi_{cr} & \psi \geq \psi_s + \psi_{cr} \\ \theta_s + k(\psi - \psi_s) & |\psi - \psi_s| < \psi_{cr} \\ 0 & \psi \leq \psi_s - \psi_{cr} \end{cases} \quad (20)$$

where k is the proportional gain, θ_s the steady-state heat transfer coefficient, ψ_s the desired reactor temperature, and $k\psi_{cr}$ a constant corresponding to half range of the coolant flow valve. The object of control is to maintain the outlet temperature, ψ, at ψ_s.

As a low frequency inlet disturbance we will consider a step change in the inlet temperature ψ_0 at $t = 0$. High frequency fluctuations in inlet concentration and temperature enter the right-hand sides of Equations 18 and 19 additively. Thus, we add the random variables $\xi_1(t)$ and $\xi_2(t)$,

$$d\phi/dt = 1 - \phi - \exp[\beta - (\gamma/\psi)]\phi + \xi_1 \quad (21)$$

$$d\psi/dt = \psi_0 - \psi + \exp[\beta - (\gamma/\psi)]\phi - \theta(\psi - \psi_s) + \xi_2 \quad (22)$$

The measured output from the reactor is the temperature, ψ, which may, in general, have a noisy component $\eta(t)$,

$$y(t) = \psi(t) + \eta(t) \quad (23)$$

where θ in Equation 20 now depends on $y(t)$ rather than $\psi(t)$. If there exists a pure time delay of magnitude α_1 in the observation, Equation 23 is replaced by

$$y(t) = \psi(t - \alpha_1) + \eta(t) \quad (24)$$

For a pure time delay of magnitude α_2 in the controller, Equation 20 is replaced by

$$\theta = \begin{cases} \theta_s + k\psi_{cr} & y(t - \alpha_2) \geq \psi_s + \psi_{cr} \\ \theta_s + k[y(t - \alpha_2) - \psi_s] & |y(t - \alpha_2) - \psi_s| < \psi_{cr} \\ 0 & y(t - \alpha_2) \leq \psi_s - \psi_{cr} \end{cases} \quad (25)$$

The original steady state of the reactor corresponds to $\psi = \psi_s$ and $\theta = \theta_s$ (no control). The following parameters are used: $\theta_s = 1$, $\psi_s = 1.75$, $k\psi_{cr} = 1$, $\beta = 25$, $\gamma = 50$, $\psi_0(t < 0) = 1.75$. For these parameters there are three steady states for the CSTR with no control. The initial steady state was chosen as $\phi_s = 0.5$ and $\psi_s = 2.0$, the unstable steady state. The inlet temperature ψ_0 for $t > 0$ is taken as 1.85. For a particular value of the gain k the new steady state(s) can be computed. The difference of $\psi(t \to \infty)$ from ψ_s is the offset. In this study $k = 20$ was used, for which there are three steady states, the unstable one resulting in the smallest offset.

We wish to compare the performance of the CSTR in the schemes of Figures 1 to 4. Thus, it is necessary to simulate each of the situations depicted by means of computer experiments. In particular, the dynamical and measurement noise must be simulated by appropriate expressions. The response of the CSTR with no noise and no time delay can be obtained from the solution of Equations 18 to 20 with $\phi(0) = \phi_s$, $\psi(0) = \psi_s$. The response of the CSTR with no noise and a delay α_2 in the loop can be obtained from the solution of Equations 18, 19, and 25 with $y(t - \alpha_2)$ simply replaced by $\psi(t - \alpha_2)$. With no filter the location of the pure time delay in the loop is immaterial. These responses are referred to as the *deterministic* responses.

To simulate the noisy dynamics of Figure 1, Equations 20 to 22 are integrated with $\phi(0) = \phi_s$, $\psi(0) = \psi_s$, and

$$\xi_1(t) = A_1 \cos \omega_1 t$$
$$\xi_2(t) = A_2 \cos \omega_2 t \quad (26)$$

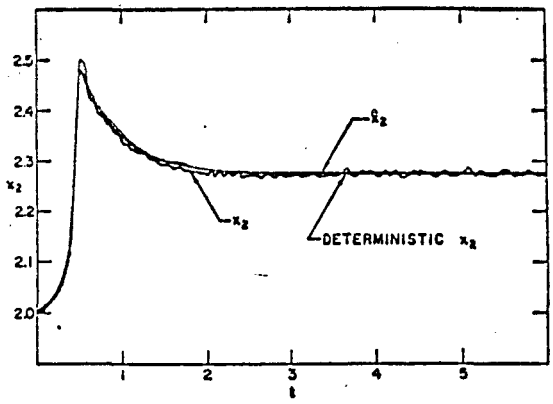

Figure 5. Transient response of CSTR and filter with controller off

where the values $A_1 = 1.5$, $A_2 = 1.0$, $\omega_1 = 20\pi$, and $\omega_2 = 40\pi$ are used. To produce the noisy observation, $\psi(t)$ from Equation 22 is used in Equation 23 with $\eta(t)$ as a normally distributed random variable with zero mean and variance of 0.1. To simulate the response when a pure time delay α_1 exists, Equations 21 and 22 are integrated with Equations 24 and 26. When the delay α_2 exists, Equations 21 and 22 are integrated with Equations 23, 25, and 26. These responses are referred to as the *unfiltered* responses.

Next we wish to simulate the response of the entire loop when a filter is placed after the observation. If we let $x^T = (x_1, x_2) = (\phi, \psi)$ and $\xi^T = (\xi_1, \xi_2)$, Equations 21 and 22 can be written in the general form of Equation 1, where $n = 2$, $m = 1$, and $r = 1$. The filter output is x_1, x_2, P_{11}, P_{12}, and P_{22} ($P_{21} = P_{12}$). Q is a scalar and R is taken as a diagonal matrix with elements R_{11} and R_{22}. Since the state equations are nonlinear, no direct statistical interpretation can be ascribed to Q and R. As mentioned previously, however, in the linear case Q and R are the covariances of η and ξ. So if we have some *a priori* knowledge as to these covariances, these values represent reasonable choices for Q and R even in the nonlinear case. From the results of an earlier study (Bellman *et al.*, 1966) it is apparent that the performance of the filtering equations depends significantly on the choices of $\mathbf{x}(0)$, $\mathbf{P}(0)$, Q, and R. In order to examine the convergence of the filter equations, Equations 21 to 23 were considered with $\theta = \theta_s$ (no control)—i.e., the pure transient response of the CSTR to a step change in ψ_0 in the presence of dynamical and measurement noise. Since at $t = 0$ we know that the system is at $(x_{1_s}, x_{2_s}) = (\phi_s, \psi_s)$, the most reasonable choice for $[x_1(0), x_2(0)]$ is (ϕ_s, ψ_s). Several cases were examined in which $\mathbf{P}(0)$, Q, and R were varied. One example is shown in Figure 5, where $Q = 1$, $R_{11} = 5$, $R_{22} = 10$, $P_{11_s} = 1$, $P_{12_s} = 1$, $P_{22_s} = 4$. The true and estimated values are almost identical over the entire time of integration. Other cases not shown converged more or less the same as in Figure 5; however, if Q is too small convergence is not obtained. These values are used in the remainder of the study.

Computational Simulation

First we consider the control schemes of Figures 1 and 3 (no time delays). The simulation of the scheme in Figure 1 has been described. For the filtered system (Figure 3) the state, observation, and filter (Equations 2, 8, 9, and 10) are solved simultaneously from $t = 0$. The outlet temperature

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

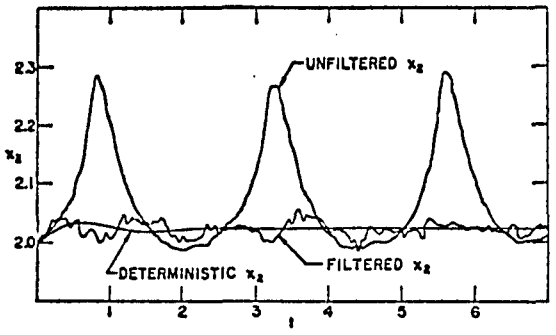

Figure 6. Comparison of filtered and unfiltered responses with no time delays

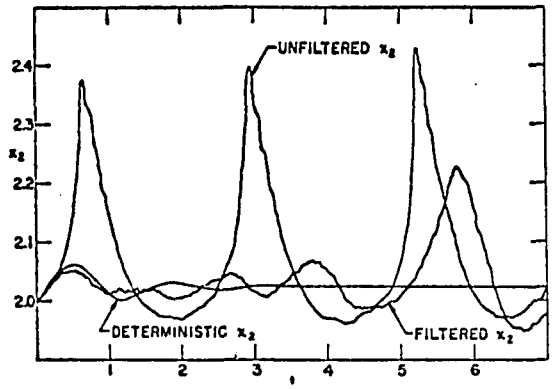

Figure 7. Comparison of filtered and unfiltered responses with $\alpha_1 = 0.1$

responses are shown in Figure 6. The deterministic $x_2(t)$ is the response of the reactor temperature with no noise. The unfiltered $x_2(t)$ is the response of the reactor temperature corresponding to Figure 1. The filtered $x_2(t)$ is the response corresponding to Figure 3. The comparison of interest is between the filtered and unfiltered $x_2(t)$. We see that performance has been substantially improved in the filtered case.

Next we consider the case of time delays (Figures 2 and 4). As noted, we treat α_1 and α_2 separately to compare the effect of the time delay location. The deterministic $x_2(t)$ is the response of the reactor with no noise. The unfiltered response for $\alpha_1 = 0$ and $\alpha_2 = 0$ is obtained by the simultaneous integration of Equations 21 and 22 with Equations 20 and 24 from $t = 0$. The filtered response is obtained by integration of Equations 3 and 11 to 13. The responses for $\alpha_1 = 0.1$ and $\alpha_2 = 0$ are shown in Figure 7. It appears that the unfiltered response is exhibiting unstable or limit cycle behavior while the filtered response is not. We discuss this matter subsequently. The estimation of $\alpha_1 = 0, \alpha_2 = 0.1$ is depicted in Figure 8. The deterministic response for $\alpha_1 = 0.1$ and $\alpha_2 = 0$ is obviously identical to that for $\alpha_1 = 0, \alpha_2 = 0.1$.

The location of the delay in the loop is seen to have some effect on the filtered response. When a pure time delay occurs in the observation, the combined effect of the delay and the observation noise causes larger oscillations in the response than when a delay of the same magnitude occurs in the controller. In the former case the filter produces estimates delayed by α_1, $\hat{\mathbf{x}}(t - \alpha)$. The responses for $\alpha_1 = 0$,

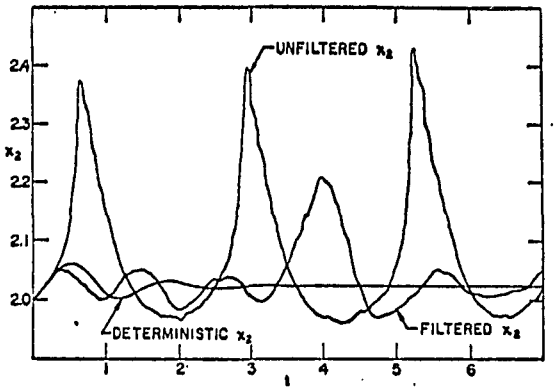

Figure 8. Comparison of filtered and unfiltered responses with $\alpha_2 = 0.1$

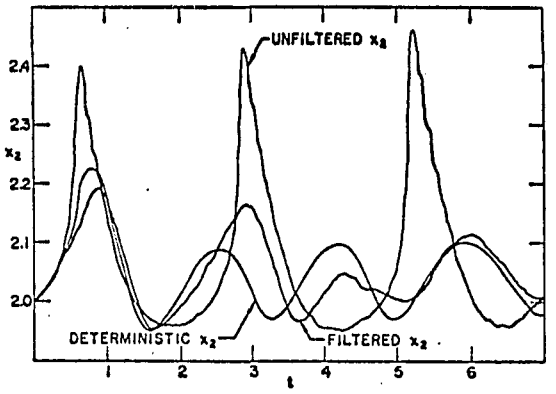

Figure 9. Comparison of filtered and unfiltered responses with $\alpha_2 = 0.15$

$\alpha_2 = 0.15$ are presented in Figure 9. Whereas in Figure 8 the deterministic x_2 experiences decaying oscillations, now with α_2 increased to 0.15, the deterministic x_2 undergoes sustained oscillations. It has been shown that limit cycle behavior is obtained for certain combinations of k and α in the deterministic system with proportional control (Seinfeld, 1969). The unfiltered x_2 exhibits the same oscillatory behavior as in Figure 8. A comparison of the deterministic and unfiltered x_2 in Figure 9 shows that that noise makes the oscillations more severe. The filtered x_2 is kept more closely to the deterministic x_2.

Effect of Noise on Control of CSTR

In each of the above cases, both dynamical and measurement noise has been considered. Dynamical noise enters the process as inputs and the differential equations as random forcing terms. The CSTR acts as a natural filter as long as the principal frequency band of the power spectrum of the noise is much greater than the characteristic frequency of the CSTR (the reciprocal of the time constant, q/V). Thus, in the absence of observational errors, the unfiltered reactor response with high frequency dynamical noise is not too different from the response with no noise at all. If the frequency of the dynamical noise approaches the characteristic frequency of the system, the dynamical noise affects the system like an additional disturbance in the input. The

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

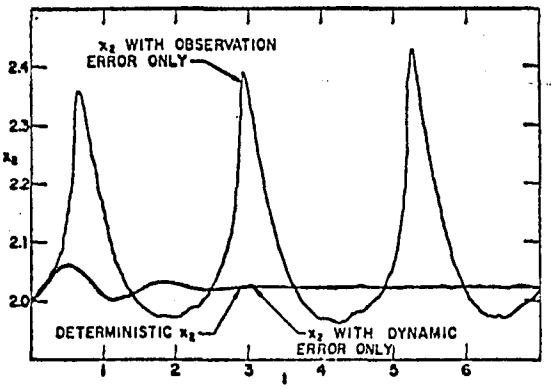

Figure 10. Effect of dynamical and observation noise on CSTR, $\alpha_2 = 0.1$

effects of dynamical and measurement noise are depicted in Figure 10 for the case of no filter and $\alpha_1 = 0$, $\alpha_2 = 0.1$. The deterministic $x_2(t)$ is the same as in Figure 8. The response x_2 for dynamical error only is seen to be close to the deterministic x_2, confirming the natural filtering characteristic of the CSTR. The response x_2 with observational error only is seen to be much more violent. The x_2 curve with both dynamical and measurement noise is the same as the *unfiltered* x_2 in Figure 8. It is obvious that the measurement noise is the key factor in stochastic control. In this example, the Gaussian measurement noise has caused the entire loop to enter a limit cycle, whereas in the absence of this noise the system is driven to the desired state. For other combinations of k and α_2, the noise could cause the system to go to one of the stable steady states. It is in a case of this type that filtering is of most usefulness.

Summary

The object of this work has been to present and examine schemes for the control of noisy nonlinear dynamical systems. The addition of a filter significantly improves performance when the amplitude of noise is large. The actual choice of whether or not to include a filter depends on the trade-off between the improved performance of regulation and the cost of computer use.

If pure time delays become large, one might try to compensate by placing a predictor after the filter. For example, if the filter output is $\hat{\mathbf{x}}(t - \alpha_1)$, the predictor would integrate the deterministic system equations from $t = t - \alpha_1$ to $t = t$ to produce $\mathbf{x}(t)$ at each instant. This scheme was actually tried in this study. The increased performance was not commensurate with the additional computing requirements, and for moderate time lags a predictor is probably unnecessary.

Nomenclature

A_1, A_2 = error amplitudes
c = concentration, lb. moles/cu. ft.
C_p = specific heat of reaction mixture, B.t.u./lb.-°F.
D = over-all heat transfer coefficient, B.t.u./min.-°F.
E = activation energy, B.t.u./lb. mole
$\mathbf{f}$ = n-dimensional vector function
$\mathbf{g}$ = r-dimensional vector function
$\mathbf{h}$ = m-dimensional vector function
I = performance index
k = proportional gain
$\mathbf{P}$ = $n \times n$ covariance matrix
q = flow rate, cu. ft./min.
$\mathbf{Q}$ = $m \times m$ weighting matrix
R = gas constant, B.t.u./lb. mole-°R.
$\mathbf{R}$ = $n \times n$ weighting matrix
s = time variable, min.
t = time variable
T = temperature °R.
$\mathbf{u}$ = r-dimensional control vector
V = volume of reactor, cu. ft.
$\mathbf{w}$ = n-dimensional vector function
$\mathbf{x}$ = n-dimensional state vector
$\mathbf{y}$ = m-dimensional output vector

Greek Letters

$\alpha, \alpha_1, \alpha_2$ = time lags
β = constant defined in Equation 13
γ = constant defined in Equation 13
ξ = n-dimensional noise vector
η = m-dimensional noise vector
θ = dimensionless heat transfer coefficient
ρ = fluid density
τ = time
ϕ = dimensionless concentration
ψ = dimensionless temperature
ω_1, ω_2 = noise frequencies

Subscripts

c = coolant fluid
cr = coolant rate
o = initial or inlet
s = steady state

Superscript

$\hat{}$ = estimated value

Literature Cited

Aoki, M., "Optimization of Stochastic Systems," Academic Press, New York, 1967.
Aris, R., Amundson, N. R., *Chem. Eng. Sci.* **7**, 121 (1958a).
Aris, R., Amundson, N. R., *Chem. Eng. Sci.* **9**, 250 (1958b).
Athans, M., Wishner, R. P., Bertolini, A., Joint Automatic Control Conference, Session 11, Ann Arbor, Mich., 1968.
Bellman, R. E., Kagiwada, H. H., Kalaba, R. E., Sridhar, R., *J. Astronaut. Sci.*, 13(3), 110 (1966).
Detchmendy, D. M., Sridhar, R., *J. Basic Eng.* **88D**, 362 (1966).
Gavalas, G. R., Seinfeld, J. H., Sridhar, R., *J. Basic Eng.*, submitted for publication, 1969.
Kalman, R. E., Bucy, R. S., *J. Basic Eng.* **83D**, 95 (1961).
Koivo, A. J., Stoller, R. L., Joint Automatic Control Conference, Ann Arbor, Mich., p. 116, 1968.
Kushner, H. J., *J. Math. Anal. Appl.* **11**, 78 (1965).
Kwakernaak, H., *I.E.E.E. Trans. Automatic Control* **AC-12**,(2), 169 (1967).
Meditch, J. S., Boeing Research Laboratories, Doc. D1-82-0693, April 1968.
Newton, G. C., Gould, L. A., Kaiser, J. F., "Analytical Design of Linear Feedback Controls," Wiley, New York, 1957.
Seinfeld, J. H., *Intern. J. Control*, in press, 1969.
Solodovnikov, V. V., "Introduction to the Statistical Analysis of Automatic Control Systems," Dover, New York, 1960.
Sworder, D. D., *Intern. J. Control* **6**(2), 179 (1967).

Received for review November 20, 1968
Accepted January 16, 1969

Work supported in part by National Science Foundation Grant GK-3342.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX II-B

The effect of the measurement error on the proportional constant of the CSTR system in Appendix II-A is illustrated by means of the describing function approach. The dynamic equation of the CSTR with measurement error only can be written, from equations (21) and (22) in Appendix II-A, as

$$\frac{dx_1}{dt} = 1 - \phi_s - x_1 - \exp(\beta - \frac{\gamma}{\psi_s + x_2})(\phi_s + x_1) \tag{1}$$

$$\frac{dx_2}{dt} = \psi_o - \psi_s - x_2 + \exp(\beta - \frac{\gamma}{\psi_s + x_2})(\phi_s + x_1) - \theta(\psi_s - \psi_c + x_2) \tag{2}$$

where

$$x_1 = \phi - \phi_s \tag{3}$$

$$x_2 = \psi - \psi_s \tag{4}$$

$$y = x_2 + \eta \tag{5}$$

$$\theta = \begin{cases} 2 & , \quad y \geq 0.05 \\ 1 + ky & , \quad -0.05 < y < 0.05 \\ 0 & , \quad y \leq -0.05 \end{cases} \tag{6}$$

where η is the measurement error. The corresponding dynamic response gives the similar results as those shown in Figure 6 in which case $\xi_1 = \xi_2 \neq 0$, because the dynamic noise effect is negligible as shown in Figure 10. For the evaluation of the effective gain, the following assumptions are made:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(1) $\eta = N(0,0.01)$

(2) The system can be separated into two parts; one is the nonlinear controller, the other is the remaining linear part as shown in Figure 11.

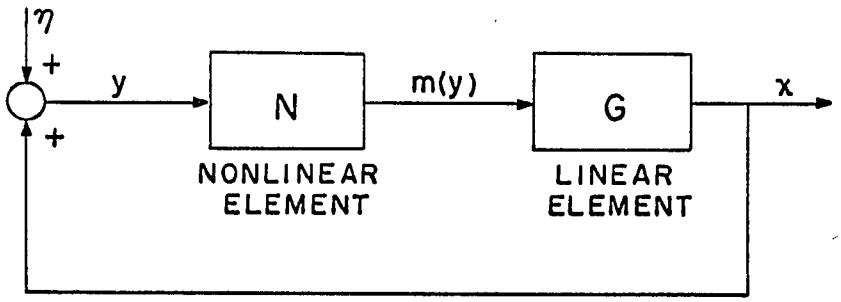

Fig. 11. Division of the system into linear and nonlinear portions

The assumption (2) implies that the present approach is applicable to the system linearized around a steady state. The input signal to the controller is the noisy observation y, where the system output x_2 can be considered as a deterministic quantity, since the linear part behaves like a low pass filter. For a random input y the controller output $m(y)$ is assumed

$$m(y) = 1 + k_{eq}\, y + f_d(y) \tag{7}$$

The motivation of the above equation is that the expression $\theta = 1 + ky$ is valid for $-0.05 < y < 0.05$. Thus around $x_2 = 0$ the above approximation covers almost the whole range of η without serious error caused by the saturation points. Here k_{eq} is so determined as to minimize $E\{f_d^2(y)\}$, where

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$E\{f_d^2(y)\} = E\{(m(y) - 1 - k_{eq}y)^2\}$$

$$= E\{m^2(y)\} - 2\,k_{eq}\,E\{y\,m(y)\} + k_{eq}^2\,E\{y^2\} + 2k_{eq}\,E\{y\} - 2E\{m(y)\} + 1 \qquad (8)$$

Hence $\dfrac{\partial E\{f_d^2(y)\}}{\partial k_{eq}} = 0$ gives

$$k_{eq} = \frac{E\{y\,m(y)\} - E\{y\}}{E\{y^2\}} \qquad (9)$$

Using the assumption (1) and $y = x_2 + \eta$, we obtain

$$k_{eq} = 0.384k \qquad (10)$$

For $k = 20$, we have $k_{eq} = 7.7$. In other words, if $x_2 = 0$, the effective proportional constant is only 38.4% of the original value because of measurement noise. As mentioned by Aris and Amundson in Appendix II-A, k should be greater than 9.1 to avoid the limit cycle for the given system. Therefore, we can expect a sustained oscillation. However, this calculation is not enough to justify the limit cycle behavior of Figure 6 or Figure 10, since we do not know the overall effective k for given operation period and for the whole range of the value. The numerical results for the cases of Figure 6 and Figure 10 up to $t_f = 21$ shows sustained oscillation for both cases.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter III

OPTIMAL LEAST SQUARE FILTERING AND INTERPOLATION IN DISTRIBUTED PARAMETER SYSTEMS

1. Introduction

The estimation of states and parameters in noisy dynamical systems has important applications in identification, optimal and adaptive control. While this problem has been studied extensively for systems described by O.D.E.'s, relatively little has appeared for distributed systems. To derive the filtering and smoothing equations, two different approaches, namely "Direct Method" and "Indirect Method" described in Chapter I have been applied. Yet no previous studies have considered the recursive estimation of constant parameters in the system and boundary conditions and the estimation of states in P.D.E.'s when the boundary conditions contain dynamical noise.

In this study we derive least square filtering and interpolation algorithms for states and parameters in nonlinear distributed systems with unknown additive volume, boundary and observation noise, including volume and boundary inputs governed by stochastic O.D.E.'s. The optimal control approach suggested by Detchmendy and Sridhar[11] for lumped systems is applied with the extended invariant imbedding technique. The solution procedures can be summarized as follows:

(1) Formulate the stochastic minimization problem with the least square error functional.

(2) Reformulate the stochastic minimization problem as a deterministic optimal control problem.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(3) Apply calculus of variation to obtain necessary conditions for optimality (two point boundary value problem).

(4) Apply the invariant imbedding technique to convert the two point boundary value problem into the initial value problem (Hamilton-Jacobi type[53]).

(5) Solve the resulting Hamilton-Jacobi type equations to obtain the estimation equations.

In the present study the one-dimensional case is considered, but the present approach can be directly extended to any dimensions. Also, the recursive estimation of constant parameters appearing in the system and boundary conditions can be handled readily.

In Appendix III-A the indirect approach used by Pell and Aris[52] for a linear system is extended to nonlinear cases without additive boundary noises to derive the nonlinear filter. In the derivation the finite differential difference approximation along the spatial axis is used to convert P.D.E. systems into O.D.E. systems to utilize the known results for lumped systems. Then a limiting operation is performed to obtain the nonlinear filter in P.D.E. form. Also, simplification of the covariance equation is shown in a numerical example.

2. Problem Statement

We consider the class of systems governed by the nonlinear partial differential equation ,

$$x_t(r,t) = f(r,t,x,x_r,x_{rr},a(t)) + \xi_1(r,t) \tag{2.1}$$

defined for $t \geq 0$ on the normalized domain (0,1), where $x(r,t)$ is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

n-vector state and $\xi_1(r,t)$ is an unknown n-vector volume disturbance. x_t, x_r, x_{rr} denote $\partial x/\partial t$, $\partial x/\partial r$, and $\partial^2 x/\partial r^2$, respectively. The ℓ_1-vector input $a(t)$ is governed by

$$\frac{da}{dt} = A(t,a(t)) + \xi_2(t) \tag{2.2}$$

and the boundary conditions of the system are given in the s-vector ($s \leq n$) functions,

$$g_o(t,x,x_r) + \xi_3(t) = 0 \quad , \quad r = 0 \tag{2.3}$$

$$g_1(t,x,x_r,b(t)) + \xi_4(t) = 0 \quad , \quad r = 1 \tag{2.4}$$

with the ℓ_2-vector input $b(t)$ governed by

$$\frac{db}{dt} = B(t,b(t)) + \xi_5(t) \tag{2.5}$$

where $\xi_i(t)$, $i = 2,\cdots,5$, are independent zero-mean random processes with unknown statistical characteristics. We assume that in the absence of noise, $\xi_i = 0$, $i = 1,\cdots,5$ the problem (2.1) - (2.5) is well posed. Observations of the system consist of the m-vector $y(r,t)$, related to the state by

$$y(r,t) = h(r,t,x(r,t)) + \eta(r,t) \tag{2.6}$$

where $\eta(r,t)$ is an m-vector of unknown measurement noise.

Based on the observations $y(r,t)$ in the interval $0 \leq t \leq T$ and $r\varepsilon[0,1]$, it is required to estimate $x(r,t)$, $a(t)$, and $b(t)$ at some time t_1. If $t_1 = T$, this is the filtering estimate, and if $t_o \leq t_1 \leq T$, it is the interpolating estimate. For any admissible

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

estimates $x(r,t)$, $a(t)$, and $b(t)$, $0 \le t \le T$, which are continuous with piecewise continuous derivatives, the criterion of estimation is defined by the least square error functional

$$I = \int_0^T \Bigg\{ \int_0^1 \int_0^1 (y(r,t) - h(r,t,x))^T Q(r,s,t)(y(s,t) - h(s,t,x)) dr ds$$

$$+ \int_0^1 \int_0^1 \Big(x_t(r,t) - f(r,t,x,x_r,x_{rr},a(t)) \Big)^T R_1(r,s,t) \Big(x_t(s,t)$$

$$- f(s,t,x,x_s,x_{ss},a(t)) \Big) dr\, ds + (y(0,t) - h(0,t,x))^T Q(0,0,t)(y(0,t)$$

$$- h(0,t,x)) + g_o(t,x,x_r)^T R_3(t)\, g_o(t,x,x_r) + g_1(t,x,x_r,b)^T$$

$$R_4(t)\, g_1(t,x,x_r,b) + (\dot{a}(t) - A(t,a))^T R_2(t)(\dot{a}(t) - A(t,a))$$

$$+ (\dot{b}(t) - B(t,b))^T R_5(t)(\dot{b}(t) - B(t,b)) \Bigg\} dt \qquad (2.7)$$

The weighting matrices $Q(r,s,t)$, $R_1(r,s,t)$, $R_i(t)$, $i = 2,\cdots,5$ are continuous with respect to their arguments and positive definite. Also $Q(r,s,t)$ and $R_1(r,s,t)$ are assumed symmetric with respect to r and s . The necessary positive-definiteness of the above weighting matrices in a quadratic error criterion of the form (2.7) has been shown by Russell and Lukes[56] for the existence of an optimal control. In addition, if

$$\int_0^1 R_1(r,s,t)\, u(s,t)\, ds = v(r,t) \qquad (2.8)$$

with an inverse operation

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\int_0^1 \tilde{R}_1(r,s,t)\, v(s,t)\, ds = u(r,t) \tag{2.9}$$

then[67]

$$\int_0^1 R_1(r,s,t)\, \tilde{R}_1(s,\rho,t)\, ds = \delta(r-\rho) \tag{2.10}$$

In what follows we denote $\tilde{R}_1(r,s,t)$ by $R_1^{-1}(r,s,t)$.

If we desire to estimate constant parameter vectors a and b appearing in the volume and boundary conditions, it is only necessary to let $A(t,a) = 0$ and $B(t,b) = 0$ in (2.2) and (2.5). It will be seen this is the proper way of treating the recursive estimation of constant parameters in partial differential equations, i.e., through the definition of auxiliary ordinary differential equations of the form (2.2) and (2.5).

3. Optimal Least Square Filtering

The filtering problem is to determine $x(r,T)$, $a(T)$, and $b(T)$ such that the functional (2.7) is minimized. We reformulate this problem as an optimal control problem, an approach with the advantage of not requiring statistical assumptions on the disturbances[11]. We desire to minimize I with respect to $x(r,t)$, $u_1(r,t)$, $u_i(t)$, $i = 2,\cdots,5$

$$I = \int_0^T \left\{ \int_0^1 \int_0^1 (y(r,t) - h(r,t,x))^T\, Q(r,s,t)(y(s,t)-h(s,t,x))\, drds \right.$$

$$\left. + \int_0^1 \int_0^1 u_1(r,t)^T\, R_1(r,s,t)\, u_1(s,t)\, drds + (y(0,t) - h(0,t,x))^T \right.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$Q(0,0,t)(y(0,t) - h(0,t,x)) + \sum_{i=2}^{5} u_i^T(t)\, R_i(t)\, u_i(t) \Big\} \, dt \qquad (3.1)$$

subject to the constraints

$$x_t(r,t) = f(r,t,x,x_r,x_{rr},a) + u_1(r,t) \qquad (3.2)$$

$$\frac{da}{dt} = A(t,a) + u_2(t) \qquad (3.3)$$

$$g_o(t,x,x_r) + u_3(t) = 0 \,, \quad r = 0 \qquad (3.4)$$

$$g_1(t,x,x_r,b) + u_4(t) = 0 \,, \quad r = 1 \qquad (3.5)$$

$$\frac{db}{dt} = B(t,b) + u_5(t) \qquad (3.6)$$

Note that the initial and terminal states are free, since we will not in general know the initial states $x(r,0)$, $a(0)$ and $b(0)$. The necessary conditions for optimality for (3.1) - (3.6) can be obtained from the Euler equations and transversality conditions, and are

$$x_t(r,t) = f(r,t,x,x_r,x_{rr},a) - \frac{1}{2} \int_0^1 R_1^{-1}(r,s,t)\, \lambda(s,t)\, ds \qquad (3.7)$$

$$g_o(t,x,x_r) - \frac{1}{2} R_3^{-1}(t)\, \mu_o(t) = 0 \qquad (3.8)$$

$$g_1(t,x,x_r,b) - \frac{1}{2} R_4^{-1}(t)\, \mu_1(t) = 0 \qquad (3.9)$$

$$\frac{da}{dt} = A(t,a) - \frac{1}{2} R_2^{-1}(t)\, \tau(t) \qquad (3.10)$$

$$\frac{db}{dt} = B(t,b) - \frac{1}{2} R_5^{-1}(t)\, \sigma(t) \qquad (3.11)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\lambda_t(r,t) = 2\int_0^1 h_x^T(r,t,x)\, Q(r,s,t)(y(s,t) - h(s,t,x))\, ds$$

$$- f_x^T\lambda + [f_{x_r}^T \lambda]_r - [f_{x_{rr}}^T \lambda]_{rr} \tag{3.12}$$

$$\frac{d\tau}{dt} = -\int_0^1 f_a^T\lambda(r,t)\, dr - A_a^T\tau \tag{3.13}$$

$$\frac{d\sigma}{dt} = - B_b^T\sigma - g_{1_b}^T \mu_1(t) \tag{3.14}$$

$$\lambda(r,0) = \lambda(r,T) = 0 \tag{3.15}$$

$$\tau(0) = \tau(T) = 0 \tag{3.16}$$

$$\sigma(0) = \sigma(T) = 0 \tag{3.17}$$

$$g_{o_x}^T \mu_o - f_{x_r}^T \lambda + [f_{x_{rr}}^T \lambda]_r - 2h_x^T Q(0,0,t)(y(0,t) - h(0,t,x)) = 0 \,, \quad r = 0 \tag{3.18}$$

$$g_{o_{x_r}}^T \mu_o - f_{x_{rr}}^T \lambda = 0 \,, \quad r = 0 \tag{3.19}$$

$$g_{1_x}^T \mu_1 + f_{x_r}^T \lambda - [f_{x_{rr}}^T \lambda]_r = 0 \,, \quad r = 1 \tag{3.20}$$

$$g_{1_{x_r}}^T \mu_1 + f_{x_{rr}}^T \lambda = 0 \,, \quad r = 1 \tag{3.21}$$

where $\lambda(r,t)$, $\mu_o(t)$, $\mu_1(t)$, $\tau(t)$, and $\sigma(t)$ are Lagrange multipliers, or adjoint variables. $\mu_o(t)$ and $\mu_1(t)$ can be expressed in terms of $\lambda(0,t)$ and $\lambda(1,t)$. If $g_{0_{x_r}} \neq 0$ and $g_{1_{x_r}} \neq 0$,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\mu_o(t) = g_{o_{x_r}}^{-1} f_{x_{rr}}^T \lambda \quad , \quad r = 0 \tag{3.22}$$

$$\mu_1(t) = -g_{1_{x_r}}^{-1} f_{x_{rr}}^T \lambda \quad , \quad r = 1 \tag{3.23}$$

where $g_{o_{x_r}}^{-1}$ and $g_{1_{x_r}}^{-1}$ can be interpreted as the left inverse when $s \neq n$

$$\begin{aligned} g_{o_{x_r}}^{-1} &= \left(g_{o_{x_r}} g_{o_{x_r}}^T\right)^{-1} g_{o_{x_r}} \\ g_{1_{x_r}}^{-1} &= \left(g_{1_{x_r}} g_{1_{x_r}}^T\right)^{-1} g_{1_{x_r}} \end{aligned} \tag{3.24}$$

If $g_{o_{x_r}} = g_{1_{x_r}} = 0$

$$\mu_o(t) = g_{o_x}^{-1} \left\{-[f_{x_{rr}}^T \lambda]_r + 2h_x^T Q(0,0,t)(y(0,t) - h(0,t,x))\right\}, r = 0 \tag{3.25}$$

$$\mu_1(t) = g_{1_x}^{-1} [f_{x_{rr}}^T \lambda]_r \quad , \quad r = 1 \tag{3.26}$$

In the remainder of the study we assume $g_{o_{x_r}} \neq 0$ and $g_{1_{x_r}} \neq 0$. Thus the necessary conditions are given by (3.7) - (3.17), (3.22), (3.23) and

$$g_{o_x}^T g_{o_{x_r}}^{-1} f_{x_{rr}}^T \lambda - f_{x_r}^T \lambda + [f_{x_{rr}}^T \lambda]_r - 2h_x^T Q(0,0,t)(y(0,t) - h(0,t,x)) = 0 \,, \quad r = 0 \tag{3.27}$$

$$g_{1_x}^T g_{1_{x_r}}^{-1} f_{x_{rr}}^T \lambda - f_{x_r}^T \lambda + [f_{x_{rr}}^T \lambda]_r = 0 \,, \quad r = 1 \tag{3.28}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The necessary conditions constitute a two-point boundary value problem, the solution of which is the optimal smoothing estimates $x(r,t)$, $a(t)$, and $b(t)$. Initial and final conditions ($t = 0$ and $t = T$) are given for all adjoint variables, whereas $x(r,0)$, $x(r,T)$, $a(0)$, $a(T)$, $b(0)$ and $b(T)$ are free. For filtering the solution of the nonlinear two-point boundary value problem (3.7)- (3.17), (3.22), (3.23), (3.27) and (3.28) is desired for all $T \geq 0$. Thus, it is necessary to convert the two point boundary value problem into an initial value problem with T as an independent variable.

If the original optimal control problem is well-posed, there exists a unique $x(r,T)$, $a(T)$ and $b(T)$ when the final conditions $\lambda(r,T) = \tau(T) = \sigma(T) = 0$ are satisfied. Let us consider a more general class of problems, namely those in which $\lambda(r,T) = c^{(\lambda)}(r)$, $\tau(T) = c^{(\tau)}$ and $\sigma(T) = c^{(\sigma)}$. The solution to the general class of problems can be denoted

$$x(r,T) = \psi(r,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{3.29}$$

$$a(T) = \psi^{(a)}(T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{3.30}$$

$$b(T) = \psi^{(b)}(T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{3.31}$$

The solution we desire is $\psi(r,T,0,0,0)$, $\psi^{(a)}(T,0,0,0)$ and $\psi^{(b)}(T,0,0,0)$.

Our objective is to determine the initial value problem governing ψ, $\psi^{(a)}$ and $\psi^{(b)}$. The technique for converting a boundary value problem into an initial value problem by imbedding the desired

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

problem in a more general class of problems is termed invariant imbedding and has received considerable attention for ordinary differential equations. We will employ this technique on the present problem.

Let us represent (3.7), (3.10), (3.11), (3.12), (3.13) and (3.14) by

$$x_t = \alpha(r,t,\lambda,x,a) \tag{3.32}$$

$$\frac{da}{dt} = \rho(t,a,\tau) \tag{3.33}$$

$$\frac{db}{dt} = \eta(t,b,\sigma) \tag{3.34}$$

$$\lambda_t = \beta(r,t,\lambda,x,a) \tag{3.35}$$

$$\frac{d\tau}{dt} = \gamma(t,x,\lambda,a,\tau) \tag{3.36}$$

$$\frac{d\sigma}{dt} = \theta(t,x,\lambda,b,\sigma) \tag{3.37}$$

For a final time $T + \Delta$ we can write

$$\begin{aligned}\psi(r,T+\Delta,\, c^{(\lambda)}(r) + \Delta c^{(\lambda)}(r),\, c^{(\tau)} + \Delta c^{(\tau)},\, c^{(\sigma)} + \Delta c^{(\sigma)}) \\ = \psi(r,T,c^{(\lambda)}(r),\, c^{(\tau)},\, c^{(\sigma)}) + \psi_T\Delta + \int_0^1 \frac{\delta\psi}{\delta c^{(\lambda)}(r)}\,\Delta c^{(\lambda)}(r)\,dr \\ + \psi_c(\tau)\,\Delta c^{(\tau)} + \psi_c(\sigma)\,\Delta c^{(\sigma)} + 0(\Delta^2)\end{aligned} \tag{3.38}$$

where $\frac{\delta\psi}{\delta c^{(\lambda)}(r)}$ is a functional derivative[70]. We also can write

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\psi(r,T+\Delta,\ c^{(\lambda)}(r)+\Delta c^{(\lambda)}(r),\ c^{(\tau)}+\Delta c^{(\tau)},\ c^{(\sigma)}+\Delta c^{(\sigma)})$$

$$= x(r,T)+\alpha(r,T,c^{(\lambda)}(r),\ \psi(r,T,c^{(\lambda)}(r),\ c^{(\tau)},c^{(\sigma)}),$$

$$\psi^{(a)}(T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}))\Delta+0(\Delta^2) \tag{3.39}$$

In addition, we let

$$\Delta c^{(\lambda)}(r)=\beta(r,T,c^{(\lambda)}(r),\ \psi(r,T,c^{(\lambda)},\ c^{(\tau)},\ c^{(\sigma)}),\psi^{(a)}(T,c^{(\lambda)}(r),$$

$$c^{(\tau)},cc^{(\sigma)}))\Delta \tag{3.40}$$

$$\Delta c^{(\tau)}=\gamma(T,\ \psi(r,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),\ c^{(\lambda)}(r),\ \psi^{(a)}(T,c^{(\lambda)}(r),$$

$$c^{(\tau)},\ c^{(\sigma)}),\ c^{(\tau)})\Delta \tag{3.41}$$

$$\Delta c^{(\sigma)}=\theta(T,\ \psi(r,T,c^{(\lambda)},\ c^{(\sigma)}),\ c^{(\lambda)}(r),\ \psi^{(b)}(T,c^{(\lambda)}(r),\ c^{(\tau)},$$

$$c^{(\sigma)}),\ c^{(\sigma)})\Delta \tag{3.42}$$

Combining (3.38) and (3.39) and taking the limit $\Delta \to 0$, we obtain the Hamilton-Jacobi type equation

$$\psi_T+\int_0^1 \frac{\delta\psi}{\delta c^{(\lambda)}(r)}\,\beta(r,T,c^{(\lambda)}(r),\psi,\psi^{(a)})\,dr+\psi_{c^{(\tau)}}\,\gamma(T,\psi,c^{(\lambda)}(r),$$

$$\psi^{(a)},\ c^{(\tau)})+\psi_{c^{(\sigma)}}\,\theta(T,\psi,c^{(\lambda)}(r),\ \psi^{(b)},\ c^{(\sigma)}) \tag{3.43}$$

$$=\alpha(r,T,c^{(\lambda)}(r),\psi,\psi^{(a)})$$

Similarly, we can obtain

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\psi_T^{(a)} + \int_0^1 \frac{\delta\psi^{(a)}}{\delta c^{(\lambda)}(r)} \beta(r,T,c^{(\lambda)}(r),\psi,\psi^{(a)})\, dr + \psi^{(a)}_{c^{(\tau)}}\, \gamma(T,\psi,c^{(\lambda)}(r),$$

$$\psi^{(a)},\, c^{(\tau)}) + \psi^{(a)}_{c^{(\sigma)}}\, \theta(T,\psi,c^{(\lambda)}(r),\, \psi^{(b)},\, c^{(\sigma)}) = \rho(T,c^{(\tau)},\psi^{(a)}) \tag{3.44}$$

and

$$\psi_T^{(b)} + \int_0^1 \frac{\delta\psi^{(b)}}{\delta c^{(\lambda)}(r)} \beta(r,T,c^{(\lambda)}(r),\psi,\psi^{(a)})dr + \psi^{(b)}_{c^{(\tau)}}\, \gamma(T,\psi,c^{(\lambda)}(r),$$

$$\psi^{(a)},c^{(\tau)}) + \psi^{(b)}_{c^{(\sigma)}}\, \theta(T,\psi,c^{(\lambda)}(r),\, \psi^{(b)},\, c^{(\sigma)}) = \eta(T,\psi^{(b)},\, c^{(\sigma)}) \tag{3.45}$$

The desired initial value problems for ψ, $\psi^{(a)}$ and $\psi^{(b)}$ are given by (3.43) - (3.45). Let us assume solutions of the form

$$\psi(r,T,c^{(\lambda)}(r),c^{(\tau)},\, c^{(\sigma)}) = \hat{x}(r,T) - \frac{1}{2}\int_0^1 P^{(vv)}(r,s,T)c^{(\lambda)}(s)ds$$

$$- \frac{1}{2} P^{(va)}(r,T)\, c^{(\tau)} - \frac{1}{2} P^{(vb)}(r,T)\, c^{(\sigma)} \tag{3.46}$$

$$\psi^{(a)}(T,c^{(\lambda)}(r),\, c^{(\tau)},\, c^{(\sigma)}) = \hat{a}(T) - \frac{1}{2}\int_0^1 P^{(av)}(s,T)\, c^{(\lambda)}(s)ds$$

$$- \frac{1}{2} P^{(aa)}(T)\, c^{(\tau)} - \frac{1}{2} P^{(ab)}(T)\, c^{(\sigma)} \tag{3.47}$$

$$\psi^{(b)}(T,c^{(\lambda)}(r),\, c^{(\tau)},\, c^{(\sigma)}) = \hat{b}(T) - \frac{1}{2}\int_0^1 P^{(bv)}(s,T)\, c^{(\lambda)}(s)ds$$

$$- \frac{1}{2} P^{(ba)}(T)\, c^{(\tau)} - \frac{1}{2} P^{(bb)}(T)\, c^{(\sigma)} \tag{3.48}$$

When $c^{(\lambda)}(r) = c^{(\tau)} = c^{(\sigma)} = 0$, the assumed solutions reduce to the optimal estimates, $\hat{x}(r,T)$, $\hat{a}(T)$, and $\hat{b}(T)$. Thus (3.46) - (3.48) can

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be viewed as first order linearizations about the optimal estimates in the deviations $c^{(\lambda)}(r)$, $c^{(\tau)}$, and $c^{(\sigma)}$. This type of assumed solution to the Hamilton-Jacobi type equations was used in the lumped parameter case[11]. In the linear white noise case (3.46) - (3.48) yield the exact solutions of (3.43) - (3.45) and

$$P^{(vv)}(r,s,T) = E\{(x(r,T) - \hat{x}(r,T))(x(s,T) - \hat{x}(s,T))^T\}$$

$$P^{(va)}(r,T) = E\{(x(r,T) - \hat{x}(r,T))(a(T) - \hat{a}(T))^T\}$$

$$P^{(vb)}(r,T) = E\{(x(r,T) - \hat{x}(r,T))(b(T) - \hat{b}(T))^T\}$$

$$P^{(aa)}(T) = E\{(a(T) - \hat{a}(T))(a(T) - \hat{a}(T)^T\}$$

$$P^{(ab)}(T) = E\{(a(T) - \hat{a}(T))(b(T) - \hat{b}(T))^T\}$$

$$P^{(bb)}(T) = E\{(b(T) - \hat{b}(T))(b(T) - \hat{b}(T))^T\}$$

Thus

$$P^{(va)} = P^{(av)^T}, \quad P^{(vb)} = P^{(bv)^T}, \quad P^{(ab)} = P^{(ba)^T}$$

$$P^{(vv)}(r,s,T) = P^{(vv)}(s,r,T)^T \tag{3.50}$$

In the nonlinear case these functions do not have a direct statistical interpretation. The equations governing these functions are determined by substituting (3.46) - (3.48) into (3.43) - (3.45), linearizing each of the nonlinear terms about $\hat{x}(r,T)$, $\hat{a}(T)$, and $\hat{b}(T)$ up to first order in $c^{(\lambda)}$, $c^{(\tau)}$, and $c^{(\sigma)}$, and equating coefficients of terms of like order in $c^{(\lambda)}$, $c^{(\tau)}$ and $c^{(\sigma)}$. It is this linearization that enables the explicit determination of the governing differential

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

equations for $\hat{x}(r,T)$, $\hat{a}(T)$, $\hat{b}(T)$ and all the P functions. Substituting (3.46) - (3.48) into (3.43) and linearizing to first order we obtain

$$\hat{x}_T(r,T) - f(r,T,\hat{x},\hat{x}_r,\hat{x}_{rr},\hat{a}) - \int_0^1 \int_0^1 P^{(vv)}(r,s,T) h_x^T(s,T,\hat{x})\, Q(s,\zeta,\hat{x})$$

$$(y(\zeta,T) - h(\zeta,T,\hat{x}))ds\, d\zeta - \frac{1}{2}\int_0^1 \Omega^{(\lambda)}(r,sT)c^{(\lambda)}(s)ds - \frac{1}{2}\Omega^{(\tau)}(r,T)c^{(\tau)}$$

$$- \frac{1}{2}\Omega^{(\sigma)}(r,T)c^{(\sigma)} - \frac{1}{2}\{P^{(vv)}(r,s,T)\, f_{x_s}^T\, c^{(\lambda)}(s)$$

$$- P^{(vv)}(r,s,T)[f_{x_{ss}}^T c^{(\lambda)}(s)]_s + P_s^{(vv)}(r,s,T)\, f_{x_{ss}}^T c^{(\lambda)}(s)\}\Big|_{s=0}^{s=1}$$

$$+ O(c^{(\lambda)^2}, c^{(\tau)^2}, c^{(\sigma)^2} \qquad (3.51)$$

where

$$\Omega^{(\lambda)}(r,s,T) = P_T^{(vv)}(r,s,T) - \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)$$

$$P^{(vv)}(\nu,s,T)d\zeta\, d\nu - P^{(vv)}(r,s,T)\, \hat{f}_x^T(s) - P_s^{(vv)}(r,s,T)\, \hat{f}_{x_s}^T(s)$$

$$- P_{ss}^{(vv)}(r,s,T)\, \hat{f}_{x_{ss}}^T(s) - \hat{f}_x(r)\, P^{(vv)}(r,s,T) - \hat{f}_{x_r}(r)\, P_r^{(vv)}(r,s,T)$$

$$- \hat{f}_{x_{rr}}(r)\, P_{rr}^{(vv)}(r,s,T) - P^{(va)}(r,T)\, \hat{f}_a^T(s) - \hat{f}_a(r)\, P^{(av)}(s,T)$$

$$+ P^{(vb)}(r,T)\, g_{1_b}^T g_{1_{x_s}}^{-1}\, \hat{f}_{x_{ss}}^T \delta(s-1) - R_1^{-1}(r,s,T) \qquad (3.52)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\Omega^{(\tau)}(r,T) = P_T^{(va)}(r,T) - \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)\, P^{(va)}(\nu,T)\,d\zeta\, d\nu$$
$$- \hat{f}_x(r)\, P^{(va)}(r,T) - \hat{f}_{x_r}(r)\, P_r^{(va)}(r,T) - \hat{f}_{x_{rr}}(r)\, P_{rr}^{(va)}(r,T)$$
$$- P^{(va)}(r,T)\, \hat{A}_a^T - \hat{f}_a(r)\, P^{(aa)}(T) \qquad (3.53)$$

$$\Omega^{(\sigma)}(r,T) = P_T^{(vb)}(r,T) - \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)\, P^{(vb)}(\nu,T)\,d\zeta\, d\nu$$
$$- \hat{f}_x(r)\, P^{(vb)}(r,T) - \hat{f}_{x_r}(r)\, P_r^{(vb)}(r,T) - \hat{f}_{x_{rr}}(r,T)\, P_{rr}^{(vb)}(r,T)$$
$$- P^{(vb)}(r,T)\, \hat{B}_b^T - \hat{f}_a(r)\, P^{(ab)}(T) \qquad (3.54)$$

$$S(\zeta,\nu,T) = [h_x^T(\zeta,T,\hat{x})\, Q(\zeta,\nu,T)(y(\nu,T) - h(\nu,T,\hat{x}))]_x \qquad (3.55)$$

and $\hat{f}(r)$ denotes $f(r,T,\hat{x},\hat{x}_r,\hat{x}_{rr},\hat{a})$ in the above equations. In order to evaluate the last three terms of (3.51), we need the boundary conditions on $P^{(vv)}(r,s,T)$. From (3.8), (3.9), (3.15), (3.22) and (3.23) we have the following relationships when T is the final time:

$$g_o(T,\hat{x},\hat{x}_r) = 0 \quad , \quad r = 0 \qquad (3.56)$$

$$g_1(T,\hat{x},\hat{x}_r,b) = 0 \quad , \quad r = 1 \qquad (3.57)$$

If we consider the imbedded final time case, i.e., $c^{(\lambda)}(r) \neq 0$, $c^{(\tau)} \neq 0$, and $c^{(\sigma)} \neq 0$, we can obtain the following after substituting (3.22), (3.23), (3.29) - (3.31) and (3.46) - (3.48) into (3.8) and (3.9) and expanding about $\hat{x}$, $\hat{a}$, and $\hat{b}$:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$g_1(T,\hat{x},\hat{x}_r,\hat{b}) - \frac{1}{2}\int_0^1 [\hat{g}_{1_x} P^{(vv)}(r,s,T) + \hat{g}_{1_{x_r}} P_r^{(vv)}(r,s,T)$$

$$+ \hat{g}_{1_b} P^{(bv)}(s,T) - R_4^{-1}(T)\ \hat{g}_{1_{x_s}}^{-1}\ \hat{f}_{x_{ss}}^T\ \delta(s-1)]\ c^{(\lambda)}(s)ds -$$

$$- \frac{1}{2}[\hat{g}_{1_x} P^{(va)}(r,T) + \hat{g}_{1_{x_r}} P_r^{(va)}(r,T) + \hat{g}_{1_b} P^{(ba)}(T)]\ c^{(\tau)}$$

$$- \frac{1}{2}[\hat{g}_{1_x} P^{(vb)}(r,T) + \hat{g}_{1_{x_r}} P_r^{(vb)}(r,T) + \hat{g}_{1_b} P^{(bb)}(T)]\ c^{(\sigma)}$$

$$+ 0(c^{(\lambda)^2}, c^{(\tau)^2}, c^{(\sigma)^2}) = 0\ ,\ r = 1 \qquad (3.58)$$

In order to use (3.56) and (3.57) as the boundary conditions for $\hat{x}(r,T)$ for the imbedded final time $T + \Delta$ and to satisfy (3.46) - (3.48), we need each coefficient of $c^{(\lambda)}(r)$, $c^{(\tau)}$ and $c^{(\sigma)}$ in (3.58) to become identically zero. The same applies to the $r = 0$ case. Thus we have all the necessary boundary conditions,

$$\hat{g}_{o_x} P^{(vv)}(r,s,T) + \hat{g}_{o_{x_r}} P_r^{(vv)}(r,s,T) + R_3^{-1}(T)\ \hat{g}_{o_{x_s}}^{-1}\ \hat{f}_{x_{ss}}^T\ \delta(s) = 0\ ,$$

$$r = 0 \qquad (3.59)$$

$$\hat{g}_{1_x} P^{(vv)}(r,s,T) + \hat{g}_{1_{x_r}} P_r^{(vv)}(r,s,T) + \hat{g}_{1_b} P^{(vb)^T}(s,T)$$

$$- R_4^{-1}(T)\ \hat{g}_{1_{x_s}}^{-1}\ \hat{f}_{x_{ss}}^T\ \delta(s-1) = 0\ ,\ r = 0 \qquad (3.60)$$

$$\hat{g}_{o_x} P^{(va)}(r,T) + \hat{g}_{o_{x_r}} P_r^{(va)}(r,T) = 0\ ,\ r = 0 \qquad (3.61)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\hat{g}_{1_x} P^{(va)}(r,T) + \hat{g}_{1_{x_r}} P_r^{(va)}(r,T) + \hat{g}_{1_b} P^{(ab)^T}(T) = 0, \quad r = 1 \tag{3.62}$$

$$\hat{g}_{o_x} P^{(vb)}(r,T) + \hat{g}_{o_{x_r}} P_r^{(vb)}(r,T) = 0 , \quad r = 0 \tag{3.63}$$

$$\hat{g}_{1_x} P^{(vb)}(r,T) + \hat{g}_{1_{x_r}} P_r^{(vb)}(r,T) + \hat{g}_{1_b} P^{(bb)}(T) = 0 , \quad r = 1 \tag{3.64}$$

Using (3.27), (3.28), (3.50), (3.59), and (3.60) the last three terms of (3.51) can be evaluated

$$\begin{aligned} P^{(vv)}(r,s,T)\, f_{x_s}^T c^{(\lambda)}(s) - P^{(vv)}(r,s,T)\, [f_{x_{ss}}^T c^{(\lambda)}(s)]_s \\ + P_s^{(vv)}(r,s,T)\, f_{x_{ss}}^T c^{(\lambda)}(s) \\ = -P^{(vb)}(r,T)\, \hat{g}_{1_b}^T \hat{g}_{1_{x_s}}^{-1} f_{x_{ss}}^T c^{(\lambda)}(s) + 0(c^{(\lambda)^2}), \quad s = 1 \end{aligned} \tag{3.65}$$

$$\begin{aligned} -P^{(vv)}(r,s,T)\, f_{x_s}^T c^{(\lambda)}(s) + P^{(vv)}(r,s,T)[f_{x_{ss}}^T c^{(\lambda)}(s)]_s \\ -P_s^{(vv)}(r,s,T)\, f_{x_{ss}}^T c^{(\lambda)}(s) = 2P^{(vv)}(r,s,T)\, h_x^T(s,T,\hat{x})\, Q(0,0,T)(y(0,T) \\ - h(0,T,\hat{x})) - P^{(vv)}(r,s,T)\, S(0,0,t)\{ \int_0^1 P^{(vv)}(s,\nu,T) c^{(\lambda)}(\nu)\, d\nu \\ + P^{(va)}(s,T) c^{(\tau)} + P^{(vb)}(s,T) c^{(\sigma)}\} + 0(c^{(\lambda)^2}) , \quad s = 0 \end{aligned} \tag{3.66}$$

Combining (3.51) with (3.65) and (3.66) we obtain the differential equations governing $\hat{x}(r,T)$, $P^{(vv)}(r,s,T)$, $P^{(va)}(r,T)$ and $P^{(vb)}(r,T)$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\hat{x}_T(r,T) = f(r,T,\hat{x},\hat{x}_r,\hat{x}_{rr},\hat{a}) + \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T) h_x^T(\zeta,T,x)\, Q(\zeta,\nu,T)$$

$$(y(\nu,T) - h(\nu,T,\hat{x}))\, d\zeta\, d\nu$$

$$+ P^{(vv)}(r,0,T)\, h_x^T(0,T,\hat{x})\, Q(0,0,T)(y(0,T) - h(0,T,\hat{x})) \qquad (3.67)$$

$$P_T^{(vv)}(r,s,T) = \hat{f}_x(r)\, P^{(vv)} + P^{(vv)}\, \hat{f}_x^T(s) + \hat{f}_{x_r}(r)\, P_r^{(vv)} + P_s^{(vv)} \hat{f}_{x_s}^T(s)$$

$$+ \hat{f}_{x_{rr}}(r)\, P_{rr}^{(vv)} + P_{ss}^{(vv)}\, \hat{f}_{x_{ss}}^T(s) + \hat{f}_a(r)\, P^{(va)^T}(s,T)$$

$$+ P^{(va)}(r,T)\, \hat{f}_a^T(s) + \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)\, P^{(vv)}(\nu,s,T) d\zeta\, d\nu$$

$$+ P^{(vv)}(r,0,T)\, s(0,0,T)\, P^{(vv)}(0,s,T) + R_1^{-1}(r,s,T) \qquad (3.68)$$

$$P_T^{(va)}(r,T) = \hat{f}_x(r)\, P^{(va)} + \hat{f}_{x_r}(r)\, P_r^{(va)} + \hat{f}_{x_{rr}}(r)\, P_{rr}^{(va)} + \hat{f}_a(r)\, P^{(aa)}$$

$$+ P^{(va)}\, \hat{A}_a^T + \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)\, P^{(va)}(\nu,T)\, d\zeta\, d\nu$$

$$+ P^{(vv)}(r,0,T)\, S(0,0,T)\, P^{(va)}(0,T) \qquad (3.69)$$

$$P_T^{(vb)}(r,T) = \hat{f}_x(r)\, P^{(vb)} + \hat{f}_{x_r}(r)\, P_r^{(vb)} + \hat{f}_{x_{rr}}(r)\, P_{rr}^{(vb)} + \hat{f}_a(r)\, P^{(ab)}$$

$$+ P^{(vb)}\, \hat{B}_b^T + \int_0^1 \int_0^1 P^{(vv)}(r,\zeta,T)\, S(\zeta,\nu,T)\, P^{(vb)}(\nu,T)\, d\zeta\, d\nu$$

$$+ P^{(vv)}(r,0,T)\, S(0,0,T)\, P^{(vb)}(0,T) \qquad (3.70)$$

Similarly, we can substitute (3.46) - (3.48) into (3.44) and (3.45), linearize and collect coefficients of like powers of $c^{(\lambda)}$, $c^{(\tau)}$ and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$c^{(\sigma)}$ to obtain the differential equations governing $\hat{a}(T)$, $\hat{b}(T)$, $P^{(ab)}(T)$, $P^{(aa)}(T)$, and $P^{(bb)}(T)$. The resulting equations are

$$\frac{d\hat{a}(T)}{dT} = A(T,\hat{a}) + \int_0^1 \int_0^1 P^{(av)}(\zeta,T)\, h_x^T(\zeta,T,x)\, Q(\zeta,\nu,T)(y(\nu,T) - h(\nu,T,\hat{x}))\, d\zeta\, d\nu + P^{(av)}(0,T)\, h_x^T(0,T,\hat{x})\, Q(0,0,T)\, (y(0,T) - h(0,T,\hat{x})) \tag{3.71}$$

$$\frac{d\hat{b}(T)}{dT} = B(T,\hat{b}) + \int_0^1 \int_0^1 P^{(bv)}(\zeta,T)\, h_x^T(\zeta,T,\hat{x})\, Q(\zeta,\nu,T)(y(\nu,T) - h(\nu,T,\hat{x}))d\zeta\, d\nu + P^{(bv)}(0,T)h_x^T(0,T,\hat{x})\, Q(0,0,T)(y(0,T) - h(0,T,\hat{x})) \tag{3.72}$$

$$\frac{dP^{(aa)}(T)}{dT} = \hat{A}_a P^{(aa)} + P^{(aa)}\, \hat{A}_a^T + \int_0^1 \int_0^1 P^{(av)}(\zeta,T)\, S(\zeta,\nu,T)\, P^{(va)}(\nu,T)d\zeta\, d\nu + P^{(av)}(0,T)\, S(0,0,T)\, P^{(va)}(0,T) + R_2^{-1}(T) \tag{3.73}$$

$$\frac{dP^{(ab)}(T)}{dT} = \hat{A}_a\, P^{(ab)} + P^{(ab)}\, \hat{B}_b^T + \int_0^1 \int_0^1 P^{(av)}(\zeta,T)\, S(\zeta,\nu,T)P^{(vb)}(\nu,T)d\zeta d\nu + P^{(av)}(0,T)\, S(0,0,T)\, P^{(vb)}(0,T) \tag{3.74}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\frac{dP^{(bb)}(T)}{dT} = \hat{B}_b\, P^{(bb)} + P^{(bb)}\, \hat{B}_b^T + \int_0^1 \int_0^1 P^{(vb)^T}(\zeta,T)\, S(\zeta,\nu,T)$$

$$P^{(vb)}(\nu,T)\, d\zeta\, d\nu + R_5^{-1}(T) \qquad (3.75)$$

Equations (3.67) - (3.85) constitute the distributed nonlinear filter for (2.1) - (2.6). It is easy to check that $P^{(bv)} = P^{(vb)^T}$, $P^{(av)} = P^{(va)^T}$, and $P^{(ab)} = P^{(ba)^T}$.

The filter obtained can now be summarized:

	Equation	Initial Conditions	Boundary Conditions
Estimates			
$\hat{x}(r,T)$	(3.67)	$\hat{x}(r,0)$	(3.56) (3.57)
$\hat{a}(T)$	(3.71)	$\hat{a}(0)$	none
$\hat{b}(T)$	(3.72)	$\hat{b}(0)$	none
Covariances			
$P^{(vv)}(r,s,T)$	(3.68)	$P^{(vv)}(r,s,0)$	(3.59) (3.60) with (3.50)
$P^{(va)}(r,T)$	(3.69)	$P^{(va)}(r,0)$	(3.61) (3.62)
$P^{(vb)}(r,T)$	(3.70)	$P^{(vb)}(r,0)$	(3.63) (3.64)
$P^{(aa)}(T)$	(3.73)	$P^{(aa)}(0)$	none
$P^{(ab)}(T)$	(3.74)	$P^{(ab)}(0)$	none
$P^{(bb)}(T)$	(3.75)	$P^{(bb)}(0)$	none

4. Optimal Least Square Interpolation

The sequential interpolation (fixed-time smoothing) problem is interpreted as choosing $x(r,t_1)$, $a(t_1)$ and $b(t_1)$, where $t_1 \varepsilon [0,T]$ and $T \geq t_1$, which minimize the error criterion (2.7). This

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

statistical minimization problem can be reformulated as an optimal control problem as shown in the previous section. In this section the approach of Kagiwada et al[29] for sequential interpolation in nonlinear lumped systems is extended to nonlinear distributed systems.

Reformulating the original problem as an optimal control problem, we want to determine $x(r,t_1)$, $a(t_1)$ and $b(t_1)$ to minimize (3.1) subject to constraints (3.2) - (3.6). If we have the optimal solution $x^*(r,t)$, $a^*(t)$ and $b^*(t)$, $t \in [0,T]$ which minimizes (3.1) with (3.2) - (3.6), then the optimal solution will satisfy the necessary conditions for optimality, (3.7) - (3.21). In addition, the optimal least square interpolation solution which minimizes (3.1) coincides with $x^*(r,t)|_{t_1}$, $a^*(t)|_{t_1}$ and $b^*(t)|_{t_1}$ from the assumed uniqueness of the optimal solution. Hence we have the same necessary conditions for optimality for the interpolation problem as for the filtering problem.

It is necessary to determine the Cauchy type representation of the interpolation solution on the basis of the two point boundary value problem. If we consider the imbedded final time case where $\lambda(r,T) = c^{(\lambda)}(r)$, $\tau(T) = c^{(\tau)}$ and $\sigma(T) = c^{(\sigma)}$, the interpolation solution can be written as

$$x(r,t_1) = \phi(t_1,T,r,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{4.1}$$

$$a(t_1) = \phi^{(a)}(t_1,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{4.2}$$

$$b(t_1) = \phi^{(b)}(t_1,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{4.3}$$

Therefore the desired solution becomes $\phi(t_1,T,r,0,0,0)$,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$\phi^{(a)}(t_1,T,0,0,0)$ and $\phi^{(b)}(t_1,T.0,0,0)$. Using (3.32) - (3.37), we can write the following relationship for the final time $T + \Delta$

$$\lambda(T + \Delta) = \lambda(r,T) + \lambda_T(r,T)\,\Delta + 0(\Delta^2)$$

$$= c^{(\lambda)}(r) + \beta(r,T,c^{(\lambda)}(r),\ \phi(T,T,r,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),$$

$$\phi^{(a)}(T,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}))\Delta + 0(\Delta^2) \qquad (4.4)$$

Also we have

$$\phi(t_1,T + \Delta,\ r,\ \lambda(r,T+\Delta),\ \tau(T+\Delta),\ \sigma(T+\Delta))$$

$$= \phi(t_1,T,r,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \qquad (4.5)$$

Similarly, we have

$$\tau(T + \Delta) = c^{(\tau)} + \gamma(T,\phi(T,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),\ c^{(\lambda)}(r),$$

$$\phi^{(a)}(T,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),\ c^{(\tau)})\Delta + 0(\Delta^2) \qquad (4.6)$$

$$\sigma(T + \Delta) = c^{(\sigma)} + \theta(T,\phi(T,T,r,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),\ c^{(\lambda)}(r),$$

$$\phi^{(b)}(T,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}),\ c^{(\sigma)})\Delta + 0(\Delta^2) \qquad (4.7)$$

$$\phi^{(a)}(t_1,T+\Delta,\ \lambda(r,T+\Delta),\ \tau(T+\Delta),\ \sigma(T+\Delta))$$

$$= \phi^{(a)}(t_1T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \qquad (4.8)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\phi^{(b)}(t_1, T+\Delta),\ \lambda(r, T+\Delta),\ \tau(T+\Delta),\ \sigma(T+\Delta))$$

$$= \phi^{(b)}(t_1, T, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)} \tag{4.9}$$

Substituting (4.4), (4.6) and (4.7) into (4.5), and taking the Taylor expansion with limiting operation $\Delta \to 0$, we have

$$\phi_T + \int_0^1 \frac{\delta\phi}{\delta c^{(\lambda)}(r)} \beta(r, T, c^{(\lambda)}(r),\ \psi, \psi^{(a)})\, dr + \phi_{c^{(\tau)}} \gamma(T, \psi, c^{(\lambda)}(r),$$

$$\cdot\ \psi^{(a)}, c^{(\tau)}) + \phi_{c^{(\sigma)}}\ \theta(T, \psi, c^{(\lambda)}(r),\ \psi^{(b)},\ c^{(\sigma)}) = 0 \tag{4.10}$$

Similarly we obtain

$$\phi_T^{(a)} + \int_0^1 \frac{\delta\phi^{(a)}}{\delta c^{(\lambda)}(r)} \beta(r, T, c^{(\lambda)}(r), \psi, \psi^{(a)})\, dr + \phi^{(a)}_{c^{(\tau)}}\ \gamma(T, \psi, c^{(\lambda)}(r),$$

$$\psi^{(a)},\ c^{(\tau)}) + \phi^{(a)}_{c^{(\sigma)}}\ \theta(T, \psi, c^{(\lambda)}(r),\ \psi^{(b)},\ c^{(\sigma)}) = 0 \tag{4.11}$$

$$\phi_T^{(b)} + \int_0^1 \frac{\delta\phi^{(b)}}{c^{(\lambda)}(r)} \beta(r, T, c^{(\lambda)}(r), \psi, \psi^{(a)})\, dr + \phi^{(b)}_{c^{(\tau)}}\ \gamma(T, \psi, c^{(\lambda)}(r),$$

$$\psi^{(a)},\ c^{(\tau)}) + \phi^{(b)}_{c^{(\sigma)}}\ \theta(T, \psi, c^{(\lambda)}(r),\ \psi^{(b)},\ c^{(\sigma)}) = 0 \tag{4.12}$$

where

$$\psi(T, r, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \phi(T, T, r, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}$$

$$\psi^{(a)}(T, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \phi^{(a)}(T, T, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)})$$

$$\psi^{(b)}(T, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \phi^{(b)}(T, T, c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) \tag{4.13}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Equations (4.10) - (4.12) are the desired initial value problem together with (3.43) - (3.45) and (4.13). Consequently we can consider only ϕ, $\phi^{(a)}$, and $\phi^{(b)}$, since (3.43) - (3.45) generate the previous filter results. Let us assume the solutions of the form, as a first order approximation

$$\phi(t_1,T,r,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \hat{z}(r,t_1,T) - \frac{1}{2}\int_0^1 W^{(vv)}(r,s,t_1,T)\ c^{(\lambda)}(s)\ ds - \frac{1}{2} W^{(va)}(r,t_1,T)\ c^{(\tau)} - \frac{1}{2} W^{(vb)}(r,t_1,T)\ c^{(\sigma)} \tag{4.14}$$

$$\phi^{(a)}(t_1,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \hat{e}(t_1,T) - \frac{1}{2}\int_0^1 W^{(av)}(s,t_1,T)c^{(\lambda)}(s)ds - \frac{1}{2} W^{(aa)}(t_1,T)c^{(\tau)} - \frac{1}{2} W^{(ab)}(t_1,T)\ c^{(\sigma)} \tag{4.15}$$

$$\phi^{(b)}(t_1,T,c^{(\lambda)}(r),\ c^{(\tau)},\ c^{(\sigma)}) = \hat{d}(t_1,T) - \frac{1}{2}\int_0^1 W^{(bv)}(s,t_1,T)c^{(\lambda)}(s)ds - \frac{1}{2} W^{(ba)}(t_1,T)c^{(\tau)} - \frac{1}{2} W^{(bb)}(t_1,T)\ c^{(\sigma)} \tag{4.16}$$

From (4.13) we have

$$\begin{aligned} \hat{z}(r,T,T) &= \hat{x}(r,T) \\ \hat{e}(T,T) &= \hat{a}(T) \\ \hat{d}(T,T) &= \hat{b}(T) \end{aligned} \tag{4.17}$$

$$W^{(vv)}(r,s,T,T) = P^{(vv)}(r,s,T) \qquad W^{(aa)}(T,T) = P^{(aa)}(T)$$

$$W^{(va)}(r,T,T) = P^{(va)}(r,T) \qquad W^{(ab)}(T,T) = P^{(ab)}(T)$$

$$W^{(vb)}(r,T,T) = P^{(vb)}(r,T) \qquad W^{(ba)}(T,T) = P^{(ba)}(T)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$W^{(av)}(r,T,T) = P^{(av)}(r,T) \qquad W^{(bb)}(T,T) = P^{(bb)}(T)$$

$$W^{(bv)}(r,T,T) = P^{(bv)}(r,T)$$

In the interpolation case we do not have the similar relationship to (3.50), since we are dealing with the quantities at t_1 and at T. If the system is linear and with Gaussian white noise, then (4.14) - (4.16) yield the exact solution of (4.10) - (4.12) with the following statistical interpretations:

$$W^{(vv)}(r,s,t_1,T) = E\{(x(r,t_1) - \hat{z}(r,t_1,T))(x(s,T) - \hat{x}(s,T))^T\}$$

$$W^{(va)}(r,t_1,T) = E\{(x(r,t_1) - \hat{z}(r,t_1,T))(a(T) - \hat{a}(T))^T\}$$

$$W^{(ab)}(t_1,T) = E\{(a(t_1) - \hat{e}(t_1,T))(b(T) - \hat{b}(T))^T\}, \text{ etc.} \qquad (4.18)$$

To obtain the governing differential equations for the W's we can follow the same procedure as in the filtering case. Substituting (4.14) - (4.16) into (4.10) and applying Taylor expansion, we can obtain

$$\hat{z}_T(r,t_1,T) - \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\, h_x^T(\zeta,T,\hat{x})\, Q(\zeta,\nu,T)(y(\nu,T) - h(\nu,T,\hat{x}))\, d\zeta\, d\nu - \frac{1}{2}\int_0^1 \Gamma^{(\lambda)}(r,s,T)\, c^{(\lambda)}(s)\, ds$$

$$- \frac{1}{2}\Gamma^{(\tau)}(r,T)\, c^{(\tau)} - \frac{1}{2}\Gamma^{(\sigma)} c^{(\sigma)} - \frac{1}{2}\{W^{(vv)}(r,s,t_1,T)c^{(\lambda)}(s)f_{x_s}(s)$$

$$- W^{(vv)}(r,s,t_1,T)[c^{(\lambda)}(s)f_{x_{ss}}(s)]_s + W_s^{(vv)}(r,s,t_1,T)$$

$$c^{(\lambda)}(s)f_{x_{ss}}\}\Big|_{s=0}^{s=1} + 0((c^{(\lambda)})^2, (c^{(\tau)})^2, (c^{(\sigma)})^2) = 0 \qquad (4.19)$$

where

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\Gamma^{(\lambda)}(r,s,T) = W_T^{(vv)}(r,s,t_1,T) - \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\, S(\zeta,\nu,T)$$

$$P^{(vv)}(\nu,s,T)\, d\zeta\, d\nu - W^{(vv)}(r,s,t_1,T)\, \hat{f}_x^T(s) - W_s^{(vv)}(r,s,t_1,T)\, \hat{f}_{x_s}^T(s)$$

$$- W_{ss}^{(vv)}(r,s,t_1,T)\, \hat{f}_{x_{ss}}^T(s) - W^{(va)}(r,t_1,T)\, \hat{f}_a^T(s)$$

$$+ W^{(vb)}(r,t_1,T)\, \hat{g}_{1_b}^T\, \hat{g}_{1_{x_s}}^{-1}\, f_{x_{ss}}^T(s)\, \delta(s-1) \qquad (4.20)$$

$$\Gamma^{(\tau)}(r,t_1,T) = W_T^{(va)}(r,t_1,T) - \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\, S(\zeta,\nu,T)$$

$$P^{(va)}(\nu,T)\, d\zeta\, d\nu - W^{(va)}(r,t_1,T)\, \hat{A}_a^T \qquad (4.21)$$

$$\Gamma^{(\sigma)}(r,t_1,T) = W_T^{(vb)}(r,t_1,T) - \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\, S(\zeta,\nu,T)$$

$$P^{(vb)}(\nu,T)\, d\zeta\, d\nu - W^{(vb)}(r,t_1,T)\, \hat{B}_b^T \qquad (4.22)$$

To evaluate the last three terms of (4.19) we need the boundary condition on $W^{(vv)}$ at $s = 0$ and $s = 1$. But we cannot handle the boundary conditions as (3.58) because of (4.18). From the analogy to the linear system with Gaussian white noise, i.e., (4.18), we can assume the boundary conditions for $W^{(vv)}(r,s,t_1,T)$ at $s = 0$ and $s = 1$ in the form

$$W^{(vv)}(r,s,t_1,T)\, \hat{g}_{o_x}^T + W_s^{(vv)}(r,s,t_1,T)\, \hat{g}_{o_{x_s}}^T = 0\,, \quad s = 0 \qquad (4.23)$$

$$W^{(vv)}(r,s,t_1,T)\, \hat{g}_{1_x}^T + W_s^{(vv)}(r,s,t_1,T)\, \hat{g}_{1_{x_s}}^T + W^{(vb)}(r,t_1,T)\, \hat{g}_{1_b}^T = 0\,,$$

$$s = 1 \qquad (4.24)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and

$$g_o(t_1,\hat{z},\hat{z}_r) = 0 \ , \quad r = 0 \tag{4.25}$$

$$g_1(t_1,\hat{z},\hat{z}_r,\hat{d}) = 0 \ , \quad r = 1 \tag{4.26}$$

Combining (3.27), (3.28), (4.23) and (4.24) with (4.19) we can obtain the following differential equations for $\hat{z}(r,t_1,T)$, $W^{(vv)}(r,s,t_1,T)$, $W^{(va)}(r,t_1,T)$ and $W^{(vb)}(r,t_1,T)$,

$$\hat{z}_T(r,t_1,T) = \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\ h_x^T(\zeta,T,\hat{x})\ Q(\zeta,\nu,T)\ (y(\nu,T) - h(\nu,T,\hat{x}))\ d\zeta\ d\nu + W^{(vv)}(r,0,t_1,T)\ h_x^T(0,t,\hat{x})\ Q(0,0,T)\ (y(0,T) - h(0,T,\hat{x})) \tag{4.27}$$

$$W_T^{(vv)}(r,s,t_1,T) = W^{(vv)}(r,s,t_1,T)\ \hat{f}_x^T(s) + W_s^{(vv)}(r,s,t_1,T)\ \hat{f}_{x_s}^T(s) + W_{ss}^{(vv)}(r,s,t_1,T)\ \hat{f}_{x_{ss}}^T(s) + W^{(va)}(r,t_1,T)\ \hat{f}_a^T(s) + \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\ S(\zeta,\nu,T)\ P^{(vv)}(\nu,s,T)\ d\zeta\ d\nu + W^{(vv)}(r,0,t_1,T)\ S(0,0,T)\ P^{(vv)}(0,s,T) \tag{4.28}$$

$$W_T^{(va)}(r,t_1,T) = W^{(va)}(r,t_1,T)\ \hat{A}_a^T + \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\ S(\zeta,\nu,T)\ P^{(va)}(\nu,T)\ d\zeta\ d\nu + W^{(vv)}(r,0,t_1,T)\ S(0,0,T)\ P^{(va)}(0,T) \tag{4.29}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$W_T^{(vb)}(r,t_1,T) = W^{(vb)}(r,t_1,T)\ \hat{B}_b^T + \int_0^1 \int_0^1 W^{(vv)}(r,\zeta,t_1,T)\ S(\zeta,\nu,T)$$

$$P^{(vb)}(\nu,T)\ d\zeta\ d\nu + W^{(vv)}(r,0,t_1,T)\ S(0,0,T)\ P^{(vb)}(0,T) \qquad (4.30)$$

Similarly, we can obtain the differential equations governing $\hat{e}(t_1,T)$, $W^{(av)}(s,t_1,T)$, $W^{(aa)}(t_1,T)$ and $W^{(ab)}(t_1,T)$ from (4.11), (4.14) - (4.16)

$$\hat{e}_T(t_1,T) = \int_0^1 \int_0^1 W^{(av)}(\zeta,t_1,T)\ h_x^T(\zeta,T,\hat{x})\ Q(\zeta,\nu,T)\ (y(\nu,T)$$

$$- h(\nu,T,\hat{x}))\ d\zeta\ d\nu + W^{(av)}(0,t_1,T)\ h_x^T(0,T,\hat{x})\ Q(0,0,T)\ (y(0,T)$$

$$- h(0,T,\hat{x})) \qquad (4.31)$$

$$W_T^{(av)}(s,t_1,T) = W^{(aa)}(t_1,T)\ \hat{f}_a^T(s) + W^{(av)}(s,t_1,T)\ \hat{f}_x^T(s)$$

$$+ W_s^{(av)}(s,t_1,T)\ \hat{f}_{x_s}^T(s) + W_{ss}^{(av)}(s,t_1,T)\ \hat{f}_{x_{ss}}^T(s)$$

$$+ \int_0^1 \int_0^1 W^{(av)}(\zeta,t_1,T)\ S(\zeta,\nu,T)\ P^{(vv)}(\nu,s,T)\ d\zeta\ d\nu$$

$$+ W^{(av)}(0,t_1,T)\ S(0,0,T)\ P^{(vv)}(0,s,T) \qquad (4.32)$$

$$W_T^{(aa)}(t_1,T) = W^{(aa)}(t_1,T)\ \hat{A}_a^T + \int_0^1 \int_0^1 W^{(av)}(\zeta,t_1,T)\ S(\zeta,\nu,T)$$

$$P^{(va)}(\nu,T)\ d\zeta\ d\nu + W^{(av)}(0,t_1,T)\ S(0,0,T)\ P^{(va)}(0,T) \qquad (4.33)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$W_T^{(ab)}(t_1,T) = W^{(ab)}(t_1,T)\hat{B}_b^T + \int_0^1 \int_0^1 W^{(av)}(\zeta,t_1,T)\ S(\zeta,\nu,T)$$

$$P^{(vb)}(\nu,T)\ d\zeta\ d\nu + W^{(av)}(0,t_1,T)\ S(0,0,T)\ P^{(vb)}(0,T) \qquad (4.34)$$

where the following boundary conditions on $W^{(av)}(s,t_1,T)$ at $s = 0$ and $s = 1$ are assumed for (4.32)

$$W^{(av)}(s,t_1,T)\ g_{o_x}^T + W_s^{(av)}(s,t_1,T)\ g_{o_{x_s}}^T = 0\ , \quad s = 0 \qquad (4.35)$$

$$W^{(av)}(s,t_1,T)\ g_{1_x}^T + W_s^{(av)}(s,t_1,T)\ g_{1_{x_s}}^T + W^{(ab)}(t_1,T)\ g_{1_b}^T = 0\ , \quad s = 1 \qquad (4.36)$$

Combining (4.14) - (4.16) with (4.12) and following the same procedure as before,

$$\hat{d}_T(t_1,T) = \int_0^1 \int_0^1 W^{(bv)}(\zeta,t_1,T)\ h_x^T(\zeta,T,x)\ Q(\zeta,\nu,T)\ (y(\nu,T)$$

$$- h(\nu,T,\hat{x}))\ d\zeta\ d\nu + W^{(bv)}(0,t_1,T)\ h_x^T(0,T,\hat{x})\ Q(0,0,T)\ (y(0,T)$$

$$- h(0,T,\hat{x})) \qquad (4.37)$$

$$W_T^{(bv)}(s,t_1,T) = W^{(ba)}(t_1,T)\ \hat{f}_a^T + W^{(bv)}(s,t_1,T)\ \hat{f}_x^T(s)$$

$$+ W_s^{(bv)}(s,t_1,T)\ \hat{f}_{x_s}^T(s) + W_{ss}^{(bv)}(s,t_1,T)\ \hat{f}_{x_{ss}}^T(s)$$

$$+ \int_0^1 \int_0^1 W^{(bv)}(\zeta,t_1,T)\ S(\zeta,\nu,T)\ P^{(vv)}(\nu,s,T)\ d\zeta\ d\nu$$

$$+ W^{(bv)}(0,t_1,T)\ S(0,0,T)\ P^{(vv)}(0,s,T) \qquad (4.38)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$W_T^{(ba)}(t_1,T) = W^{(ba)}(t_1,T)\ \hat{A}_a^T + \int_0^1 \int_0^1 W^{(bv)}(\zeta,t_1,T)\ S(\zeta,\nu,T)$$

$$P^{(va)}(\nu,T)\ d\zeta\ d\nu + W^{(bv)}(0,t_1,T)\ S(0,0,T)\ P^{(va)}(0,T) \qquad (4.39)$$

$$W_T^{(bb)}(t_1,T) = W^{(bb)}(t_1,T)\ \hat{B}_b^T + \int_0^1 \int_0^1 W^{(bv)}(\zeta,t_1,T)\ S(\zeta,\nu,T)$$

$$P^{(vb)}(\nu,T)\ d\zeta\ d\nu + W^{(bv)}(0,t_1,T)\ S(0,0,T)\ P^{(vb)}(0,T) \qquad (4.40)$$

where the boundary condition of $W^{(bv)}(s,t_1,T)$ is taken as

$$W^{(bv)}(s,t_1,T)\ \hat{g}_{o_x}^T + W_s^{(bv)}(s,t_1,T)\ \hat{g}_{o_{x_s}}^T = 0\ , \quad s = 0 \qquad (4.41)$$

$$W^{(bv)}(s,t_1,T)\ \hat{g}_{1_x}^T + W_s^{(bv)}(s,t_1,T)\ \hat{g}_{1_{x_s}}^T + W^{(bb)}(t_1,T)\ \hat{g}_{1_b}^T = 0\ ,$$

$$s = 1 \qquad (4.42)$$

This completes the derivation of the interpolation equations for the nonlinear distributed system. The initial conditions for (4.27) - (4.42) can be obtained from (4.17) if we take $T = t_1$. Thus to solve the interpolation problem at t_1 we have to integrate the filter equations first up to t_1. Then we have to solve the filter and the interpolation equations simultaneously for $T > t_1$.

The interpolation equations can be summarized:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

	Equation	Initial Condition	Boundary Condition
Estimates			
$\hat{z}(r,t_1,T)$	(4.27)	$\hat{x}(r,t_1)$	(4.25) (4.26)
$\hat{e}(t_1,T)$	(4.31)	$\hat{a}(t_1)$	none
$\hat{d}(t_1,T)$	(4.37)	$\hat{b}(t_1)$	none
Covariance			
$W^{(vv)}(r,s,t_1,T)$	(4.28)	$P^{(vv)}(r,s,t_1)$	(4.23) (4.24)
$W^{(va)}(r,t_1,T)$	(4.29)	$P^{(va)}(r,t_1)$	none
$W^{(vb)}(r,t_1,T)$	(4.30)	$P^{(vb)}(r,t_1)$	none
$W^{(av)}(s,t_1,T)$	(4.32)	$P^{(av)}(r,t_1)$	(4.35) (4.36)
$W^{(bv)}(s,t_1,T)$	(4.38)	$P^{(bv)}(r,t_1)$	(4.41) (4.42)
$W^{(aa)}(t_1,T)$	(4.33)	$P^{(aa)}(r,t_1)$	none
$W^{(ab)}(t_1,T)$	(4.34)	$P^{(ab)}(t_1)$	none
$W^{(ba)}(t_1,T)$	(4.39)	$P^{(ba)}(t_1)$	none
$W^{(bb)}(t_1,T)$	(4.40)	$P^{(bb)}(t_1)$	none

5. Examples

5.1 Example 1. It is required to perform filtering and sequential interpolation for the heat conduction system

$$x_t(r,t) = 0.1\, x_{rr} + 0.1\, x^2 + \xi_1(r,t) \tag{5.1}$$

$$x(0,t) - 0.05\, x_r = \xi_3(t)\ , \quad r = 0 \tag{5.2}$$

$$x_r = \xi_4(t)\ , \quad r = 1 \tag{5.3}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

with unknown initial condition

$$x(r,0) = 2 \sin \pi r \tag{5.4}$$

and noisy observations generated by

$$y(r_i,t) = x(r_i,t)\ [1 + \eta(t)], \quad i = 1,2,3 \tag{5.5}$$

where $r_1 = 0.25$, $r_2 = 0.50$, $r_3 = 0.75$. The dynamical disturbances are generated by

$$\begin{aligned} \xi_1(r,t) &= 0.1\ G(0,0.5) \\ \xi_3(t) &= \xi_4(t) = 0.15\ G(0,1) \\ \eta(t) &= 0.1\ G(0,1) \end{aligned} \tag{5.6}$$

where $G(0,\sigma)$ is a normally distributed random variable with mean zero and standard deviation σ .

The filter equations for $Q = 1$ are

$$\hat{x}_T(r,T) = 0.1\hat{x}_{rr} + 0.1\hat{x}^2 + \sum_{i=1}^{3} P^{(vv)}(r,r_i,T)\ [y(r_i,T) - \hat{x}(r_i,T)] \tag{5.7}$$

$$\hat{x}(0,T) - 0.05\hat{x}_r = 0 , \quad r = 0 \tag{5.8}$$

$$\hat{x}_r = 0 , \quad r = 1 \tag{5.9}$$

$$\begin{aligned} P_T^{(vv)}(r,s,T) &= 0.2\hat{x}(r,T)\ P^{(vv)}(r,s,T) + 0.2P^{(vv)}(r,s,T)\ \hat{x}(s,T) \\ &+ 0.1P_{rr}^{(vv)}(r,s,T) + 0.1P_{ss}^{(vv)}(r,s,T) - \sum_{i=1}^{3} P^{(vv)}(r,r_i,T)P^{(vv)}(r_i,s,T) \\ &+ R_1^{-1} \end{aligned} \tag{5.10}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$P^{(vv)}(0,s,T) - 0.05\, P_r^{(vv)}(0,s,T) - 20\, R_3^{-1}\, \delta(s) = 0 \qquad (5.11)$$

$$P_r^{(vv)}(1,s,T) - R_4^{-1}\, \delta(s-1) = 0$$

Initial conditions for (5.7) and (5.10) were chosen as

$$\hat{x}(r,0) = 0 \qquad (5.12)$$

$$P^{(vv)}(f,s,0) = 25 \exp(-0.5|r-s|) \qquad (5.13)$$

The additional interpolation equations are

$$\hat{z}_T(r,t_1,T) = \sum_{i=1}^{3} W^{(vv)}(r,r_i,t_1,T)\, [y(r_i,T) - \hat{x}(r_i,T)] \qquad (5.14)$$

$$W_T^{(vv)}(r,s,t_1,T) = 0.2W^{(vv)}(r,s,t_1,T)\, \hat{x}(s,T) + 0.1W_{ss}^{(vv)}(r,s,t_1,T) - \sum_{i=1}^{3} W^{(vv)}(r,r_i,t_1,T)\, P^{(vv)}(r_i,s,T) \qquad (5.15)$$

$$W^{(vv)}(r,0,t_1,T) - 0.05W_s^{(vv)}(r,0,t_1,T) = 0 \qquad (5.16)$$

$$W_s^{(vv)}(r,1,t_1,T) = 0 \qquad (5.17)$$

$$W^{(vv)}(r,s,t_1,t_1) = P^{(vv)}(r,s,t_1) \qquad (5.18)$$

$$\hat{z}(r,t_1,t_1) = \hat{x}(r,t_1) \qquad (5.19)$$

Numerical solution of (5.7) - (5.19) was carried out using quasilinearization and the Crank-Nicholson method[43,54] and the alternating direction method[54] for (5.10). The Dirac delta function

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

was approximated by $1/\Delta r$, the mesh spacing. The results are shown in Figures 1 and 2 for $Q = 1$, $R_1^{-1} = 0.5$, $R_3^{-1} = R_4^{-1} = 1.0$. The filter estimates shown in Figure 1 converge rapidly to the true (undisturbed) trajectories. Figure 2 presents a comparison of the true profile at $t = 0.4$, the filter estimate at $T = 0.4$, $\hat{x}(r,0.4)$, and the interpolating filter estimate of $\hat{x}(r,0.4)$ at $T = 2.0$, $\hat{z}(r,0.4,2.0)$. The additional observations collected from $t = 0.4$ to $t = 2.0$ are useful in improving the estimate at $t = 0.4$ through the use of the interpolating filter.

5.2 Example 2. We desire to estimate the state and the constant parameter a in the hyperbolic system, representing a plug flow tubular chemical reactor

$$x_t(r,t) + x_r(r,t) = - ax^2 \tag{5.20}$$

$$\frac{da}{dt} = 0 \tag{5.21}$$

$$x(0,t) = 1 \tag{5.22}$$

with unknown steady state solution

$$x(r,0) = (1 + ar)^{-1} \tag{5.23}$$

and unknown true value of $a = 2$. The observations are

$$y(r_i,t) = x(r_i,t)\ (1+0.1\ G(0,1)),\quad i = 1,2,3 \tag{5.24}$$

with $r_1 = 0.25$, $r_2 = 0.5$, and $r_3 = 0.75$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The corresponding filter equations for $Q = 1$ are

$$\hat{x}_T + \hat{x}_r = - \hat{a}\, \hat{x}^2 + \sum_{i=1}^{3} P^{(vv)}(r,r_i,T)\ (y(r_i,T) - \hat{x}(r_i,T)) \tag{5.25}$$

$$\frac{d\hat{a}}{dT} = \sum_{i=1}^{3} P^{(av)}(r_i,T)\ (y(r_i,T) - \hat{x}(r_i,T)) \tag{5.26}$$

$$\begin{aligned} P_T^{(vv)}(r,s,T) = &- 2\hat{a}\, \hat{x}(r,T)\, P^{(vv)}(r,s,T) - 2\hat{a}\, P^{(vv)}(r,s,T)\, \hat{x}(s,T) \\ &- \hat{x}^2(r,T)\, p^{(av)}(s,T) - P^{(av)}(r,T)\, x^2(s,T) - P_r^{(vv)}(r,s,T) \\ &- P_s^{(vv)}(r,s,T) - \sum_{i=1}^{3} P^{(vv)}(r,r_i,T)\, P^{(vv)}(r_i,s,T) \end{aligned} \tag{5.27}$$

$$\begin{aligned} P_T^{(av)}(r,T) = &-2\hat{a}\, \hat{x}(r,T)\, P^{(av)}(r,T) - P_r^{(av)}(r,T) \\ &- \hat{x}^2(r,T)\, P^{(aa)}(T) - \sum_{i=1}^{3} P^{(vv)}(r,r_i,T)\, P^{(av)}(r_i,T) \end{aligned} \tag{5.28}$$

$$\frac{dP^{(aa)}}{dT} = - \sum_{i=1}^{3} P^{(av)}(r_i,T)^2 \tag{5.29}$$

with

$$\begin{aligned} P^{(vv)}(0,s,T) &= 0 \\ P^{(av)}(0,T) &= 0 \end{aligned} \tag{5.30}$$

and

$$\begin{aligned} \hat{x}(r,0) &= 0 \\ \hat{a}(0) &= 1 \end{aligned}$$

$$P^{(vv)}(r,s,0) = 20 \sin(0.8\, \pi r) \sin(0.8\, \pi s)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$P^{(av)}(r,0) = 15 \sin(0.8\pi r)$$

$$P^{(aa)}(0) = 20 \qquad (5.31)$$

The numerical results for $\hat{x}(0.5,T)$ and $\hat{a}(T)$ are shown in Figure 3. Although convergence is slower than in Example 1, the results obtained confirm the applicability of the filter for estimating parameters in distributed systems.

6. Remarks

For discrete spatial measurements we define a new $Q_d(\zeta,\nu,T)$ as shown by Meditch[46] as

$$Q_d(\zeta,\nu,T) = \sum_{k=1}^{M} \sum_{\ell=1}^{M} Q_d'(r_k,r_\ell,T)\ \delta(\zeta - r_k)\ \delta(\nu - r_\ell) \qquad (6.1)$$

where

$$Q_d'(r_k,r_\ell,T) = \frac{1}{M^2} Q(r_k,r_\ell,T) \qquad (6.2)$$

Thus the integrations become discrete summations and Q_d becomes an $(nM) \times (nM)$ matrix.

In the case of no boundary noise, we put $R_3^{-1} = R_4^{-1} = 0$. When $g_{o_{x_r}} = 0$ or $g_{1_{x_r}} = 0$ with boundary noise, (3.25) and (3.26) give $\delta'(s)$ or $\delta'(s-1)$ in (3.59) and (3.60). However, in the linear case, Green's functions or eigenfunctions can be used to avoid the delta functions in the boundary conditions of $P^{(vv)}(r,s,T)$. Then the present results coincide with previous results[39,40]. Also, if we assume (2.1) is valid for the closed interval $[0,1]$ and at $r = 1, \hat{g}_{1_b} P^{(bv)}(r,T) \simeq P^{(vv)}(1,r,T)$, then the present results reduce to those of Appendix III-A.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

7. Figures

Figure 1. True and Filtered Values of $x(r,T)$ at Three Selected Locations for Example 1.

Figure 2. Comparison of True Profile at $t = 0.4$ with the Filter Estimate, $\hat{x}(r,0.4)$, and the Interpolating Filter Estimate at $T = 2.0$, $\hat{z}(r,0.4,2.0)$, for Example 1.

Figure 3. True and Filtered Values of $x(0.5,T)$ and a for Example 2.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

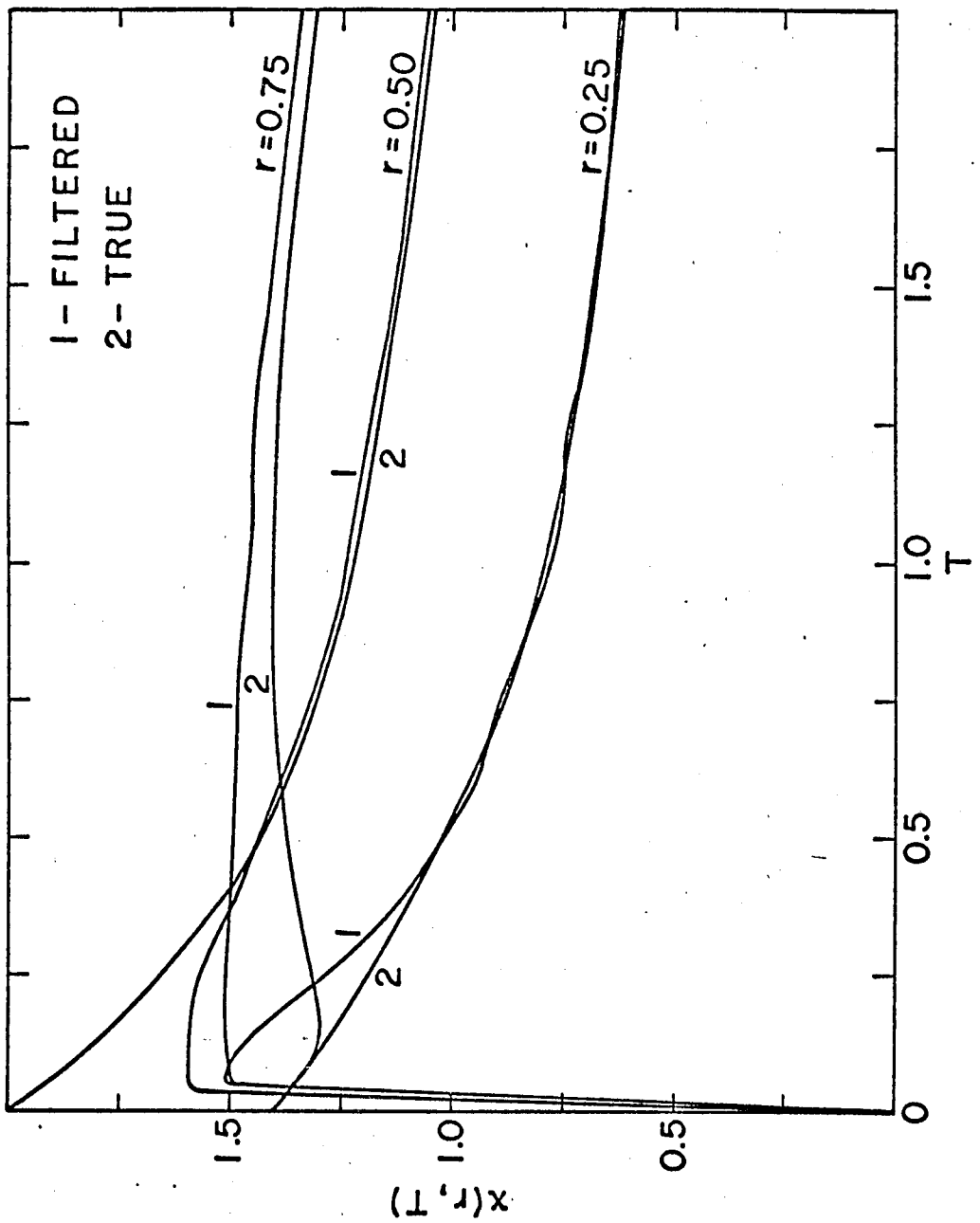

FIGURE 1

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

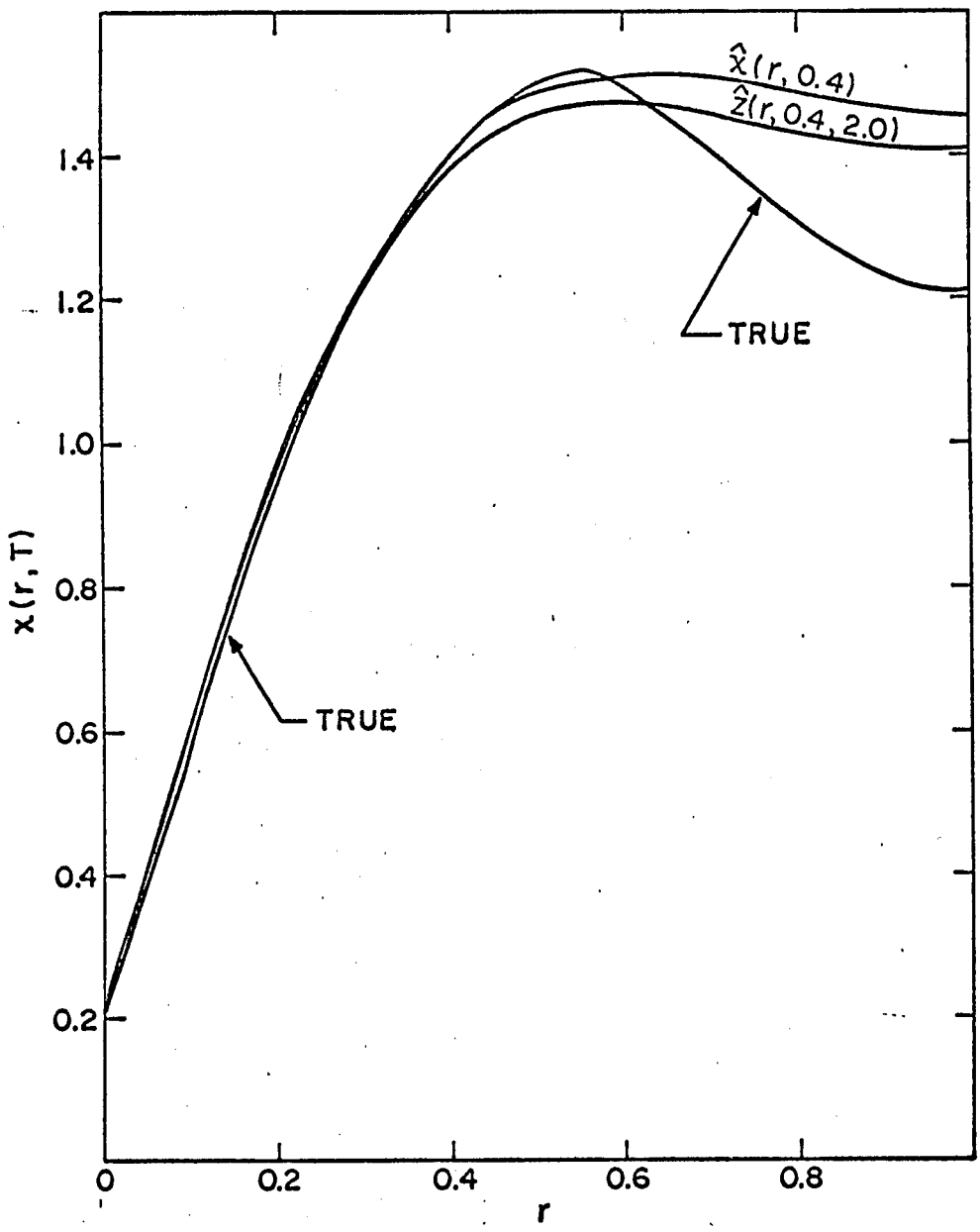

FIGURE 2

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

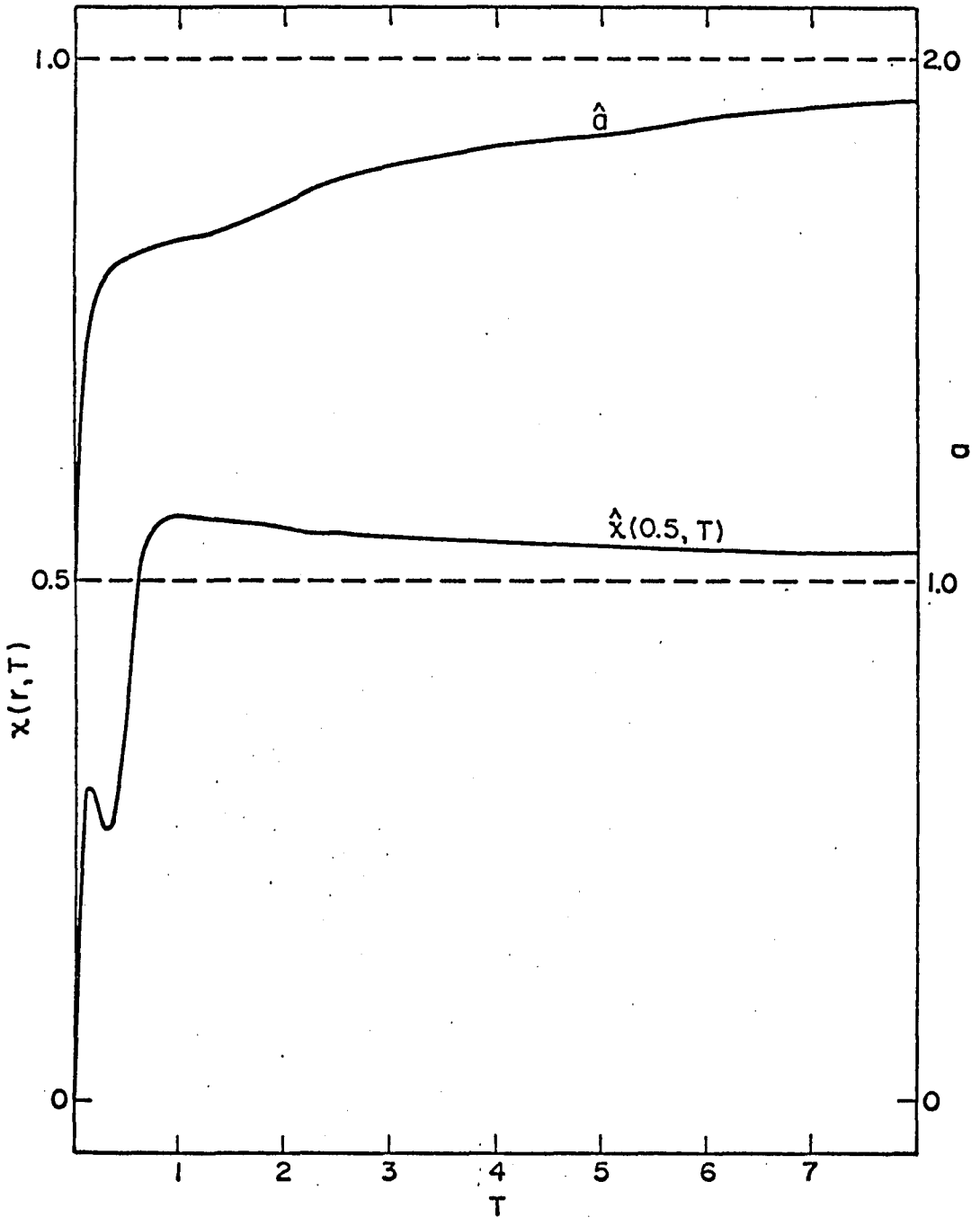

FIGURE 3

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix III-A

SESSION PAPER 22-C

NONLINEAR FILTERING IN DISTRIBUTED PARAMETER SYSTEMS *

J. H. Seinfeld, G. R. Gavalas, and M. Hwang

Chemical Engineering Laboratory
California Institute of Technology
Pasadena, California 91109

ABSTRACT

A general nonlinear filter is derived for systems described by partial differential equations which contain random disturbances in initial and boundary conditions, as dynamical inputs and measurement errors. Observations are assumed to be made continuously in time at an arbitrary number of discrete spatial locations. As an example, the filter is used to estimate the state in a nonlinear hyperbolic system describing a tubular flow chemical reactor.

INTRODUCTION

Estimation of the state of a nonlinear dynamic system has important engineering applications in modeling and adaptive control. With few exceptions, work on the estimation of the state of dynamic systems has been concentrated on systems described by ordinary differential equations. A large number of practical processes are described by partial differential equations, for which the problem of state estimation in the presence of noisy inputs and measurement errors is important. For example, estimating the effect of random disturbances on a transmission line, estimating the composition, temperature profiles, and catalyst activity in a packed bed catalytic reactor, and determining reservoir pressures from data at selected well locations represent applications involving the estimation of the state and parameters of a distributed system.

Exact solutions to the filtering problem for nonlinear lumped parameter systems have not been obtained, although a number of approximate nonlinear filters have been proposed (7). The purpose of this paper is to propose a nonlinear filter applicable to systems described by partial differential equations. Several studies have appeared recently on filtering for linear partial differential equations: Balakrishnan (1), Meditch (4), Thau (9), Tzafestas and Nightingale (10, 11), and Pell and Aris (5,6). A problem of similar nature, filtering for linear systems with time delays, was considered by Kwakernaak (3).

Nonlinear filters for distributed systems have been presented by Seinfeld (8) and Tzafestas and Nightingale (12). The former study (8) presented a filter applicable to nonlinear hyperbolic and parabolic systems with spatially continuous measurements. On the basis of a least square estimation criterion, a Hamilton-Jacobi equation was derived and solved approximately by a linearization in the region of the optimal estimate. The filter obtained is only an approximation to the more exact filter for spatially continuous measurements obtained in the latter study (12) by the same technique. Both of these filters assume deterministic boundary conditions.

In the present study a general filter is derived for systems described by nonlinear partial differential equations with unknown stochastic inputs in the system and the boundary conditions and noisy measurements carried out at an arbitrary number of discrete spatial locations. The approach of Pell and Aris (5,6) is used, namely conversion of the distributed system to a lumped system by finite difference approximations, application of a lumped parameter filter, and then performing a limiting operation on the spatial increment. Each of the linear and nonlinear filters cited above are special cases of the general nonlinear distributed filter presented here. The nonlinear filter is applied to estimate the state in a nonlinear hyperbolic system which describes the time-dependent behavior of a tubular chemical reactor.

DERIVATION OF THE FILTER

The problem is to estimate continuously the state of a distributed system, based on continuous noisy measurements carried out at a discrete number of spatial locations. We will consider the general class of systems on the reduced spatial domain, $r\epsilon[0,1]$, described by the vector

$$x_t = f(t,r,x,x_r,x_{rr}) + \xi_1(r,t) \qquad (1)$$

where $x(r,t)$ is the n-dimensional state vector, $\xi_1(r,t)$ is an n-dimensional vector of unknown random inputs, assumed to have zero mean, and x_t, x_r, and x_{rr} are partial derivatives.

The initial condition for (1) is

$$x(r,0) = x_0(r) \qquad (2)$$

which, in general, will not be known exactly. The general boundary conditions at $r = 0$ and $r = 1$ can be expressed as follows:

$$g_0(t,a_0(t),\ x(r,t),\ x_r(r,t))\Big|_{r=0} = 0 \qquad (3)$$

$$\frac{da_0(t)}{dt} = w_0(t,a_0(t)) + \xi_2(t) \qquad (4)$$

$$a_0(0) = a_{00} \qquad (5)$$

$$g_1(t,a_1(t),\ x(r,t),\ x_r(r,t))\Big|_{r=1} = 0 \qquad (6)$$

$$\frac{da_1}{dt} = w_1(t,a_1(t)) + \xi_3(t) \qquad (7)$$

$$a_1(0) = a_{10} \qquad (8)$$

Conditions (3) and (6) are completely general and include ℓ_0 and ℓ_1-dimensional inputs, $a_0(t)$ and $a_1(t)$, each of which is governed by an ordinary differential equation containing random excitations, $\xi_2(t)$ and $\xi_3(t)$. $a_0(t)$ and $a_1(t)$ account for the existence of controlled or uncontrolled inputs at the boundary of

* Presented at the Joint Automatic Control Conference, Georgia Institute of Technology, Atlanta, Georgia, 1970.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the process.

In the following derivations it is convenient to rewrite the boundary conditions (3), (6) in the form,

$$x(r,t)\big|_{r=0} = v\big(t,a_0(t),\ x(r,t),\ x_r(r,t)\big)\big|_{r=0} \quad (9)$$

$$x(r,t)\big|_{r=1} = k\big(t,a_1(t),\ x(r,t),\ x_r(r,t)\big)\big|_{r=1} \quad (10)$$

Observations of the system are carried out at m discrete points in the spatial domain, $r_1,r_2,\cdots,r_m$. At each point a q-dimensional vector of observations is made, $y(r_i,t)$, $i = 1,2,\cdots,m$. Let us define the mn-dimensional vector, $x_{ob}(t)$, consisting of the states at each measurement point,

$$x_{ob}(t) \triangleq \Big[x(r_1,t)^T,\ x(r_2,t)^T,\cdots,\ x(r_m,t)^T\Big]^T \quad (11)$$

Then the observations can be included in a p-dimensional vector y(t) and related to the state by

$$y(t) = h\big(t,x_{ob}(t)\big) + \eta(t) \quad (12)$$

where η is a p-dimensional vector of unknown measurement errors. The significance of y(t) can be seen from the following example. For a process with two states, each of which is measured directly at two points, n = 2, m = 2 and q = 2, the observations are described by

$$y_i(r_1,t) = x_i(r_1,t) + (\text{errors}) \quad i = 1,2$$

$$y_i(r_2,t) = x_i(r_2,t) + (\text{errors}) \quad i = 1,2$$

From (11), $x_{ob}(t) = [x_1(r_1,t),\ x_2(r_1,t),\ x_1(r_2,t),\ x_2(r_2,t)]^T$. If we let $h\big(t,x_{ob}(t)\big) = x_{ob}(t)$, then $y(t) = [y_1(r_1t),\ y_2(r_1,t),\ y_1(r_2,t),\ y_2(r_2,t)]^T$, and p = 4. In general, however, p is not necessarily equal to mq because a combination of values at different points may be the observed quantity rather than individual state measurements at each point. Thus, (12) represents a completely general representation of the measurements on the distributed system.

Approximation of the Distributed System

The filtering problem is the following: Given the observations y(t) from t = 0 to t = T, what is the best estimate of the state of the system, x(r,t), at t = T? The first step in solving the filtering problem is to approximate (1) by a set of ordinary differential equations using finite differences. The r-interval, [0,1], is divided into N parts and (1) is rewritten as

$$\frac{dx(i)}{dt} = f\Big(i\Delta,t,x(i),\ \frac{x(i)-x(i-1)}{\Delta},\ \frac{x(i+1)-2x(i)+x(i-1)}{\Delta^2}\Big) + \xi_1(i\Delta,t) \quad (13)$$

$$= f(i) + \xi_1(i\Delta,t)$$

where

$$\Delta = 1/N$$

$$x(i) = x(i\Delta,t)\ ;\ i = 0,1,\cdots,N \quad (14)$$

Let us now define X(t) as the (N+1) n-dimensional vector,

$$X(t) \triangleq [x(0,t)^T,\ x(\Delta,t)^T,\cdots,\ x(1,t)^T]^T \quad (15)$$

Then (13) can be written

$$\frac{dX}{dt} = F\big(t,X(t)\big) + \xi(t) \quad (16)$$

where

$$F\big(t,X(t)\big) \triangleq [f^T(0),\ f^T(1),\cdots,\ f^T(N)]^T \quad (17)$$

and

$$\xi(t) = [\xi_1{}^T(0,t),\ \xi_1{}^T(\Delta,t),\cdots,\ \xi_1{}^T(1,t)]^T \quad (18)$$

where $\xi_1(0,t) = \tilde{v}_{a_0}(t)\ \xi_2(t)$ and $\xi_1(1,t) = \tilde{k}_{a_1}(t)\xi_3(t)$, and $\tilde{v}_{a_0} = (I-v_x)^{-1}v_{a_0}$ and $\tilde{k}_{a_1} = (I-k_x)^{-1}k_{a_1}$.

The observations take the form,

$$y(t) = H\big(t,X(t)\big) + \eta(t) \quad (19)$$

where $H\big(t,X(t)\big)$ is obtained from $h\big(t,x_{ob}(t)\big)$ by relating $x_{ob}(t)$ to X(t). For this purpose it can be assumed that the m points of measurement coincide with mesh points since the mesh spacing Δ will be much finer than the measurement spacing. This assumption will not be required once the distributed filter has been recovered

Derivation of the Estimator Equation

With the use of finite difference approximations for the spatial derivatives, the distributed system has been transformed into a lumped system, for which the filtering problem is: Given the observations y(t) from t = 0 to t = T, what is the best estimate of the state of the system X(t), at t = T? One of the several nonlinear filters proposed for lumped parameter systems can now be applied to (16) and (19). Using the least square filter of Detchmendy and Sridhar, we obtain the nonlinear filter applicable to (16) and (19),

$$\frac{d\hat{X}}{dt} = F\big(t,\hat{X}(t)\big) + P'(t)H_X^T\ Q\Big(y-H\big(t,\hat{X}(t)\big)\Big) \quad (20)$$

$$\frac{dP'}{dt} = F_X P' + P'F_X^T + P'\Big[H_X^T\ Q\Big(y-H\big(t,\hat{X}(t)\big)\Big)\Big]_X P' + R^{-1} \quad (21)$$

where $\hat{X}$ is the least square estimate of X and Q and R are p x p and (N+1)n x (N+1) n-dimensional weighting matrices which must be specified a priori and P' is a (N+1)n x (N+1) n-dimensional matrix, corresponding to the covariance matrix of the estimate error in the Kalman filter. If the system is linear and $\xi(t)$ and $\eta(t)$ are zero-mean Gaussian white noise, then $E\{\xi(r_i,t)\ \xi(\rho_j,\tau)^T\} = R^{-1}(r_i,\rho_j,t)\ \delta(t-\tau)$ and $E\{\eta(r_i,t)\eta(\rho_j,\tau)^T\} = Q^{-1}(r_i,\rho_j,t)\ \delta(t-\tau)$ where R and Q are positive definite. As $\Delta r \to 0$, $R = R(r,\rho,t)$ and $Q = Q(r,\rho,t)$, $R^{-1}(0,0,t) = \tilde{v}_{a_0}(t)\ E\{\xi_2(t)\xi_2(t)^T\}\ \tilde{v}_{a_0}^T(t)$ and $R^{-1}(1,1,t) = \tilde{k}_{a_1}(t)\ E\{\xi_3(t)\xi_3(t)^T\}\ k_{a_1}^T(t)$. We assume that $E\{\xi(r,t)\ \eta(r,t)^T\} = 0$, i.e. the dynamic and observation errors are uncorrelated.

Let us now partition P' into submatrices, each of dimension n x n. We denote each submatrix by $P'_{ij}(t)$, where i = 0,1,···, N and j = 0,1,···, N, and define the following n x n matrices,

$$P(r,\rho,t) \triangleq P'_{ij}(t)\ ;\ r = i\Delta, \rho = j\Delta \quad (22)$$

$$P_{00}(t) \triangleq P(0,0,t) \text{ if } i = j = 0 \quad (23)$$

$$P_0(r,t) \triangleq P(r,0,t) \text{ if } j = 0,\ i \neq 0 \quad (24)$$

In the linear case these matrices are defined as the covariances of the estimate error,

$$P(r,\rho,t) = E\Big\{\big(x(r,t) - \hat{x}(r,t)\big)\big(x(\rho,t) - \hat{x}(\rho,t)\big)^T\Big\} \quad (25)$$

$$P_0(r,t) = E\Big\{\big(x(r,t) - \hat{x}(r,t)\big)\big(x(0,t) - \hat{x}(0,t)\big)^T\Big\} \quad (26)$$

$$P_{00}(t) = E\Big\{\big(x(0,t) - \hat{x}(0,t)\big)\big(x(0,t) - \hat{x}(0,t)\big)^T\Big\} \quad (27)$$

The following symmetry property, evident from (25), can also be expected to be valid in the nonlinear case,

$$P(r,\rho,t) = P^T(\rho,\ r,t) \quad (28)$$

We now take the limit as $N \to \infty$, $\Delta \to 0$, with $\Delta N = 1$ and obtain from (20)

$$\frac{\partial\hat{x}}{\partial t} = f\Big(r,t,\hat{x},\frac{\partial\hat{x}}{\partial r},\frac{\partial^2\hat{x}}{\partial r^2}\Big) + P_{ob}h_{x_{ob}}^T\ Q\big(y-h(t,\hat{x}_{ob})\big) \quad (29)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where

$$P_{ob}(t) \triangleq [P(r,r_1,t)\ ,\ P(r,r_2,t),\cdots,P(r,r_m,t)] \quad (30)$$

The estimator equation (29) is subject to the following boundary conditions, which are obtained directly from (9) and (10),

$$\dot{\hat{x}}(0,t) = v(t,\hat{a}_0(t),\ \hat{x}(0,t),\ \hat{x}_r(r,t)\big|_{r=0}) \quad (31)$$

and

$$\dot{\hat{x}}(1,t) = k(t,\hat{a}_1(t),\ \hat{x}(1,t),\ \hat{x}_r(r,t)\big|_{r=1}) \quad (32)$$

Estimator equations for a_0 and a_1 can be obtained by performing the limiting operation on the $i = 0$ component of (20). Differentiating both sides of (9) with respect to time we obtain

$$\dot{x}(0,t) = \left[v_t+v_{a_0}v_0+v_x\,\dot{x}(r,t)+v_{x_r}\dot{x}_r\right]_{r=0} + v_{a_0}\xi_2 \quad (33)$$

$$=(I - v_x)^{-1}\left[v_t + v_{a_0}v_0 + v_{x_r}\dot{x}_r\right]+ \tilde{v}_{a_0}\,\xi_2.$$

The estimate equation for $x(0,t)$, analogous to (20), can be obtained by combining (20) and (33),

$$\frac{d\hat{x}(0,t)}{dt} = \left[v_t+ v_{a_0}v_0(t,\hat{a}_0(t)) + v_x\dot{\hat{x}}(r,t)+ v_{x_r}\dot{\hat{x}}_r\right]_{r=0}$$

$$+P_{0,ob}h_{x_{ob}}^T\,Q\left(y - h(t,\hat{x}_{ob})\right) \quad (34)$$

where

$$P_{0,ob} \triangleq [P(0,r_1,t),\ P(0,r_2,t),\cdots,\ P(0,r_m,t)]$$

$$\therefore = [P_0(r_1,t)^T,\ P\ (r_2,t)^T,\cdots,\ P_0(r_m,t)^T] \quad (35)$$

Differentiating (31) with respect to t, we obtain

$$\frac{d\dot{\hat{x}}(0,t)}{dt} = v_t + v_{a_0}\frac{d\hat{a}_0}{dt} + v_x\dot{\hat{x}}\Big|_{r=0} + v_{x_r}\dot{\hat{x}}_r\Big|_{r=0} \quad (36)$$

Combining (34) and (36) we obtain the desired equation for $\hat{a}_0$,

$$\frac{d\hat{a}_0}{dt} = w_0(t,\hat{a}_0(t))+\left(v_{a_0}^Tv_{a_0}\right)^{-1}v_{a_0}^T\,P_{0,ob}\,h_{x_{ob}}^T\,Q\left(y-h(t,\hat{x}_{ob})\right) \quad (37)$$

The same procedure can be used with the boundary condition at $r = 1$, (10). In this case we obtain

$$\frac{d\hat{a}_1}{dt} = w_1(t,\hat{a}_1(t))+\left(k_{a_1}^Tk_{a_1}\right)^{-1}k_{a_1}^T\,P_{1,ob}\,h_{x_{ob}}^T\,Q\left(y-h(t,\hat{x}_{ob})\right) \quad (38)$$

where

$$P_{1,ob} \triangleq [P(1,r_1,t),\ P(1,r_2,t),\cdots,\ P(1,r_m,t)] \quad (39)$$

The estimator equations for $x(r,t)$, $a_0(t)$, and $a_1(t)$, (29), (37) and (38), require initial conditions. The initial conditions are generally our best guesses of the initial values of $x(r,t)$, $a_0(t)$ and $a_1(t)$ at $t = 0$. These initial conditions will be denoted $\hat{x}_0(r)$, $\hat{a}_0(0)$ and $\hat{a}_1(0)$, since in general the actual initial conditions $x_0(r)$, a_{00}, and a_{10} are unknown.

Derivation of the Covariance Equations

Now the equation governing $P(r,\rho,t)$ will be determined. For convenience these are termed the covariance equations even though this association is only exact for the linear case. Consider (21) for the i,j^{th} submatrix of P', $P'_{i,j}$,

$$\frac{dP'_{i,j}}{dt} = f_{i,i-1}P'_{i-1,j}+f_{i,i}P'_{i,j}+f_{i,i+1}P'_{i+1,j}+P'_{i,j-1}f_{j,j-1}^T$$

$$+ P'_{i,j}f_{j,j}^T + P'_{i,j+1}f_{j,j+1}^T$$

$$+ \sum_k^m\sum_\ell^m P'_{i,k}\left\{[H_x^T\,Q(y - H(t,\hat{x}))]_x\right\}_{k,\ell}\,P'_{\ell,j}$$

$$+ R_{ij}^{-1} \quad (40)$$

where

$$f_{i,i-1} = \frac{\partial f\left(i\Delta,t,x(i),\ \frac{x(i)-x(i-1)}{\Delta},\ \frac{x(i+1)-2x(i)+x(i-1)}{\Delta^2}\right)}{\partial x(i-1)} \quad (41)$$

which becomes

$$f_{i,i-1} = \frac{\partial f}{\partial x_r}\left(\frac{-1}{\Delta}\right)+\frac{\partial f}{\partial x_{rr}}\left(\frac{1}{\Delta^2}\right)$$

$$= - f_{x_r}(i)N + f_{x_{rr}}(i)N^2 \quad (42)$$

In a similar manner we obtain,

$$f_{i,i} = f_x(i) + f_{x_r}(i)N - 2f_{x_{rr}}(i)N^2 \quad (43)$$

$$f_{i,i+1} = f_{x_{rr}}N^2 \quad (44)$$

Then (40) becomes

$$\frac{dP'_{i,j}}{dt} = f_x(i)P'_{i,j}+P'_{i,j}f_x^T(j)+ f_{x_r}(i)\left(\frac{P'_{i,j}- P'_{i-1,j}}{\Delta}\right) +$$

$$\left(\frac{P'_{i,j}- P'_{i,j-1}}{\Delta}\right)f_{x_r}^T(j) + f_{x_{rr}}(i)\left(\frac{P'_{i+1,j}- 2P'_{i,j} + P'_{i-1,j}}{\Delta^2}\right)$$

$$+ \left(\frac{P'_{i,j+1} - 2P'_{i,j} + P'_{i,j-1}}{\Delta^2}\right)f_{x_{rr}}^T(j) + \sum_k^m\sum_\ell^m P'_{i,k}$$

$$\left\{[H_x^T\,Q\left(y - H(t,x(t))\right)]_x\right\}_{k,\ell}\,P'_{\ell,j} + R_{ij}^{-1} \quad (45)$$

We now let $N \to \infty$, $\Delta \to 0$, noting that

$$\lim_{\substack{N\to\infty\\ \Delta\to 0}}\frac{dP'_{i,j}}{dt} = \frac{\partial P(r,\rho,t)}{\partial t} \quad (46)$$

Performing the limiting operation on (45), we obtain

$$\frac{\partial P(r,\rho,t)}{\partial t}$$

$$= f_x(r)\,P(r,\rho,t) + P(r,\rho,t)f_x^T(\rho) + f_{x_r}(r)\,\frac{\partial P(r,\rho,t)}{\partial r}$$

$$+ \frac{\partial P(r,\rho,t)}{\partial \rho}\,f_{x_r}^T(\rho)$$

$$+ f_{x_{rr}}(r)\frac{\partial^2P(r,\rho,t)}{\partial r^2} + \frac{\partial^2P(r,\rho,t)}{\partial \rho^2}\,f_{x_{rr}}^T(\rho)$$

(Eq. con't)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$+ \sum_{k}^{n} \sum_{\ell}^{n} P(r,r_k,t) A_{k,\ell} P(r_\ell,\rho,t) + R^{-1}(r,\rho) \tag{47}$$

where $A_{k,\ell}$ is the n x n-dimensional submatrix defined by

$$A_{k,\ell} = \left([h^T_{x_{ob}} Q \left(y - h(t,\hat{x}_{ob}(t)) \right)]_{x_{0_b}} \right)_{k,\ell} \tag{48}$$

Now we consider the boundary conditions for (47). In order to derive these conditions we will consider two cases: (1) No noise in the boundary conditions, i.e. $\zeta_2(t) = \zeta_3(t) = 0$; (2) Noisy boundary conditions as represented in (4) and (7). We treat first the case of deterministic boundary conditions.

Let us use the following identities,

$$\dot{x}(0,t) = f(t,r,x,x_r,x_{rr})\Big|_{r=0} + g_0(t,a_0,x,x_r)\Big|_{r=0} \tag{49}$$

$$\dot{x}(0,t) = f(t,r,x,x_r,x_{rr})\Big|_{r=0} \tag{50}$$

and

$$\dot{x}(1,t) = f(t,r,x,x_r,x_{rr})\Big|_{r=1} + g_1(t,a_1,x,x_r)\Big|_{r=1} \tag{51}$$

$$\dot{x}(1,t) = f(t,r,x,x_r,x_{rr})\Big|_{r=1} \tag{52}$$

Combining (49) and (47) evaluated at r = 0 and then combining (50) and (47) at r = 0 yields two equations for $\frac{\partial P(0,\rho,t)}{\partial t}$. When these are equated we obtain the r = 0 boundary condition for (47),

$$g_{0_x}(t,\hat{a}_0,\hat{x},\hat{x}_r)\, P(0,\rho,t) + g_{0_{x_r}}(t,\hat{a}_0,\hat{x},\hat{x}_r)\, \frac{\partial P(r,\rho,t)}{\partial r}\Big|_{r=0} = 0 \tag{53}$$

$$\rho \in [0,1]$$

The corresponding boundary condition at r = 1 is

$$g_{1_x}(t,\hat{a}_1,\hat{x},\hat{x}_r)\, P(1,\rho,t) + g_{1_{x_r}}(t,\hat{a}_1,\hat{x},\hat{x}_r)\, \frac{\partial P(r,\rho,t)}{\partial r}\Big|_{r=1} = 0 \tag{54}$$

$$\rho \in [0,1]$$

In the linear case these conditions can be shown to reduce to those of the linear filters (11). For the special case in which g_0 is independent of x_r, the proper boundary condition for (47) is obtained from (53) as

$$P(0,\rho,t) = 0 \tag{55}$$

Similarly, if g_1 is independent of x_r,

$$P(1,\rho,t) = 0 \tag{56}$$

Also, if g_0 and g_1 are independent of x, i.e. depend only on a_0, a_1 , and x_r, the boundary conditions from (53) and (54) are

$$\frac{\partial P(r,\rho,t)}{\partial r}\Big|_{\substack{r=0 \\ r=1}} = 0 \tag{57}$$

Now let us consider the case in which the boundary conditions of the system contain noisy inputs $a_0(t)$ and $a_1(t)$, i.e. $\zeta_2(t) \neq 0$, $\zeta_3(t) \neq 0$. In this case we will distinguish two subcases: (a) g_0 and g_1 do not depend on x_r, i.e. $g_0(t,a_0,x) = 0$ at r = 0 and $g_1(t,a_1,x) = 0$ at r = 1; (b) g_0 and g_1 depend on x_r as in (3) and (6). Since P is a function of r and ρ, where r∈[0,1] and ρ∈[0,1], and since the system (1) may in general be second order in r, we require two boundary conditions on r and two boundary conditions on ρ for (47). These conditions can be written as follows:

$$\rho = 0 \qquad P(r,0,t) \triangleq P_0(r,t) \tag{58}$$

$$\rho = 1 \qquad P(r,1,t) = P_1(r,t) \tag{59}$$

From (28) we see that $P(r,0,t) = P(0,r,t)^T$ and $P(r,1,t) = P(1,r,t)^T$ so that r = 0 and r = 1 boundary conditions can be obtained directly from (58) and (59). Let us now derive the equations satisfied by $P_0(r,t)$ and $P_1(r,t)$.

We consider subcase (a) first and rewrite (3) in the form

$$a_0(t) = \tilde{g}_0(t,a_0(t),x) \tag{60}$$

Differentiating (60) with respect to t we obtain

$$\dot{x}(0,t) = \tilde{g}_{0_x}^{-1} \left\{ (I - \tilde{g}_{0_{a_0}}) \left[w_0(\tilde{g}_0(t,a_0,x)) + \zeta_2(t) \right] - \tilde{g}_{0_t} \right\} \tag{61}$$

Thus, if $\dot{x}(0,t) = f(0) + \zeta_1(0,t)$, we obtain

$$f(0) = \tilde{g}_{0_x}^{-1} \left\{ (I - \tilde{g}_{0_{a_0}})\, w_0(\tilde{g}_0(t,a_0,x)) - \tilde{g}_{0_t} \right\} \triangleq b_0(t,a_0,x) \tag{62}$$

Similarly, letting $a_1(t) = \tilde{g}_1(t,a_1(t),x)$, the relation at r = 1 analogous to (62) is

$$f(1) = \tilde{g}_{1_x}^{-1} \left\{ (I - \tilde{g}_{1_{a_1}})\, w_1(\tilde{g}_1(t,a_1,x)) - \tilde{g}_{1_t} \right\} \triangleq b_1(t,a_1,x) \tag{63}$$

Evaluating (47) at r = 0 and r = 1 and using (62) and (63), the following equations are obtained for $P_0(r,t)$ and $P_1(r,t)$,

$$\frac{\partial P_0(r,t)}{\partial t} = f_x(r) P_0(r,t) + P_0(r,t) b^T_{0_x}$$

(Eq. con't).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$+ f_{x_r}(r)\frac{\partial P_0(r,t)}{\partial r} + f_{x_{rr}}(r)\frac{\partial^2 P_0(r,t)}{\partial r^2}$$

$$+ \sum_k^m \sum_\ell^m P(r,r_k,t)\, A_{k,\ell}\, P_0(r_\ell,t) + R^{-1}(r,0) \tag{64}$$

$$\frac{\partial P_1(r,t)}{\partial t} = f_x(r)\, P_1(r,t)$$

$$+ P_1(r,t) b_{1_x}^T + f_{x_r}(r)\frac{\partial P_1(r,t)}{\partial r} + f_{x_{rr}}(r)\frac{\partial^2 P_1(r,t)}{\partial r^2}$$

$$+ \sum_k^m \sum_\ell^m P(r,r_k,t)\, A_{k,\ell}\, P_1(r_\ell,t) + R^{-1}(r,1) \tag{65}$$

We now need boundary conditions for (64) and (65) at $r = 0$ and $r = 1$. These are obtained by evaluating (64) and (65) at $r = 0$ and $r = 1$ and are denoted by

$$r = 0 \quad P_0(0,t) \triangleq P_{00}(t) \quad P_1(0,t) \triangleq P_{01}(t) \tag{66}$$

$$r = 1 \quad P_0(1,t) \triangleq P_{01}(t)^T \quad P_1(1,t) \triangleq P_{11}(t) \tag{67}$$

$P_{00}(t)$, $P_{01}(t)$ and $P_{11}(t)$ are governed by

$$\frac{dP_{00}(t)}{dt} = b_{0_x} P_{00}(t) + P_{00}(t)\, b_{0_x}^T$$

$$+ \sum_k^m \sum_\ell^m P_0(r_k,t)^T A_{k,\ell}\, P_0(r_\ell,t) + R^{-1}(0,0) \tag{68}$$

$$\frac{dP_{01}(t)}{dt} = b_{0_x} P_{01}(t) + P_{01}(t)\, b_{1_x}^T$$

$$+ \sum_k^m \sum_\ell^m P_0(r_k,t)^T A_{k,\ell}\, P_1(r_\ell,t) + R^{-1}(0,1) \tag{69}$$

$$\frac{dP_{11}(t)}{dt} = b_{1_x} P_{11}(t) + P_{11}(t)\, b_{1_x}^T$$

$$+ \sum_k^m \sum_\ell^m P_1(r_k,t)^T A_{k,\ell}\, P_1(r_\ell,t) + R^{-1}(1,1) \tag{70}$$

Now we consider subcase (b) in which g_0 and g_1 may contain x_r. Using (9) and (10) we can write

$$\dot{x}(0,t) = f\left(t, v(t,a_0,x,x_r), x_r, x_{rr}\right)\Big|_{r=0} + \xi_1(0,t) \tag{71}$$

$$\dot{x}(1,t) = f\left(t, k(t,a_1,x,x_r), x_r, x_{rr}\right)\Big|_{r=1} + \xi_1(1,t) \tag{72}$$

Combining (71) and (72) with (47) we obtain the boundary conditions for (47) as

$$\frac{\partial P_0(r,t)}{\partial t} = f_x(r)\, P_0(r,t)$$

$$+ f_{x_r}(r)\frac{\partial P_0(r,t)}{\partial r} + f_{x_{rr}}(r)\frac{\partial^2 P_0(r,t)}{\partial r^2}$$

$$+ \left[P_0(r,t) v_x^T(\rho) f_x^T(\rho) + \frac{\partial P(r,\rho,t)}{\partial \rho}\left(f_{x_r}^T(\rho) + v_{x_r}^T(\rho) f_x^T(\rho)\right)\right.$$

$$\left. + \frac{\partial^2 P(r,\rho,t)}{\partial \rho^2} f_{x_{rr}}^T(\rho)\right]_{\rho=0} + \sum_k^m \sum_\ell^m P(r,r_k,t) A_{k,\ell} P_0(r_\ell,t) + R^{-1}(r,0) \tag{73}$$

$$\frac{\partial P_1(r,t)}{\partial t} = f_x(r) P_1(r,t)$$

$$+ f_{x_r}(r)\frac{\partial P_1(r,t)}{\partial r} + f_{x_{rr}}(r)\frac{\partial^2 P_1(r,t)}{\partial r^2}$$

$$+ \left[P(r,\rho,t) k_x^T(\rho) f_x^T(\rho) + \frac{\partial P(r,\rho,t)}{\partial \rho}\left(f_{x_r}^T(\rho) + k_{x_r}^T(\rho) f_x^T(\rho)\right)\right.$$

$$\left. + \frac{\partial^2 P(r,\rho,t)}{\partial \rho^2} f_{x_{rr}}^T(\rho)\right]_{\rho=1} + \sum_k^m \sum_\ell^m P(r,r_k,t) A_{k,\ell} P_1(r_\ell,t) + R^{-1}(r,1) \tag{74}$$

$$\frac{dP_{00}(t)}{dt} = \left[f_x(r) v_x(r) P_0(r,t)\right.$$

$$\left. + \left(f_{x_r}(r) + f_x(r) v_{x_r}(r)\right)\frac{\partial P_0(r,t)}{\partial r} + f_{x_{rr}}\frac{\partial^2 P_0(r,t)}{\partial r^2}\right]_{r=0}$$

$$+ \left[P_0^T(r,t) v_x^T(r) f_x^T(r) + \frac{\partial P_0^T(r,t)}{\partial r}\left(f_{x_r}^T(r) + v_{x_r}^T(r) f_x^T(r)\right)\right.$$

$$\left. + \frac{\partial^2 P_0^T(r,t)}{\partial r^2} f_{x_{rr}}^T(r)\right]_{r=0}$$

$$+ \sum_k^m \sum_\ell^m P_0^T(r_k,t) A_{k,\ell} P_0(r_\ell,t) + R(0,0)^{-1} \tag{75}$$

$$\frac{dP_{01}(t)}{dt} = \left[f_x(r) v_x(r) P_1(r,t)\right.$$

$$\left. + \left(f_{x_r}(r) + f_x(r) v_{x_r}(r)\right)\frac{\partial P_1(r,t)}{\partial r}\right.$$

(Eq. con't)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

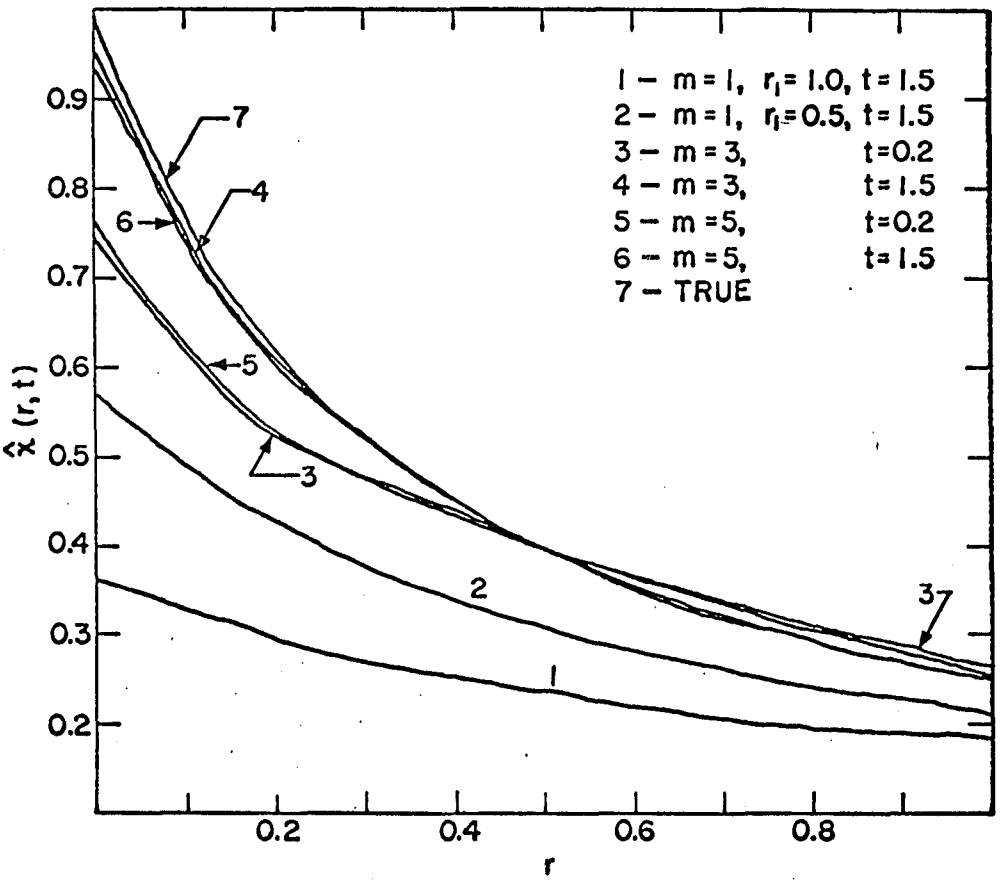

Fig. 1 $\hat{x}(r,t)$ for various values of m for the steady state profile. Observation error only.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$+ f_{x_{rr}}(r)\,\frac{\partial^2 P_1(r,t)}{\partial r^2}\Big]_{r=0} + \Big[P_0^T(r,t)k_x^T(r)f_x^T(r)$$

$$+ \frac{\partial P_0^T(r,t)}{\partial r}\left(f_{x_r}^T(r) + k_{x_r}^T(r)f_x^T(r)\right)$$

$$+ \frac{\partial^2 P_0^T(r,t)}{\partial r^2}\,f_{x_{rr}}^T(r)\Big]_{r=1}$$

$$+ \sum_k^m \sum_\ell^m P_0^T(r_k,t)\,A_{k,\ell}P_1(r_\ell,t) + R^{-1}(0,1) \qquad (76)$$

$$\frac{dP_{11}(t)}{dt} = \Big[f_x(r)\,k_x(r)P_1(r,t)$$

$$+ \left(f_{x_r}(r) + f_x(r)k_{x_r}(r)\right)\frac{\partial P_1(r,t)}{\partial r}$$

$$+ f_{x_{rr}}(r)\,\frac{\partial^2 P_1(r,t)}{\partial r^2}\Big]_{r=1} + \Big[P_1(r,t)\,k_x^T(r)\,f_x^T(r)$$

$$+ \frac{\partial P_1^T(r,t)}{\partial r}\left(f_{x_r}^T(r) + k_{x_r}^T(r)\,f_x^T(r)\right)$$

$$+ \frac{\partial^2 P_1^T(r,t)}{\partial r^2}\,f_{x_{rr}}^T(r)\Big]_{r=1}$$

$$+ \sum_k^m \sum_\ell^m P_1^T(r_k,t)\,A_{k,\ell}\,P_1(r_\ell,t) + R^{-1}(1,1) \qquad (77)$$

For a system with n state variables the distributed filter consists of the following number of equations:

$\hat{x}(r,t)$	n	P.D.E.
$\hat{a}_0(t)$	ℓ_0	O.D.E.
$\hat{a}_1(t)$	ℓ_1	O.D.E.
$P(r,\rho,t)$	n^2	P.D.E.
$P_0(r,t)$	n^2	P.D.E.
$P_1(r,t)$	n^2	P.D.E.
$P_{00}(t)$	$n(n+1)/2$	O.D.E.
$P_{01}(t)$	n^2	O.D.E.
$P_{11}(t)$	$n(n+1)/2$	O.D.E.

We note that the initial condition required for (47), namely

$$P(r,\rho,0) = P^0(r,\rho) \qquad r,\rho \in [0,1] \qquad (78)$$

when specified provides values for $P_0(r,0)$, $P_1(r,0)$, $P_{00}(0)$, $P_{01}(0)$ and $P_{11}(0)$. The choice of $P^0(r,\rho)$ is arbitrary, however in the linear case $P^0(r,\rho) = E\{(x(r,0) - \hat{x}_0(r))(x(\rho,0) - \hat{x}_0(\rho)^T)$ which might be chosen from the degree of knowledge of the initial state $x_0(r)$.

The present results are related to systems described by (1). However, a much wider class of distributed systems can be treated by a simple change of variable. Consider the distributed system defined by

$$\psi(x_t,x_{tr},x_{tt}) = f(t,r,x,x_r,x_{rr}) \qquad (79)$$

If we let $z(r,t) = x_t(r,t)$ then (79) becomes

$$\phi(z,z_r,z_t) = f(t,r,x,x_r,x_{rr}) \qquad (80)$$

If z_t appears non-transcendentally in (80), (80) is now in the form (1). We note, however, that this procedure cannot be used for systems of the form (79) if the left side of which is $\psi(x_t,x_{tr})$ only.

EXAMPLE

We consider the problem of estimating the time-dependent concentration distribution in an isothermal, plug-flow chemical reactor with a second order irreversible reaction, based on noisy measurements at finite locations along the reactor. Let x(r,t) be the dimensionless concentration of component A at time t and position r in the reactor in which A decomposes according to the second order reaction, $2A \rightarrow B + C$.

The dynamic behavior of the reactor in the presence of random excitations in the system and the initial conditions is described by

$$x_t(r,t) + x_r(r,t) = -\beta x(r,t)^2 \qquad (81)$$

$$x(r,0) = x_{ss}(r) \qquad (82)$$

$$x(0,t) = a_0(t) \qquad (83)$$

$$\dot{a}_0(t) = \xi_2(t) \;;\; a_{00} = 1 \qquad (84)$$

The inlet concentration will often not be exactly a_{00} but will fluctuate because of variations upstream of the reactor. These inlet condition fluctuations are included by the error $\xi_2(t)$ in (84) which produces a noisy $a_0(t)$.

Finally, the observations which consist of direct measurements of x at various locations r_i, $i = 1,2,\cdots,$ m are corrupted with additive experimental errors, $n(t,r_i)$.

The filtering problem can be posed as follows: Estimate the concentration distribution, x(r,t), based on noisy measurements of x(r,t) carried out a m locations along the reactor for the system governed by (81)-(84) and

$$y(r_i,t) = x(r_i,t) + n_i(t),\; i = 1,2,\cdots, m \qquad (85)$$

The filter is obtained from (29), (37), (38), (47), (64), and (68).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

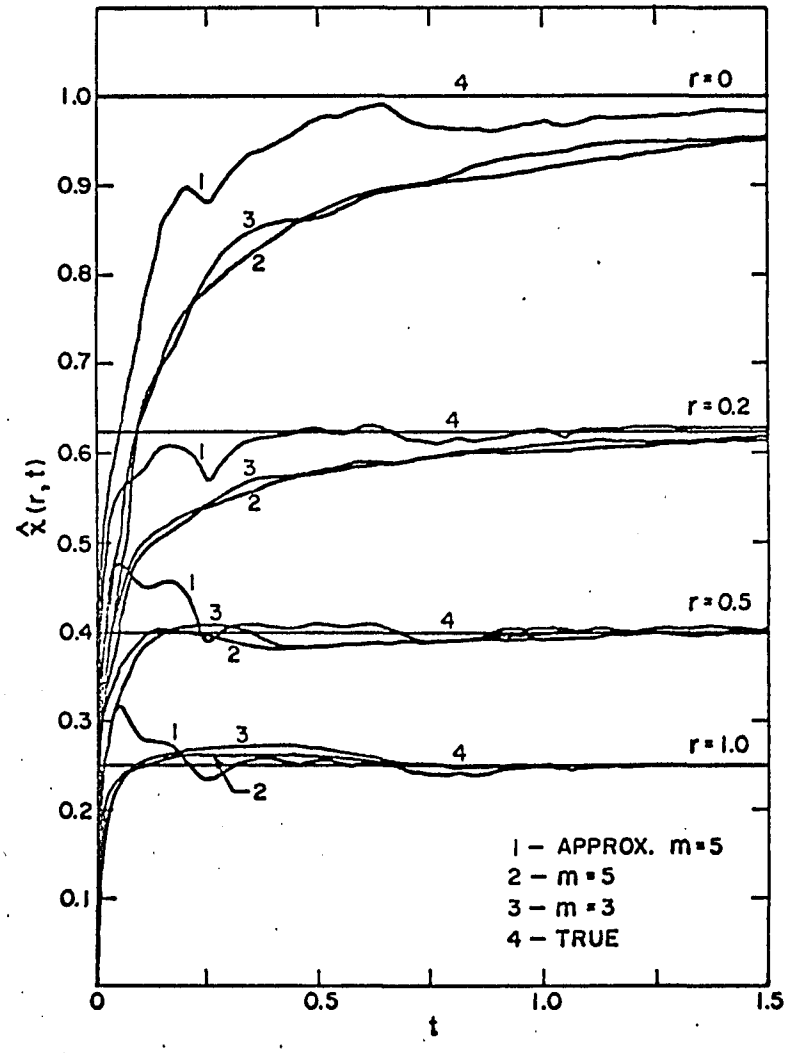

Fig. 2 Comparison of convergence of the full filter with the approximate filter at various spatial locations for the steady state profile. Observation error only.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\hat{x}_t(r,t) + \hat{x}_r(r,t) = -\beta\hat{x}(r,t)^2 + \sum_{i=1}^{m} P(r,r_i,t)[y(r_i,t) - \hat{x}(r_i,t)] \quad (86)$$

$$\hat{x}(0,t) = \hat{a}_0(t) \quad (87)$$

$$\frac{d\hat{a}_0}{dt} = \sum_{i=1}^{m} P_0(r_i,t)\,[y(r_i,t) - \hat{x}(r_i,t)] \quad (88)$$

$$P_t(r,\rho,t) + P_r(r,\rho,t) + P_\rho(r,\rho,t) = -2\beta[\hat{x}(r,t)\,P(r,\rho,t) + \hat{x}(\rho,t)\,P(r,\rho,t)] - \sum_{i=1}^{m} P(r,r_i,t)\,P(r_i,\rho,t) \quad (89)$$

$$P_{0_t}(r,t) + P_{0_r}(r,t) = -2\beta\hat{x}(r,t)\,P_0(r,t) - \sum_{i=1}^{m} P_0(r_i,t)\,P(r,r_i,t) \quad (90)$$

$$\frac{dP_{00}}{dt} = -\sum_{i=1}^{m} P_0^2(r_i,t) + R^{-1} \quad (91)$$

where Q has been taken as one. The initial conditions for (86), (88) and (89)-(91), $\hat{x}_0(r)$, $\hat{a}_0(0)$, and $P^0(r,\rho)$, are arbitrary and must be specified a priori. The errors $\xi_2(t)$ and $\eta_i(t)$ were simulated by 10 G(0,1) and 0.1 G(0,1), respectively, where G(0,1) is a normally distributed random variable with zero mean and unit standard deviation. It was assumed that the actual initial state, $x_{ss}(r)$, was unknown. The initial condition, $\hat{x}_0(r)$, was taken as a constant $\hat{x}_0$. We expect the maximum value of $P^0(r,\rho)$ to occur at $r = \rho$ with monotonically decreasing values as $|r - \rho|$ increases. Thus, the following condition was used,

$$P^0(r,\rho) = 10\exp[-|r - \rho|] \quad (92)$$

Numerical computations were carried out for the steady state input and a ramp input given by

$$\begin{aligned} \dot{a}_0(t) &= 3 + \xi_2(t) \quad 0 \le t \le 0.1 \\ &= \xi_2(t) \quad t \le 0.1 \end{aligned} \quad (93)$$

In the study, $\beta = 3$, $\hat{x}_0(r) = 0$, $\hat{a}_0(0) = 0$ and $R = 10$.

The filtering results are shown in Figs. 1 and 2 for the steady state input (83) and (84) and $m = 1,3,5$. Thus, the problem is to estimate the steady state concentration profile in the reactor from noisy measurements at various locations along the reactor when the steady state input contains time-dependent fluctuations. The inlet fluctuations cause the reactor concentration to be continuously time-varying. In each case $r_1 = 0$ and $r_m = 1$, with intermediate measurements equally spaced. Fig. 1 shows the rate of convergence of $\hat{x}(r,t)$ from the initial guess of $\hat{x}_0 = 0$ to the actual steady state profile, shown by curve 7, for $m = 1$ and $m = 3$. The initial guess of $\hat{x}_0 = 0$ was chosen as representing one that reflects no knowledge of the actual state of the system. The rate of convergence for $m = 1$ is slow, as evidenced by curves 1 and 2. The rate of convergence is increased significantly with the addition of more measurement locations, as seen by curves 3 - 6. It is interesting to note that while the filter converges much more rapidly for $m = 3$ than for $m = 1$, additional measurements to $m = 5$ do not significantly improve convergence over $m = 3$. When $m = 1$ better convergence is obtained for $r_1 = 0.5$ than for $r_1 = 1.0$. This is simply a result of the fact that information reaches $r = 0.5$ twice as soon as $r = 1.0$ because of the hyperbolic nature of the system.

Fig. 2 shows the rate of convergence of $\hat{x}(r,t)$ at four locations for $m = 3$ and 5. For both values of m, $\hat{x}(0,t)$ converges most slowly because of the combined effect of the inlet distrubances and the measurement errors at $r = 0$. Filtering results for the ramp input are shown in Fig. 3 for $m = 3$ and 5 at the same four locations as in Fig. 2. In this case, the inlet concentration undergoes a definite change, as shown by the $r = 0$ curve. Because of the second order reaction the ramp change is attenuated at increasing values of r. Again, due to the large change at $r = 0$, $\hat{x}(0,t)$ converges slowest.

Obviously, the dimensionality of the filter becomes a severe problem as n increases. For real time application to large systems it is thus desirable to eliminate some of the equations by appropriate approximations. One possible approximation is to assume the form of the solution of the equation for $P(r,\rho,t)$. In constructing the approximation we expect that $P(r,\rho,t)$ should be a maximum at $r = \rho$. In addition, $P(r,\rho,t)$ will decay from its initial distribution, approximately in an exponential manner. Thus it was assumed that

$$P(r,\rho,t) = C\exp(-\lambda t)\exp(-\mu|r - \rho|) \quad (94)$$

where C, λ and μ are constants chosen such that the filter converges.

In the present example, the following values were chosen: $C = 10$, $\lambda = 2$, $\mu = 1$. The results of filtering using (94) are shown in Fig. 2. As shown in Fig. 2 by curve 1 we see that convergence actually improved with (94) instead of (89). This can only be attributed to the particular numerical values used for the parameters in (94). The computing time required for the approximate filter was reduced to 1/3 of the time required for the full filter. Although specification of the parameters in an approximation for $P(r,\rho,t)$ may be difficult, such approximations have promise for applications to higher dimension systems. The computing time for the $m = 3$ case using the full filter to $T = 1.6$ was 3 minutes on an IBM 360/75.

SUMMARY

A general nonlinear filter has been derived for distributed parameter systems that contain noisy dynamical inputs in the system and boundary conditions and measurement errors. The filter was applied to estimate the state in a nonlinear hyperbolic system. An approximation to the solution of the covariance equations was found to be highly effective in reducing the amount of computing required for the filter. The question of convergence of the filter was not studied, a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

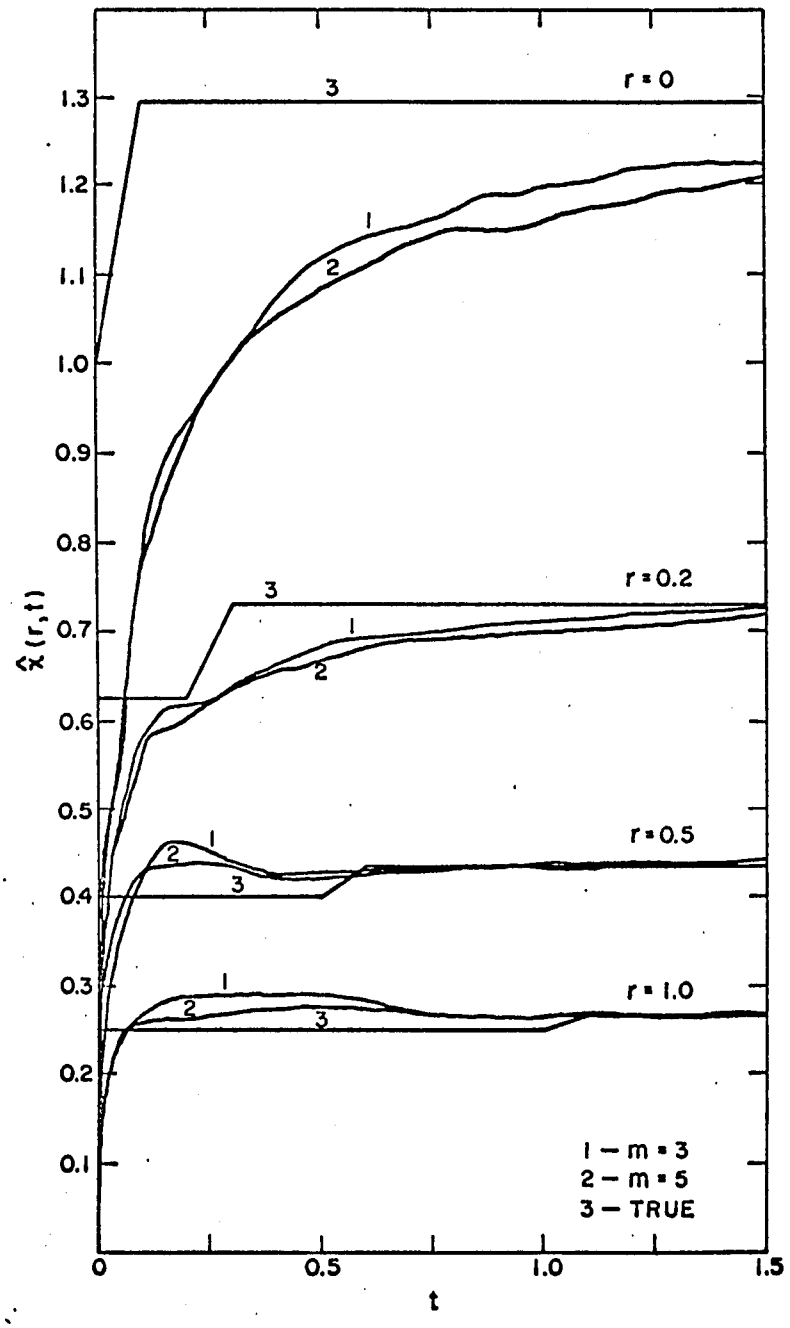

Fig. 3 $\hat{x}(r,t)$ for ramp input. Input and observation errors.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

question which is related to the observability of the state based on the noise-free observations. This point is currently under investigation.

REFERENCES

[1] Balakrishnan, A.V., "State Estimation for Infinite Dimensional Systems," Journal of Computer and System Science, Vol. 1, 1967, pp. 391-403.

[2] Detchmendy, D.M. and Sridhar, R., "Sequential Estimation of States and Parameters in Noisy Nonlinear Systems," Journal of Basic Engineering, Trans. ASME, Series D. Vol. 88, No. 2, June 1966, pp. 362-368.

[3] Kwakernaak, H., "Optimal Filtering in Linear Systems with Time Delays," I.E.E.E. Trans. on Auto. Control, Vol. AC-12, No. 2, April 1967, pp. 169-173.

[4] Meditch, J.S., "On State Estimation for Distributed Parameter Systems," Boeing Scientific Research Lab Raport D1-82-0917, Sept. 1969.

[5] Pell, T.M., Jr., "Some Problems in Chemical Reactor Analysis with Stochastic Features," Ph.D. Thesis, Department of Chemical Engineering, University of Minnesota, Feb. 1969.

[6] Pell, T.M., Jr., and Aris, R., "Some Problems in Chemical Reactor Analysis with Stochastic Features. II. Control of Linearized Distributed Systems on Discrete and Corrupted Observations," Industrial and Engineering Chemistry Fundamentals, Vol. 9, No. 1, 1970, pp. 15-20.

[7] Schwartz, L. and Stear, E.B., "A Computational Comparison of Several Nonlinear Filters," I.E.E.E. Trans. on Auto. Control, Vol. AC-13, No. 1, Feb. 1968, pp. 83-86.

[8] Seinfeld, J.H., "Nonlinear Estimation for Partial Differential Equations," Chemical Engineering Science, Vol. 24, No. 1, Jan. 1969, pp. 75-83.

[9] Thau, F.E., "On Optimal Filtering for a Class of Linear Distributed-Parameter Systems," Journal of Basic Engineering, Trans. ASME, Series D. Vol. 91, No. 2, June 1969, pp. 173-178.

[10] Tzafestas, S.G. and Nightingale, J.M., Optimal Filtering Smoothing and Prediction in Linear Distributed Parameter Systems, Proc. I.E.E., Vol. 115, No. 8, Aug. 1968, pp. 1207-1212.

[11] Tzafestas, S.G. and Nightingale, J.M., "Concerning Optimal Filtering Theory of Linear Distributed Parameter Systems," Proc. I.E.E., Vol. 115, No. 11, Nov. 1968, pp. 1737-1742.

[12] Tzafestas, S.G. and Nightingale, J.M., "Maximum Likelihood Approach to the Optimal Filtering of Distributed Parameter Systems, Proc. I.E.E. Vol.116, No. 6, June 1969, pp. 1085-1093.

Acknowledgment

This work was supported by National Science Foundation Grant GK 10136.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

8. Notation

a = constant parameter

$a(t)$ = ℓ_1-dimensional vector

$A(t,a)$ = ℓ_1-dimensional vector function

$b(t)$ = ℓ_2-dimensional vector

$B(t,b)$ = ℓ_2-dimensional vector function

c = constant

d = ℓ_2-dimensional vector

e = ℓ_1-dimensional vector

$E\{\cdot\}$ = expectation operation

f = n-dimensional vector function

g = s-dimensional vector function

$G(a,b)$ = Gaussian distribution with mean a and standard deviation b

h = m-dimensional vector function

I = performance index

J = performance index

P = covariance matrix of filtering estimation error

Q = weighting matrix

r = independent variable

R = weighting matrix

s = independent variable

t, t_1, T = time variable

u = control vector

W = covariance matrix of the interpolation estimation error

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

x = n-dimensional state vector

y = m-dimensional observation vector

z = n-dimensional vector

Greek Symbols

$\alpha(\cdot)$ = n-dimensional vector function

$\beta(\cdot)$ = n-dimensional vector function

$\gamma(\cdot)$ = ℓ_1-dimensional vector function

$\delta(t)$ = Dirac delta function

ζ = dummy variable

η = m-dimensional observation error

$\theta(\cdot)$ = ℓ_2-dimensional vector function

$\lambda(r,t)$ = n-dimensional adjoint variable

$\mu(t)$ = s-dimensional adjoint variable

ν = dummy variable

ξ = dynamical noise vector

$\rho(\cdot)$ = ℓ_1-dimensional vector function

$\sigma(t)$ = ℓ_2-dimensional adjoint variable

τ = dummy variable

$\phi(\cdot)$ = final state vector for the generalized case defined in Equation (3.29)

$\psi(\cdot)$ = state vector at $t = t_1$ for the generalized case defined in Equation (4.1)

Superscripts

$\hat{}$ = estimated value

$*$ = optimal value

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Subscripts

o	= initial or at r = 0
1	= end or at r = 1

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter IV

OBSERVABILITY OF NONLINEAR SYSTEMS

1. Introduction

A question fundamental to the analysis of physical systems is whether the state of the system can be uniquely determined from its output data. Specifically, given the dynamic description of the system and the observation process, we can ask under what conditions can the initial state of the system be determined uniquely on the basis of the observed output on a given time interval. This problem is called the inverse or observability problem. The test of a system's observability is a necessary prerequisite to the estimation of states and parameters from the output of the system.

In this study we consider the problem of determining conditions for the observability of the initial state and a vector of constant parameters in systems governed by nonlinear ordinary differential equations. New necessary and sufficient conditions are obtained for local observability in the neighborhood of a given value of the initial state and a given value of the parameter vector. In addition, a technique is presented whereby the local observability results can be used to study the entire domain of initial conditions and parameter values. The local approach is based on the extension of the necessary and sufficient conditions for observability of time-varying linear systems to nonlinear systems.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. Review of Linear Observability Theory

Definition 2.1. A state x_o is observable at t_o if, given any control $u(t)$ and the output $y(t)$, $t_o \le t \le T$, x_o can be determined uniquely. If every state x_o is observable at t_o, then we say the system is observable at t_o.

Theorem 2.1 [71] The process

$$\dot{x}(t) = A(t)\, x(t) + B(t)\, u(t) \tag{2.1}$$

$$y(t) = H(t)\, x(t) \tag{2.2}$$

with $x \in R^n$, $u \in R$, and $y \in R^m$, is observable at t_o if and only if the symmetric matrix

$$M(t_o,t_1) = \int_{t_o}^{t_1} \Phi^T(t,t_o)\, H^T(t)\, H(t)\, \Phi(t,t_o)dt \tag{2.3}$$

is positive definite for some t_1, $t_o \le t_1 \le T$, where

$$\frac{\partial \Phi(t,\zeta)}{\partial t} = A(t)\, \Phi(t,\zeta) \tag{2.4}$$

$$\Phi(t,t) = I \tag{2.5}$$

Theorem 2.2 [33] The system of (2.1) and (2.2) is completely controllable if and only if the symmetric matrix

$$W(t_o,t_1) = \int_{t_o}^{t_1} \Phi(t_o,t)\, B(t)\, B^T(t)\, \Phi^T(t_o,t)dt \tag{2.6}$$

is positive definite for some t_1, $t_1 > t_o$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Theorem 2.3 [71] The range of both $M(t_o,t)$ and $W(t_o,t)$, $t > t_o$ is monotone nondecreasing with increasing t.

Now let us extend the problem slightly by considering the conditions for observability of both x_o and a vector of constant parameters p, $p \in R^{\ell}$, in the modified form of (2.1) and (2.2)

$$\dot{x}(t) = A(t)\, x(t) + B(t)\, p \tag{2.7}$$

$$y(t) = H(t)\, x(t)$$

We want to determine the necessary and sufficient conditions for the observability of both x_o and p. These conditions are stated in Theorem 2.4.

Theorem 2.4 The initial state x_o and the parameter vector p in (2.7) and (2.8) are observable if and only if the symmetric matrix $K(t_o,t_1)$ is positive definite for some t_1, $t_o \leq t_1 \leq T$, where

$$K(t_o,t_1) = \begin{bmatrix} K_{11}(t_o,t_1) & K_{12}(t_o,t_1) \\ K_{21}(t_o,t_1) & K_{22}(t_o,t_1) \end{bmatrix} \tag{2.9}$$

and

$$K_{11}(t_o,t_1) = \int_{t_o}^{t_1} \Phi^T(t,t_o)\, H^T(t)\, H(t)\, \Phi(t,t_o)\, dt \tag{2.10}$$

$$K_{12}(t_o,t_1) = \int_{t_o}^{t_1} \Phi^T(t,t_o)\, H^T(t)\, H(t)\, \Phi(t,t_o)\, q(t)\, dt \tag{2.11}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$K_{21}(t_o,t_1) = \int_{t_o}^{t_1} q^T(t)\ \Phi^T(t,t_o)\ H^T(t)\ H(t)\ \Phi(t,t_o)\ dt \qquad (2.12)$$

$$K_{22}(t_o,t_1) = \int_{t_o}^{t_o} q^T(t)\ \Phi^T(t,t_o)\ H^T(t)\ H(t)\ \Phi(t,t_o)\ q(t)dt \qquad (2.13)$$

$$q(t) = \int_{t_o}^{t} \Phi(t_o,\zeta)\ B(\zeta)\ d\zeta \qquad (2.14)$$

The proof of the necessary part of Theorem 2.4 proceeds as follows. The solution of (2.7) is

$$x(t) = \Phi(t,t_o)\ x_o + \int_{t_o}^{t} \Phi(t,\zeta)\ B(\zeta)\ p\ d\zeta \qquad (2.15)$$

Substituting into (2.8)

$$y(t) = H(t)\ \Phi(t,t_o)x_o + \int_{t_o}^{t} H(t)\ \Phi(t,\zeta)\ B(\zeta)\ p\ d\zeta \qquad (2.16)$$

which can be rewritten as

$$y(t) = [H(t)\ \Phi(t,t_o),\ H(t)\ \Phi(t,t_o)\ q(t)] \begin{bmatrix} x_o \\ p \end{bmatrix} \qquad (2.17)$$

Solving for x_o and p we obtain

$$\begin{bmatrix} x_o \\ p \end{bmatrix} = K(t_o,t_1)^{-1} \int_{t_o}^{t} [H(t)\ \Phi(t,t_o),\ H(t)\Phi(t,t_o)q(t)]^T\ y(t)dt \qquad (2.18)$$

Hence, the positive definiteness of $K(t_o,t_1)$ for some t_1, $t_o < t_1 \leq T$ is a necessary condition for the existence of a unique

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(x_o,p) . The proof of sufficiency proceeds exactly as for Theorem 2.1 and is presented in references [33,71].

The observability of x_o at t_o only requires the existence of $K_{11}(t_o,t_1)^{-1}$ which is identical to $M(t_o,t_1)^{-1}$ as expected. The observability of p only requires the existence of $K_{22}(t_o,t_1)^{-1}$. Note, however, that because p is time invariant, $K_{22}(t_o,t_1)$ is different from $W(t_o,t_1)$, although the concepts of controllability of (2.1) and observability of p in (2.7) and (2.8) are closely related . The range of $K(t_o,t)$, $t > t_o$ is monotone nondecreasing with increasing t.

For completeness we state the following theorem.

Theorem 2.5. [60] The system of (2.1) and (2.2) is observable on $[t_o,T]$ if and only if $Q(t)$ does not have rank less than n on any subinterval of $[t_o,T]$ where

$$Q(t) = [S_o(t),\ S_1(t),\ \cdots,\ S_{n-1}(t)] \tag{2.19}$$

$$S_{k+1}(t) = [A^T(t) + I\,\frac{d}{dt}]\,S_k(t) \tag{2.20}$$

$$S_o(t) = H^T(t) \tag{2.21}$$

If any state x_o in (2.1) or any state x_o and any parameter vector p in (2.7) are observable at t_o , the systems are observable at t_o .

If $\dot{y}(t)$ is available in addition to $y(t)$, $t_o \le t \le T$, we can modify (2.2) as

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\tilde{y}(t) = \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} H(t) & 0 \\ \dot{H}(t)+H(t)A(t) & H(t)B(t) \end{bmatrix} \begin{bmatrix} x(t) \\ p \end{bmatrix} \tag{2.22}$$

The modified form of (2.17) is

$$\tilde{y}(t) = H(t) \begin{bmatrix} x_o \\ p \end{bmatrix} \tag{2.23}$$

where

$$\tilde{H}(t) = \begin{bmatrix} H(t)\ \Phi(t,t_o) & H(t)\Phi(t,t_o)q(t) \\ (\dot{H}(t)+H(t)A(t))\Phi(t,t_o) & (\dot{H}(t)+H(t)A(t))\Phi(t,t_o)q(t) \\ & \quad + H(t)B(t) \end{bmatrix} \tag{2.24}$$

The necessary and sufficient conditions for observability of x_o and p at t_o when both $y(t)$ and $\dot{y}(t)$ are known are given by Theorem 2.4 with $K(t_o,t_1)$ replaced by

$$\tilde{K}(t_o,t_1) = \int_{t_o}^{t_1} \tilde{H}^T(t)\ \tilde{H}(t)\ dt \tag{2.25}$$

3. Local Observability of Nonlinear Systems

We now consider the class of systems governed by

$$\dot{x}(t) = f(t,x(t),p) \tag{3.1}$$

$$y(t) = h(t,x(t)) \tag{3.2}$$

where $(t,x) \in S \subset R^1 \times R^n$, $p \in R_p \subset R^\ell$, $y \in R_y \subset R^m$ and $t \in [t_o,T]$. We assume that S is compact, R_p is a linear space, and f and $h \in C^1$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In addition, f and h are assumed to have continuous first order partial tial derivatives with respect to their arguments. The observability question, namely, under what conditions can x_o and p be uniquely determined from $y(t)$, $t \in [t_o,T]$, can also be stated as under what conditions there exists a one-to-one correspondence between $(x_o,p) \in W$ and $y(t)$, $t_o \leq t \leq T$, where W is the domain of initial states x_o and parameter vectors p.

Lee and Markus[42] obtained necessary and sufficient conditions for observability of nonlinear systems in the neighborhood of the origin by applying the results for linear, time invariant systems. Roitenberg[55] considered the construction of a Lyapunov function for the linearized system to study the observability of nonlinear systems. Kostyukovskii[35,36] and Griffith and Kumar[23] determined conditions for observability of nonlinear systems from the one-to-one mapping conditions from $x(t) = [x_1(t),\cdots,x_{n+\ell}(t)]$ to $\tilde{y}(t) = [y(t), y^{(1)}(t), \cdots,y^{(n+\ell-1)}(t)]$ where $y^{(i)}(t)$ is the i^{th} order time derivative of $y(t)$. Therefore $\tilde{y}(t)$ is considered as another variable in R^{mn} and it is required that $f \in C^{n+\ell-1}$ and $y \in C^{n+\ell}$. The Jacobian matrix $\partial\tilde{y}/\partial x$ is required, however its nonsingularity is not sufficient, in general, for a unique mapping from R^n to R^{mn}. The approach is somewhat analogous to Theorem 2.5, in that repeated time differentiations of the output are required.

In this study we obtain necessary and sufficient conditions for observability of both x_o and p in the system 3.1 and 3.2. The conditions are obtained for the observability in the neighborhood of (x_o,p_o), $N(x_o,p_o)$, by application of Theorem 2.4 to the linearized

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

trajectory of (3.26) and (3.27) from $x(t_o) = x_o$ and $p = p_o$. Thus, we extend the work of Lee and Markus[42] from local observability about the origin (equilibrium point of the system) to any point in the entire initial condition and parameter domain of (x_o,p). The approach enables the examination of the observability of any (x_o,p) and avoids the stringent differentiability requirements of references [23,35,36].

In order to justify linearization of (3.1) we will first require an embedding theorem for differential equations presented by Hestenes and Guinn[25]. The theorem is quoted without proof.

Theorem 3.1. [25,45] Let $x(t)$, $t_o \leq t \leq T$ be a solution of (3.1) with initial condition x_o and parameter value p_o. For each set of (ξ,p) satisfying the relationships

$$\|\xi - x_o\| < \sigma \qquad \|p - p_o\| < \sigma' \tag{3.3}$$

there is a unique solution $v(t,\xi,p)$, $t_o \leq t \leq T$ of

$$\dot{v}(t) = f(t,v(t),p) \tag{3.4}$$

$$v(t_o) = \xi \tag{3.5}$$

satisfying the inequality

$$\|v(t,\xi,p) - x(t)\| < \varepsilon \qquad t_o \leq t \leq T \tag{3.6}$$

where $\sigma' > 0$, and $G > \sigma'$

$$G = \sup_{0 \leq t \leq T} \left\| \int_{t_o}^{t} \{f(s,x(s),p) - f(s,x(s),p_o)\}\, ds \right\| \tag{3.7}$$

and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\sigma = \frac{\varepsilon}{2} \exp\left[- \int_{t_o}^{T} L(s)\, ds \right] \tag{3.8}$$

$$\|f(t,x,p) - f(t,v,p)\| \leq L(t)\ \|x - v\| \tag{3.9}$$

for all (t,x), $(t,v) \in S$ and all admissible $p \in R_p$.

Theorem 3.1 is crucial to our analysis since it establishes precisely the conditions under which the perturbed trajectory $v(t,\xi,p)$ remains close to $x(t)$, $t_o \leq t \leq T$. Our notion of observability in a neighborhood about (x_o,p_o) will derive its validity from Theorem 3.1.

Let us consider a reference trajectory $x(t)$ with initial condition $x(t_o) = x_o$ and $p = p_o$. Perturbation of x_o and p_o , $\xi = x_o + \delta x_o$ and $p = p_o + \delta p_o$, produces a trajectory $v(t,\xi,p) = x(t) + \delta x(t)$. Then $\delta x(t)$ and $\delta y(t)$ are governed by

$$\delta\dot{x}(t) = A(t)\ \delta x(t) + B(t)\ \delta p_o + 0(\varepsilon,\sigma') \tag{3.10}$$

$$\delta x(t_o) = \delta x_o \tag{3.11}$$

$$\delta y(t) = H(t)\ \delta x(t) + 0(\varepsilon) \tag{3.12}$$

where

$$A(t) = f_x(t,x(t),p_o) \tag{3.13}$$

$$B(t) = f_p(t,x(t),p_o) \tag{3.14}$$

$$H(t) = h_x(t,x(t)) \tag{3.15}$$

From Theorem 3.1, (3.10) and (3.11) are unique expressions for $\delta x(t)$ and $\delta y(t)$. For $0 < \sigma, \sigma', \varepsilon << 1$, we obtain

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\delta\dot{x}(t) = A(t)\,\delta x(t) + B(t)\,\delta p_o\,, \quad \delta x(0) = \delta x_o \tag{3.16}$$

$$\delta y(t) = H(t)\,\delta x(t) \tag{3.17}$$

If we can determine the necessary and sufficient conditions under which δx_o and δp_o can be determined uniquely in the neighborhood of (x_o,p_o), $N(x_o,p_o)$, from $\delta y(t)$, $t_o \leq t \leq T$, then we have the desired result. However, Theorem 2.4 can be applied directly to (3.16) and (3.17). Let us denote $K(t_o,t_1)$ by $K(t_o,t_1;x_o,p_o)$, since its value clearly depends on the reference trajectory of (3.1). The positive definiteness of $K(t_o,t_1;x_o,p_o)$ is then necessary and sufficient for the observability of (3.1) and (3.2) at t_o in the neighborhood of (x_o,p_o). If x_o is known, the positive definiteness of $K_{22}(t_o,t_1;x_o,p_o)$ is necessary and sufficient for the observability of p_o.

We can, in principle, examine the observability of the entire domain W by computing $K(t_o,t_1;x_o,p_o)$ at a number of grid points separated by a distance k, $k < \min(\sigma,\sigma')$. However, we must consider the possibility that two or more isolated points or sets of points in W, each of which generates a positive definite $K(t_o,t_1;x_o,p_o)$ might yield identical observations $y(t)$, $t_o \leq t \leq T$. By "isolated" we mean that the distance between the two neighborhoods is greater than ε. In such a case, even though the system is locally observable at each point, the system is unobservable at t_o. We will now state the following theorem.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Theorem 3.2. If $K(t_o,t_1;x_o,p_o)$ is positive definite for all $(x_o,p_o) \in W$, then there cannot exist two or more isolated points or sets of points in W which yield identical observations for $t_o \leq t \leq T$. Thus, if the system of (3.1) and (3.2) is locally observable on the entire domain of W, it is observable at t_o.

The proof of Theorem 3.2 proceeds as follows. For convenience, let us consider the case in which $2m \geq n+\ell$. The set of $x(t)$ that satisfy the observation relation $y(t) = h(t,x(t))$ can be denoted by the $(n+\ell-m)$ dimensional manifold $\Omega(t)$. From the assumption that $h \in C^1$, $\dot{y}(t)$ can be considered as an additional observation, related to $x(t)$ and p by

$$\dot{y}(t) = h_t(t,x) + h_x(t,x)\, f(t,x,p) \tag{3.18}$$

Let $\tilde{\Omega}(t)$ represent the $(n+\ell-m)$ dimensional manifold of the solutions $x(t)$ and p of (3.18). Hence, both $y(t)$ and $\dot{y}(t)$ assume the role of observations for $t_o \leq t \leq T$.

Let us assume that each point (x_o,p) in $\Omega(t_o)$ generates a positive definite $K(t_o,T;x_o,p)$. Since $2m \geq n+\ell$, $\Lambda(t_o) = \Omega(t_o) \cap \tilde{\Omega}(t_o)$ may contain a number of isolated points, that is, there may exist more than one value of (x_o,p) satisfying (3.2) and (3.18). Suppose $\Lambda(t_o)$ contains two points (x_o,p) and $(x_o'p')$ at each of which K is positive definite. Thus, these two points generate identical values for $y(t_o)$ and $\dot{y}(t_o)$. If we fix $y(t_o)$ and thus fix $\Omega(t_o)$, there exist bounds on the value of $\dot{y}(t_o)$, $a_* \leq \dot{y}(t_o) \leq a^*$, such that allowing $\dot{y}(t_o)$ to vary within this range of values will cause $\Lambda(t_o)$ to cover all of $\Omega(t_o)$. a_* and a^* can

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be determined from the requirement that $\Lambda(t_o)$ must be non-empty. Then we certainly can choose $\dot{y}(t_o)$ such that $||(x_o,p) - (x_o',p')|| < \varepsilon/2$ namely, so that the two points satisfying both (3.2) and (3.18) lie within their common neighborhoods. Note that all elements of $\Lambda(t)$ satisfy (3.1). However, from Theorems 2.4 and 3.1, if two points within a common neighborhood yield identical values of $y(t)$, K must be singular. Thus, by contradiction, it is not possible for $\Lambda(t_o)$ to contain two isolated points or sets of points if all points on $\Omega(t_o)$ are locally observable. The proof is easily generalized to the case in which $2m < n+\ell$.

From $y(t_o) = h(t_o,x_o)$ we can determine the possible manifold of x_o values $\Omega(t_o)$. From (3.18) evaluated at t_o we can determine another possible manifold of x_o and p values, $\tilde{\Omega}(t_o)$. We can then test the local observability of the system by simply calculating K at every point in $\Lambda(t_o)$. A criterion determining when, for all practical purposes, the system is unobservable may be set. For example, if $\det K < \delta$ [16] then the system may be considered unobservable.

If these procedures indicate that the system is observable for the given value of T, we can then examine the effect of reducing T on observability. Each neighborhood in $\Lambda(t_o)$ has its own characteristic time T^* such that $K(t_o,T;x_o,p_o), T \geq T^*$ is nonsingular. The overall characteristic observation time for the system would be

$$T^*_{ob} = \sup_{x_o,p_o \in \Lambda} T^*(x_o,p_o) \qquad (3.19)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4. Examples

Example 1. Let us consider the consecutive chemical reactions

$$A \xrightarrow{p_1} B \xrightarrow{p_2} C$$

If we let x_1 and x_2 denote the concentrations of A and B, respectively, the dynamic description of the system is

$$\dot{x}_1(t) = -p_1 x_1(t) \tag{4.1}$$

$$\dot{x}_2(t) = p_1 x_1(t) - p_2 x_2(t) \tag{4.2}$$

where p_1 and p_2 are the rate constants for the steps shown. We assume p_2 is known, and the observability of the system will be defined for the determination of x_{1_o}, x_{2_o} and p_o from given observations.

Consider first the observation of the concentration of species A only

$$y(t) = x_1(t) \tag{4.3}$$

At $t = 0$, if we have observations $y(0)$ and $\dot{y}(0) = -p_{1_o} x_{1_o}$, we can determine both x_{1_o} and p_{1_o}. Thus, the locus of intersection of the set of solutions of (4.3) and $\dot{y}(0) = -p_{1_o} x_{1_o}$ in the x_1, x_2, p_1 space is the line which at all points has $x_1 = x_{1_o}$ and $p_1 = p_{1_o}$. We can test $K(0,t;x_{1_o},x_{2_o},p_{1_o})$ at every point on this line and we will find that K is singular. Thus, the system of (4.1) - (4.3) is unobservable. Physically, this result is easily explained, since from observations of the concentration of species A

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

only, we can never determine the initial concentration of B .

Now consider the case in which we observe only the concentration of species B ,

$$y(t) = x_2(t) \tag{4.4}$$

At $t = 0$, $y(0) = x_{2_o}$, and $\dot{y}(0) = p_1 x_{1_o} - p_2 x_{2_o}$. The locus of intersection of the solutions of these two relations is the curve $p_{1_o} x_{1_o} = \dot{y}(0) + p_2 y(0)$ in the plane of $x_2 = x_{2_0}$ in the space of x_1, x_2, p_1 . $K(0,t;x_{1_o}, x_{2_o}, p_{1_o})$ is positive definite along this curve, so the system of (4.1), (4.2) and (4.4) is observable. Thus by measuring the intermediate component in simple consecutive reaction schemes rather than the primary component, the system can be made observable.

Example 2. We consider the following system

$$\dot{x}_1(t) = x_2(t) \tag{4.5}$$

$$\dot{x}_2(t) = x_1(t) \tag{4.6}$$

$$y(t) = x_1(t)\, x_2(t) \tag{4.7}$$

Thus,

$$\dot{y}(t) = x_1^2(t) + x_2^2(t) \tag{4.8}$$

Let us assume that $y(0) = 3$ and $\dot{y}(0) = 10$. Thus we can have (x_{1_o}, x_{2_o}) equal to (3,1) or (1,3). The solutions corresponding to these two possible initial conditions are

$$x_1(t) = 2e^t + e^{-t} \tag{4.9}$$

$$x_2(t) = 2e^t - e^{-t} \tag{4.10}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and

$$x_1(t) = 2e^{t} - e^{-t} \tag{4.11}$$

$$x_2(t) = 2e^{t} + e^{-t} \tag{4.12}$$

Each set of solutions generates the same $y(t)$ and $\dot{y}(t)$, $t \geq 0$. It is easy to show that $K(0,t;x_{1_o},x_{2_o})$ is nonsingular in the neighborhood of both (3,1) and (1,3). However, from Theorem 3.2 we know that if there exist two or more initial conditions which satisfy both (4.7) and (4.8) then the system cannot be observable on the entire domain of x_{1_o} and x_{2_o}. In fact, in this example, any initial condition on the line $x_{1_o} = x_{2_o}$ yields a singular value of K. Thus the system of (4.5) - (4.7) is unobservable at $t = 0$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5. Notation

$A(t)$ = $n \times n$ time dependent matrix

$B(t)$ = $n \times \ell$ time dependent matrix

G = constant defined by equation (3.7)-

$H(t)$ = $m \times n$ time dependent matrix

h = m-dimensional vector function

I = identity matrix

$K(t_o,t_1)$ = symmetric matrix (proposed observability matrix) given by equation (2.9)

$L(s)$ = Lipschitz constant

$M(t_1,t_1)$ = controllability matrix defined by equation (2.3)

$N(x_o,p_o)$ = neighborhood of the point (x_o,p_o)

p = constant parameter

q = vector function defined by equation (2.14)

$Q(t)$ = observability matrix defined by equation (2.19)

s = dummy variable

$S(t)$ = matrix function defined by equation (2.20)

t = time variable

$u(t)$ = ℓ-dimensional control vector

v = n-dimensional vector

$W(t_o,t_1)$ = observability matrix defined by equation (2.6)

x = n-dimensional state vector

y = m-dimensional observation vector

Greek Symbols

δ = variation

ε = constant

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ζ = dummy variable

ξ = constant

σ = constant

$\phi(t,\tau)$ = transition matrix

$\Omega(t_o)$ = initial manifold determined by $y(t_o)$

Superscripts

$\cdot$ = time derivative

$\sim$ = new quantity

$*$ = characteristic, or upper bound

Subscripts

$*$ = lower bound

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter V

CONCLUSIONS AND REMARKS

Chapter II

The objective of this chapter has been to present and study schemes for the control of noisy dynamic systems. When dynamical noise enters a process in the form of additive inputs, the system acts as a natural filter as long as the principal frequency band of the noise is much greater than the characteristic frequency of the system. The key factor in stochastic feedback control, however, is noise due to measurement error. Addition of a filter significantly improves the controller performance when the noise level is high. The proposed scheme can, in principle, be applied to distributed parameter systems with simple controller function. In practice, the scheme can be profitably employed when the improvement of system performance justifies the cost of additional computation.

Chapter III

General nonlinear filtering and fixed-point smoothing (interpolation) equations have been derived for distributed parameter systems. The system and the boundary conditions contain noisy inputs which are described by stochastic O.D.E.'s. Additive dynamical and observation noise can also be present both in the interior and the boundary. The observation process can be either discrete or continuous along the spatial axis, but is continuous with regard to time. The results obtained have been applied to estimate the state and the parameter of a nonlinear hyperbolic system and the state of a parabolic system. An

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

approximation to the solution of the covariance equations was found to be highly effective in reducing the computation required for the filter.

Future work might be to develop corresponding statistical approaches to derive a nonlinear filter with the present results as a guideline. A systematic approximation of the solution of the covariance equations can be investigated for on-line applications. Filter convergence which is related to observability should also be studied. The optimal choice of measurement devices for accuracy and good convergence of the filter can then be investigated.

Chapter IV

New necessary and sufficient conditions for the local observability of nonlinear lumped parameter systems have been obtained. The local observability for any initial condition can be determined by computing the proposed observability matrix, and global observability can be examined by extending the local result. Although the present result provides a computational method for determining the observability condition, it requires excessive computation for general problems. The approach can be extended to some limited classes of nonlinear P.D.E. systems if there exists an effective and general numerical technique for evaluation of the Green's function for the linearized system. This extension requires consideration of questions of well-posedness and of the perturbation theory of nonlinear P.D.E. systems, for which existing theory is inadequate.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

1. Aoki, M., Optimization of Stochastic Systems, Academic Press, New York, 1967.

2. Aoki, M., "On observability of stochastic discrete-time dynamic systems", J. Franklin Inst 286, 36-58 (1968).

3. Astrom, K. J., Introduction to Stochastic Control Theory, Academic Press, New York, 1970.

4. Athans, M., Wishner, R. P. and Bertolin, A., "Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements", Joint Automatic Control Conf., Ann Arbor, Michigan, 1968, pp. 364-382.

5. Balakrishnan, A. V. and Lions, J. L., "State estimation for infinite dimensional systems", J. Computer & System Sc. 1, 391-403 (1967).

6. Bejczy, A. K. and Sridhar, R., "Analytical methods for performance evaluation of nonlinear filters", Jet Propulsion Lab., California Institute of Technology, March 1970.

7. Bryson, A. E., Jr. and Ho, Y. C., Applied Optimal Control, Blaisdell Pub. Co., Waltham, Mass., 1969.

8. Bucy, R. S. and Joseph, P. D., Filtering for Stochastic Processes with Applications to Guidance, Interscience Publishers, New York, 1968.

9. Butkovskiy, A. G., Distributed Control Systems, Elsevier, New York, 1969.

10. Cox, H., "On the estimation of state variables and parameters for noisy dynamic systems", IEEE Trans. Automatic Control 9, 5-12 (1964).

11. Detchmendy, D. D. and Sridhar, R., "Sequential estimation of states and parameters in noisy nonlinear dynamical systems", J. Basic Eng. Trans.ASME 89, 362-368 (1966).

12. Deutsch, R., Estimation Theory, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1965.

13. Dreyfus, S. E., "Some types of optimal control of stochastic systems", SIAM J. Control 2, 120-134 (1964).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

14. Early, B. N., "Stochastic optimal control", Ph.D. thesis, California Institute of Technology, Pasadena, California, 1970

15. Figueiredo, R. J. P. de and Dyer, L. W., "Extensions of discrete stochastic approximation with dynamics and applications to nonlinear filtering", System Report No. 20-5, Rice University, Houston, Texas, August 1967.

16. Franklin, J. L., Matrix Theory, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1968.

17. Fuller, A. T., "Analysis of nonlinear stochastic systems by means of the Fokker-Planck equations", Int. J. Control 9, 603-655 (1969).

18. Gavalas, G. R. and Seinfeld, J. H., "Sequential estimation of states and kinetic parameters in tubular reactors with catalyst decay", Chem. Eng. Sci. 24, 625-636 (1969).

19. Gelfand, I. M. and Fomin, S. V., Calculus of Variations, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

20. Gibson, J. E., Nonlinear Automatic Control, McGraw-Hill Co., New York, 1963.

21. Goodson, R. E. and Klein, R. E., "A definition and some results for distributed system observability", IEEE Trans. Automatic Control 15, 165-174 (1970).

22. Griffin, R. E. and Sage, A. P., "Large and small scale sensitivity analysis of optimum estimation algorithms", IEEE Trans. Automatic Control 13, 320-329 (1968).

23. Griffith, E. W. and Kumar, K. S. P., "On the observability of nonlinear systems: I", Dept. of Electrical Eng., Univ. of Minnesota, 1970.

24. Heins, W. and Mitter, K. S., "Conjugate convex functions, duality and optimal control problems I: Systems governed by ordinary differential equations", Information Sciences 2, 211-243 (1970).

25. Hestenes, M. R. and Guinn, T., "An embedding theorem for differential equations", J. Optimization Theory and Applications 2, 87-101 (1968).

26. Jazwinski, A. H., Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

27. Johnson, F. C. and Meditch, J. S., "Review and critique of some procedures and results in nonlinear estimation", Boeing Scientific Research Lab., Report D1-82-1001, Sept. 1970.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

28. Jurdjevic, V., "Abstract control systems: controllability and observability", SIAM J. Control 8, 424-439 (1970).

29. Kagiwada, H. H., Kalaba, R. E., Schumitzky, A. and Sridhar, R., "Invariant imbedding and sequential interpolating filters for nonlinear processes", J. Basic Eng., Trans. ASME 91, 195-200 (1969).

30. Kalman, R. E., "A new approach to linear filtering and prediction problems", J. Basic Eng., Trans. ASME 82, 35-45 (1960).

31. Kalman, R. E. and Bucy, R. S., "New results in linear filtering and prediction theory", J. Basic Eng., Trans. ASME 83, 95-108 (1961).

32. Kalman, R. E., "Contributions to the theory of optimal control", Bol. Soc. Mat. Mexicana, 102-119 (1960).

33. Kalman, R. E., Ho, Y. C. and Narendra, K. S., "Controllability of linear dynamical systems", Contributions to Differential Equations 1, 189-213 (1962).

34. Kim, M. and Gajwani, S. H., "A variational approach to optimum distributed parameter systems", IEEE Trans. Automatic Control 13, 191-193 (1968).

35. Kostyukovskii, Yu. M. L., "Observability of nonlinear controlled systems", Automation and Remote Control 29, 1384-1396 (1968).

36. Kostyukovskii, Yu. M. L., "Simple conditions of observability of nonlinear controlled systems", Automation and Remote Control 29, 1575-1584 (1968).

37. Kushner, H. J., "On the dynamical equations of conditional probability density functions, with applications to optimal stochastic control theory", J. Math. Anal. & Appl. 8, 332-344 (1964).

38. Kushner, H. J., "On the differential equations satisfied by conditioned probability densities of Markov processes, with applications" SIAM J. Control 2, 106-119 (1964).

39. Kushner, H. J., "Filtering for linear distributed parameter systems", SIAM J. Control 8, 346-359 (1970).

40. Lamont, G. B. and Kumar, K. S. P., "State estimation in distributed parameter systems via least squares and invariant imbedding", to appear in J. Math. Anal. Appl.

41. Lapidus, L., "Control, stability and filtering", Ind. Eng. Chem.59, 28-38 (1967).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

42. Lee, E. B. and Markus, L., Foundations of Optimal Control Theory John Wiley, New York, 1967.

43. Lee, E. S., Quasilinearization and Invariant Imbedding, Academic Press, New York, 1968.

44. Lefschetz, S., Stability of Nonlinear Control Systems, Academic Press, New York, 1965.

45. May, L. E.,"Perturbations in fully nonlinear systems", SIAM J. Math. Anal. 1, 376-391 (1970).

46. Meditch, J. S., "On state estimation for distributed parameter systems", Boeing Scientific Research Lab. Report D1-82-0917, Sept. 1969.

47. Meditch, J. S., "Filtering and smoothing of boundary and interior measurement data for distributed parameter systems", Boeing Scientific Research Lab. Report D1-82-0995, Sept. 1970.

48. Meditch, J. S., Stochastic Optimal Linear Estimation and Control McGraw-Hill Co., New York, 1969.

49. Mitter, S. K., "Optimal control if distributed parameter systems" in Control of Distributed Parameter Systems, American Society of Mechanical Engineers, New York, 1969.

50. Nishimura, T., "Error bounds of continuous Kalman filters and the applications to orbit determination problems", IEEE Trans. Automatic Control 12, 268-275 (1967).

51. Papoulis, A., Probability, Random Variables and Stochastic Processes, McGraw-Hill Co., New York, 1965.

52. Pell, T. M., Jr. and Aris, R., "Some problems in chemical reactor analysis with stochastic features II. Control of linearized distributed systems on discrete and corrupted observations", Ind. & Eng. Chem. Fund 9, 15-20 (1970).

53. Pontryagin, L. S. et al, The Mathematical Theory of Optimal Processes, Interscience Publishers, New York, 1962.

54. Richtmyer, R. D. and Morton, K. W., Difference Methods for Initial-Value Problems, Interscience Publishers, New York, 1967.

55. Roitenberg, Y. Y., "Observability of nonlinear systems", SIAM J. Control 8, 338-345 (1970).

56. Russell, D. L. and Lukes, D. L., "The quadratic criterion for distributed systems", SIAM J. Control 7, 75-83 (1969).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

57. Schwartz, L. and Stear, E. B., "A computational comparison of several nonlinear filters", IEEE Trans. Automatic Control 13, 83-86 (1968).

58. Seinfeld, J. H., "Optimal stochastic control of nonlinear systems", AIChE J. 16, 1016-1022 (1970).

59. Seinfeld, J. H., "Nonlinear estimation theory", Ind. Eng. Chem. 62, 32-42 (1970).

60. Silverman, L. M. and Meadows, H. E., "Controllability and observability in time-variable linear systems", SIAM J. Control 5, 64-73 (1967).

61. Simon, K. W. and Stubberud, A. R., "Duality of linear estimation and control", J. Optimization Theory and Applications 6, 55-67 (1970).

62. Sorenson, H. W., "On the error behavior in linear minimum variance estimation problems", IEEE Trans. Automatic Control 12, 557-562 (1967).

63. Sorenson, H. W., "Controllability and observability of linear stochastic, time-discrete control systems", Advances in Control Systems 6, 95-160 (1968).

64. Thau, F. E., "On optimal filtering for a class of linear distributed parameter systems", J. Basic Eng., Trans. ASME 91, 173-178 (1969).

65. Tzafestas, S. G. and Nightingale, J. M., "Differential dynamic-programming approach to optimal nonlinear distributed parameter control systems", Proc. I.E.E. 116, 1079-1084 (1968).

66. Tzafestas, S. G. and Nightingale, J. M., "Optimal control of a class of linear stochastic distributed parameter systems", Proc. I.E.E. 115, 1213-1220 (1968).

67. Tzafestas, S. G. and Nightingale, J. M., "Maximum-likelihood approach to the optimum filtering of distributed parameter systems" Proc. I.E.E. 116, 1085-1093 (1969).

68. Tzafestas, S. G. and Nightingale, J. M., "Boundary and volume filtering of linear distributed parameter systems", Electronic Letters 5, 199-200 (1969).

69. Volterra, V., Theory of Functional and of Integral and Integro-Differential Equations, Dover Publications, New York, 1959

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

70. Wang, P. K. C., "Control of distributed parameter systems", Advances in Control Systems 1, 75-172 (1964).

71. Weiss, L., "On the structural theory of linear differential equations", SIAM J. Control 6, 659-680 (1968).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

PROPOSITIONS

P-I Review of Numerical Integration Techniques for Stiff Ordinary Differential Equations

P-II Control of Plug-Flow Tubular Reactors by Variation of Flow Rate

P-III Some Results on Estimation of Parameters in Ordinary Differential Equations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

P-I

Review of Numerical Integration Techniques for Stiff Ordinary Differential Equations

(Part of this work has been accepted at the candidacy examination in February 1969.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Review of Numerical Integration Techniques for Stiff Ordinary Differential Equations

John H. Seinfeld, Leon Lapidus,[1] and Myungkyu Hwang
Chemical Engineering Laboratory, California Institute of Technology, Pasadena, Calif. 91109

Ordinary differential equations with widely separated eigenvalues (stiff O.D.E.) occur often in practice and present severe numerical integration problems. The stability and accuracy problems associated with the numerical solution of such equations are outlined. Several methods, including a modified Runge-Kutta method due to Treanor, a class of implicit Runge-Kutta methods, extrapolation methods, and methods based on the inclusion of second derivatives and exponential fitting, are considered. Numerical results are given on three stiff systems for stiff and conventional methods and recommendations are made on what methods to use for particular systems.

Many physical systems give rise to ordinary differential equations (O.D.E.), the magnitudes of the eigenvalues of which vary greatly. Such situations arise in the study of the flow of a chemically reacting gas (Emanuel, 1963; Eschenroeder *et al.*, 1962), exothermic chemical reaction in a tubular reactor (Amundson, 1965; Amundson and Luss, 1968), circuit theory (Brayton *et al.*, 1966; Calahan and Abbott, 1967), and process dynamics and control (Davison, 1968; Kalman, 1966; Mah *et al.*, 1962). It is common to refer to such systems as "stiff."

Let us examine the particular problems associated with the numerical integration of stiff equations. To do this consider the linear O.D.E.

$$\mathbf{y}' = \mathbf{A}\mathbf{y} \tag{1}$$

where $\mathbf{y} = [y_1,y_2]^T$, $\mathbf{y}(0) = [2,1]^T$, and

$$\mathbf{A} = \begin{bmatrix} -500.5 & 499.5 \\ 499.5 & -500.5 \end{bmatrix} \tag{2}$$

The solution of Equation 1 is

$$y_1(x) = 1.5e^{-x} + 0.5e^{-1000x} \tag{3}$$

$$y_2(x) = 1.5e^{-x} - 0.5e^{-1000x} \tag{4}$$

where the eigenvalues of $\mathbf{A}$ are $\lambda_1 = -1000$ and $\lambda_2 = -1$. Both y_1 and y_2 have a rapidly decaying component, corresponding to λ_1, which very quickly becomes insignificant. After a brief initial phase of the solution in which the λ_1 component is not negligible, we would like to proceed, if we were integrating Equation 1 numerically, with a step length h which is determined only by the component of the solution corresponding to λ_2. However, for a stable numerical solution most methods require that $|h\lambda_1|$ and $|h\lambda_2|$ be bounded by a single small number, the order of 1 to 10. The stability of numerical integration of Equation 1 will be governed by $|-1000h|$—for example, for Euler's method it is necessary that $|-1000h| < 2$, giving the maximum stable $h = 0.002$. Thus, 500 integration steps would be required to reach $x = 1$.

Although the component of the solution corresponding to λ_1 is of no practical interest, the criterion of absolute stability (defined precisely below) forces us to use an extremely small value of h over the entire range of integration. As a result, the computation time necessary to integrate a highly stiff system can become excessive.

The purpose of this paper is twofold. First, we outline the problems of numerical stability and accuracy in the integration of stiff O.D.E. Second, we present several numerical integration algorithms for stiff O.D.E. together with detailed numerical results on the use of these algorithms on example systems.

Stability and Accuracy in Integration of Stiff O.D.E.

Numerical Stability of Linear Multistep Methods. An important and extensive class of numerical integration formulas is represented by the general linear multistep method

$$y_{n+1} = \alpha_1 y_n + \ldots + \alpha_k y_{n+1-k} + h[\beta_0 y'_{n+1} + \ldots + \beta_k y'_{n+1-k}] \tag{5}$$

where $y_n, y_{n+1}, \ldots$, are the numerically computed approximation to the exact solutions, $\mathbf{y}(x_n), \mathbf{y}(x_{n+1}), \ldots$, of the O.D.E.

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y}) \tag{6}$$

at equidistant points, $x_n = x_0 + nh$, $x_{n+1} = x_0 + (n+1)h$, etc. If $\beta_0 = 0$, Equation 5 is explicit, and if $\beta_0 \neq 0$, it is implicit. It is convenient to develop the concepts of numerical stability with reference to this class of methods, although the ideas are completely general and are applicable to all classes of numerical integration methods.

In the numerical solution of an O.D.E., a sequence of approximations $\mathbf{y}_n$ to the true solution $\mathbf{y}(x_n)$ is generated. The stability of a numerical method refers to the behavior of the difference or accumulated error $\mathbf{y}(x_n) - \mathbf{y}_n$ as n becomes large. Extensive treatments of numerical stability are given by Dahlquist (1956, 1963a,b) and Henrici (1962). In this section we outline only those aspects bearing on the stiff problem.

Consider for the moment the numerical integration of the scalar form of Equation 1

$$y' = \lambda y;\ y(x_0) = 1 \tag{7}$$

by Equation 5. The characteristic equation of Equation 5 is

[1] Department of Chemical Engineering, Princeton University, Princeton, N. J. 08540

Reprinted from I&EC FUNDAMENTALS, Vol. 9, No. 2, p. 266, May 1970
Copyright 1970 by the American Chemical Society and reprinted by permission of the copyright owner

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\mu^k - \sum_{i=1}^{k} \alpha_i \mu^{k-i} - h\lambda \sum_{i=0}^{k} \beta_i \mu^{k-i} = 0 \qquad (8)$$

which is a kth-order polynomial in μ. The k solutions of Equation 8 are the characteristic roots μ_i, $i = 1,2, \ldots, k$. The numerical solution is thus

$$y_n = d_1\mu_1^n + d_2\mu_2^n + \ldots + d_k\mu_k^n \qquad (9)$$

One of the characteristic roots approximates the Taylor series expansion of the true solution, $y = \exp(\lambda x)$, with a truncation error corresponding to the order p of the method. If we let this root be μ_1, then $\mu_1 = \exp(h\lambda) + 0(h^{p+1})$ as $h \to 0$. This root, called the principal root, is the root which we wish to be represented in the numerical solution, since μ_1^n approximates $\exp(nh\lambda)$. The other $k - 1$ roots are called spurious or extraneous roots and are a result of the use of a difference equation of degree k to represent a first-order differential equation. The extraneous roots have no relation to the exact solution but, nevertheless, are unavoidable.

The characteristic roots of Equation 8 are the same as those of the difference equation for the error, $\epsilon_n = y_n - y(x_n)$—i.e.,

$$\epsilon_n = c_1\mu_1^n + c_2\mu_2^n + \ldots + c_k\mu_k^n \qquad (10)$$

For a valid numerical solution we require that ϵ_n not grow with n. A linear multistep method is called

Absolutely stable, if $|\mu_i| \leq 1 \quad i = 1,2, \ldots, k$

Relatively stable, if $|\mu_i| \leq |\mu_1| \quad i = 2,3, \ldots, k$

The single O.D.E. $y' = \lambda y$ will be called inherently stable if $\mathrm{Re}(\lambda) < 0$. In this case the exact solution is decreasing with x_n, and the important condition is absolute stability, since the numerical solution must also decrease with x_n. If, however, $\mathrm{Re}(\lambda) \geq 0$, the exact solution is growing with x_n, and we do not want $|\mu_i| \leq 1$; rather it is relative stability that is the important consideration. In other words, we will have a valid solution as long as no component of the numerical solution, μ_i^n, increases faster than the one corresponding to the principal root. The critical problems of numerical stability in stiff O.D.E. are associated with inherently stable O.D.E., $\mathrm{Re}(\lambda) < 0$, $i = 1,2, \ldots, m$, in which absolute stability is the important factor. Thus, in this paper we confine our attention to stiff O.D.E. of this type.

The value of $h\lambda$ for which $|\mu_i| = 1$ and for which a small increase in $|h\lambda|$ makes $|\mu_i| > 1$ is called the general stability boundary. Since in general λ is complex, we can let $\lambda = \bar{\lambda} \exp(i\theta)$, where $\bar{\lambda}$ and θ are real. Then we can make similar definitions of the real and imaginary stability boundaries, corresponding to the values of $\bar{\lambda}h$ and $i(\bar{\lambda}h)$, where the root condition is obeyed. Any method with a finite general stability boundary can be called conditionally stable, whereas any method with an infinite general stability boundary can be called unconditionally stable, or A-stable. Thus, a linear multistep method is A-stable if all solutions of Equation 9 tend to zero, as $n \to \infty$, when the method is applied with fixed $h > 0$ to $y' = \lambda y$, where λ is a complex constant with $\mathrm{Re}(\lambda) < 0$. An elegant theory exists for stability of linear multistep methods as $h \to 0$ (Dahlquist, 1956, 1963b; Henrici, 1962). While many interesting results have been obtained in this asymptotic case, it is not of prime concern in the stiff problem.

Dahlquist (1956, 1963a,b) has proved two important theorems relating p and k for A-stable multistep methods.

Theorem 1. An explicit k-step method cannot be A-stable.

Theorem 2. The order p of an A-stable linear multistep method cannot exceed 2. The smallest truncation error in such case is obtained for the trapezoidal rule for which $p = 2$, $k = 1$.

The concept of A-stability can be modified somewhat to allow for methods of higher accuracy. Widlund (1967) has proved that A-stable methods for which $k = p = 3$ and $k = p = 4$ exist if λ is a complex constant which lies in the wedge-shaped region, $S_\theta = \{z; |\arg(-z)| < \theta, z \neq 0\}$, $\theta \in (0, \pi/2)$.

In general, it is impossible to construct arbitrarily accurate A-stable linear multistep methods. An important point, however, is that Theorems 1 and 2 in no way restrict the construction of more accurate A-stable methods which are not of the linear multistep type. Such methods can be constructed.

Stability of Multistep Methods in Integrating Coupled O.D.E. In this section we consider the relationship between the characteristic roots of a linear multistep method and the eigenvalues of the O.D.E. Let us still confine our attention to the linear O.D.E. Equation 1. First let us consider the numerical integration of Equation 1 by Euler's method,

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}_n' \qquad (11)$$

and the trapezoidal rule (modified Euler method),

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{y}'_{n+1} + \mathbf{y}_n') \qquad (12)$$

The exact solution of Equation 1 will change over a step h by the factor $\exp(\mathbf{A}h)$—i.e., $\mathbf{y}(x_{n+1}) = \exp(\mathbf{A}h)\mathbf{y}(x_n)$. A single-step method when applied to Equation 1 can be expressed as

$$\mathbf{y}_{n+1} = \mathbf{M}(h\mathbf{A})\mathbf{y}_n \qquad (13)$$

the properties of which are determined by how well $\mathbf{M}(h\mathbf{A})$ approximates $\exp(h\mathbf{A})$. For Euler's method, $\mathbf{M}(h\mathbf{A}) = \mathbf{I} + h\mathbf{A}$, the first 2 terms in an infinite series expansion of $\exp(h\mathbf{A})$. For the trapezoidal rule, $\mathbf{M}(h\mathbf{A}) = (\mathbf{I} - 1/2h\mathbf{A})^{-1}(\mathbf{I} + 1/2h\mathbf{A})$, the first diagonal Padé approximant to $\exp(h\mathbf{A})$ (Calahan, 1967; Kelly, 1967).

For an m-dimensional linear O.D.E. and any single-step method, the stability of the algorithm depends only on the eigenvalues of the O.D.E. This can be shown as follows. Let us carry out the similarity transformation

$$\mathbf{y} = \mathbf{P}\mathbf{z} \qquad (14)$$

so that Equation 1 becomes

$$\mathbf{z}' = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{z} \qquad (15)$$

$$\mathbf{z}(0) = \mathbf{P}^{-1}\mathbf{y}_0$$

If the eigenvalues λ_i of $\mathbf{A}$ are distinct, Equation 15 reduces to

$$\mathbf{z}' = \mathbf{\Lambda}\mathbf{z} \qquad (16)$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues λ_i. The solution of Equation 1 is

$$\mathbf{y}(x) = \mathbf{P}\exp(\mathbf{\Lambda}x)\mathbf{P}^{-1}\mathbf{y}_0 \qquad (17)$$

or, in a recursive notation,

$$\mathbf{y}(x_{n+1}) = \mathbf{P}\exp(\mathbf{\Lambda}h)\mathbf{P}^{-1}\mathbf{y}(x_n) \qquad (18)$$

A single-step method can be expressed as Equation 13, so that we want

$$\mathbf{M}(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}h) = \mathbf{P}\exp(\mathbf{\Lambda}h)\mathbf{P}^{-1} \qquad (19)$$

If we consider the general rational form of a Padé approximation, which includes all forms of characteristic roots,

$$\mathbf{M}(\mathbf{B}) = \left(\sum_{i=1}^{n} b_i\mathbf{B}^i\right)^{-1}\left(\sum_{i=1}^{m} a_i\mathbf{B}^i\right) \qquad (20)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and let

$$\mathbf{B} = \mathbf{P}\underline{\Lambda}\mathbf{P}^{-1} \tag{21}$$

then

$$\begin{aligned}\mathbf{M}(\mathbf{P}\underline{\Lambda}\mathbf{P}^{-1}) &= \left(\sum_{i=1}^{n} b_i(\mathbf{P}\underline{\Lambda}\mathbf{P}^{-1})^i\right)^{-1}\left(\sum_{i=1}^{m} a_i(\mathbf{P}\underline{\Lambda}\mathbf{P}^{-1})^i\right)\\ &= \mathbf{P}\left(\sum_{i=1}^{n} b_i\underline{\Lambda}^i\right)^{-1}\left(\sum_{i=1}^{m} a_i\Lambda^i\right)\mathbf{P}^{-1}\\ &= \mathbf{P}\mathbf{M}(\Lambda)\mathbf{P}^{-1}\end{aligned} \tag{22}$$

Referring to Equation 19, we see that

$$\mathbf{M}(\Lambda h) \cong \exp(\Lambda h) \tag{23}$$

Since each matrix is diagonal, the corresponding diagonal elements of $\mathbf{M}(\Lambda h)$ approximate those of $\exp(\Lambda h)$. Each diagonal element of $\mathbf{M}$ is simply the characteristic root μ_i of the single-step method, so that Equation 23 can be written

$$\mu_i(h\lambda_i) \cong \exp(h\lambda_i)\ i = 1,2,\ldots,m \tag{24}$$

Absolute stability requires that

$$|\mu_i(h\lambda_i)| \leq 1\ i = 1,2,\ldots,m \tag{25}$$

for example, for $m = 2$ and Euler's method, Equation 25 is

$$|1 + h\lambda_1| \leq 1;\ |1 + h\lambda_2| \leq 1 \tag{26}$$

and for the trapezoidal rule,

$$\left|\frac{1 + \frac{1}{2}h\lambda_1}{1 - \frac{1}{2}h\lambda_1}\right| \leq 1;\quad \left|\frac{1 + \frac{1}{2}h\lambda_2}{1 - \frac{1}{2}h\lambda_2}\right| \leq 1 \tag{27}$$

Equation 26 is satisfied if $|h\lambda_1| \leq 2$, where λ_1 is the largest eigenvalue in absolute value. From Theorem 2 we know that Equation 27 is satisfied for all $\mathrm{Re}(\lambda_i) < 0$. The important point is that in numerically integrating a coupled set of linear O.D.E. it is sufficient to consider the method as applied to the scalar equation $y' = \lambda_i y$, where λ_i takes on the values of the eigenvalues of the O.D.E.

The analysis can easily be extended to k-step methods. The kth-degree characteristic polynomial in μ yields k roots for each of the m eigenvalues λ_j, $j = 1,2,\ldots,m$.

So far we have considered linear O.D.E. Our real aim in discussing stability of numerical integration methods of O.D.E. is to treat nonlinear O.D.E. However, there does not exist at this time a general theory of absolute stability of linear multistep methods, such that the stability behavior of different multistep methods when applied to nonlinear O.D.E. can be determined in a systematic manner. What is normally done in the nonlinear case is to let the eigenvalues of the O.D.E. be the eigenvalues of the Jacobian $\mathbf{f}_y$, a procedure valid for small h (Hildebrand, 1956). We can expect that stability limits derived on the basis of the eigenvalues of the local Jacobian matrix will not be exact. Nevertheless, we will rely on such limits as representing a good approximation to the true stability limits, which, of course, are unknown.

The problem associated with stiff systems is twofold: stability and accuracy. If a method with a finite absolute stability boundary is used, large negative real parts of some λ_i will force the step length used to be excessively small. On the other hand, if an A-stable method is used—e.g., the trapezoidal rule—the stability problem is avoided but for a reasonable step length h the solution component corresponding to the largest eigenvalue will be approximated very inaccurately. For example, in the case of Euler's method the accuracy is determined by the approximations

$$(1 + \lambda_1 h) \cong e^{\lambda_1 h}$$

$$(1 + \lambda_2 h) \cong e^{\lambda_2 h}$$

Since these approximations improve as $\lambda_i h \to 0$, the poorer approximation will be associated with the larger eigenvalue, λ_1. Similarly, for the trapezoidal rule, the characteristic root in Equation 27 improves in approximation to $\exp(h\lambda_i)$ as $h\lambda_i \to 0$. For a multistep method only the principal root μ_{1i} is an approximation to $\exp(h\lambda_i)$, the others being extraneous.

It is natural to require close approximations to $\exp(h\lambda_i)$ in the neighborhood of the origin, and this is normally the consideration in determining the principal root. At points where the λ_i have large negative real parts, the exact solution components corresponding to these eigenvalues are negligible when compared to the other solution components. It is only necessary then that the principal root also be negligible for the stiff eigenvalues.

Numerical Integration Routines for Stiff O.D.E.

We now outline several methods that have been proposed for the numerical integration of stiff O.D.E. Our treatment is limited to explicit and implicit single-step methods which have been found most efficient in actual applications. Additional methods not outlined in detail are cited.

Treanor's Method. A modified explicit Runge-Kutta method has been proposed by Treanor (1966). It assumes that Equation 6 can be approximated in an interval by the linear form

$$y_i' = -(P_i)_n y_i + (D_{1i})_n + (D_{2i})_n x + (D_{3i})_n x^2 \tag{28}$$

where P_i, D_{1i}, D_{2i}, and D_{3i} are parameters to be determined. If one applies the fourth-order Runge-Kutta method to Equation 28, the following algorithm is obtained for each component of vector $\mathbf{y}$,

$$y_{n+1/2}^{(1)} = y_n + \frac{h}{2} f_n$$

$$y_{n+1/2}^{(2)} = y_n + \frac{h}{2} f_{n+1/2}^{(1)}$$

$$y_{n+1}^{(3)} = y_n + h\,[2f_{n+1/2}^{(2)} F_2 + f_{n+1/2}^{(1)} PhF_2 + f_n(F_1 - 2F_2)] \tag{29}$$

$$\begin{aligned}y_{n+1} = y_n &+ hf_nF_1 + hv_3(Py_n + f_n) +\\ &hv_2(Py_{n+1/2}^{(1)} + f_{n+1/2}^{(1)}) +\\ &hv_3(Py_{n+1/2}^{(2)} + f_{n+1/2}^{(2)}) + hv_1(Py_{n+1}^{(3)} + f_{n+1}^{(3)})\end{aligned}$$

where

$$F_1 = \frac{e^{-Ph} - 1}{-Ph} \qquad F_2 = \frac{e^{-Ph} - 1 + Ph}{(Ph)^2} \tag{30}$$

$$F_3 = \frac{e^{-Ph} - 1 + Ph - \frac{1}{2}(Ph)^2}{-(Ph)^3}$$

and

$$v_1 = -F_3 + 4F_2\ v_2 = 2(F_2 - 2F_3)\ v_3 = 4F_3 - 3F_2 \tag{31}$$

The only undetermined parameter is P. In Equation 28 we have used P_i, whereas in Equations 29 and 30 a scalar P

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table I. Parameters for Semiimplicit Runge-Kutta Methods

Reference	p	a_1	a_2	b_1	c_1	w_1	w_2
Rosenbrock, 1963	2	$1 - \frac{\sqrt{2}}{2}$	$1 - \frac{\sqrt{2}}{2}$	$(\sqrt{2} - 1)/2$	0	0	1
	3	1.40824829	0.59175171	0.17378667	0.17378667	−0.41315432	1.41315432
Calahan, 1968	3	0.788675134	0.788675134	−1.15470054	0	0.75	0.25
Trapezoidal rule	2	1/2	1/2	0	0	1	0

has been used. The relation is as follows. The values P_i are determined as the ratio of the terms in the first two steps of Equation 29

$$P_i = -\frac{f_{n+1/2}^{(2)} - f_{n+1/2}^{(1)}}{y_{n+1/2}^{(2)} - y_{n+1/2}^{(1)}} \qquad (32)$$

where the division is defined to mean that an element in the vector in the numerator is divided by the corresponding element in the vector in the denominator. The value of P used in Equations 29 and 30 is taken to be the largest value of P_i computed from Equation 32. If this value is negative, P is set equal to zero. Using a scalar P ensures that each of the m O.D.E.'s is differenced with the same step length.

If the O.D.E. were uncoupled, the P_i would represent the local eigenvalues of the individual equations. Taking the single value of P equal to the largest P_i makes the algorithm approximate the *corresponding solution component*.

The algorithm is used as follows:

1. Select an initial step length h.
2. Compute $y_{n+1/2}^{(1)}$ and $y_{n+1/2}^{(2)}$ from the first two equations of Equation 29.
3. Compute P as the largest P_i from Equation 32. If all $P_i < 0$, set $P = 0$.
4. Compute y_{n+1} from the last equation of Equation 29.
5. If $|y_{n+1} - y_n|/|y_{n+1}| > \epsilon_{max}$, set $h = h/2$ and return to 2.
6. If $|y_{n+1} - y_n|/|y_{n+1}| < \epsilon_{min}$, set $h = 2h$ and return to 2.

A complete stability analysis of Treanor's method has been carried out by Lomax and Bailey (1967). As $h \to 0$ the method is identical to the fourth-order Runge-Kutta method. For $-2 < h\lambda < 0$, the method is stable for any value of P. If $h\lambda < -2$, the method is conditionally stable. If $Ph = 8$, the real stability boundary is -10, compared to the fourth-order Runge-Kutta value of -2.785.

The method has the advantage of improving on the stability characteristics of the fourth-order Runge-Kutta method while maintaining the same accuracy. Its major disadvantages are that it is still only conditionally stable and that because of the form of the approximation Equation 28 the method can only be used when the Jacobian matrix of the original O.D.E. has large diagonal elements, not large off-diagonal elements.

Semiimplicit Runge-Kutta Methods. Implicit Runge-Kutta methods (Butcher, 1964) are attractive for stiff systems because of being highly stable. A particularly important class of implicit Runge-Kutta methods has been developed by Rosenbrock (1963), Calahan (1968), and Allen (1969). These methods possess the dual advantages of explicit form and high stability, and are referred to as semiimplicit Runge-Kutta methods.

If we consider the autonomous form of Equation 6, $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, and define the Jacobian matrix $\mathbf{A}(\mathbf{y}) = \mathbf{f}_y$, the third-order method can be written as

$$\mathbf{k}_1 = h[\mathbf{I} - ha_1\mathbf{A}(\mathbf{y}_n)]^{-1}\mathbf{f}(\mathbf{y}_n)$$

$$\mathbf{k}_2 = h[\mathbf{I} - ha_2\mathbf{A}(\mathbf{y}_n + c_1\mathbf{k}_1)]^{-1}\mathbf{f}(\mathbf{y}_n + b_1\mathbf{k}_1) \qquad (33)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + w_1\mathbf{k}_1 + w_2\mathbf{k}_2$$

Parameter values for Equation 33 determined by Rosenbrock (1963) and Calahan (1968) for different orders p are shown in Table I. Each of the methods is A-stable. The characteristic root of Calahan's method is

$$\mu_1 = \frac{1 - 0.578\,h\lambda - 0.456\,h^2\lambda^2}{1 - 1.578\,h\lambda + 0.622\,h^2\lambda^2} \qquad (34)$$

Even though the method is A-stable, $\mu_1 \to -0.735$ as $h\lambda \to -\infty$, so that we might expect accuracy problems to arise in simulating the stiff eigenvalues.

Extrapolation Methods. The concept of Richardson extrapolation, traditionally used in Romberg's method of numerically evaluating integrals (Davis and Rabinowitz, 1967), has recently been proposed as a technique to be used in conjunction with certain single-step methods in the numerical integration of O.D.E. (Bauer *et al.*, 1963; Gragg, 1965). By forming a linear combination of the numerical results evaluated using two values of h, the leading error term in the asymptotic error, expansion of the numerical method can be eliminated (Henrici, 1964). This procedure, which can be repeated *indefinitely* with *decreasing* values of h, each time removing the leading error term, is called "extrapolation to the limit."

If we consider Euler's method, Equation 11, in which $p = 1$, the asymptotic error expansion is

$$y_n = y(x_n) + e_1(x_n)h + e_2(x_n)h^2 + e_3(x_n)h^3 + \ldots \qquad (35)$$

Values of y_n can be generated using Euler's method based on a sequence of h_k, $k = 0,1,2, \ldots$, and the $y_n(h_k)$ obtained can be denoted as $Y_0^{(k)}$. In this case the total interval length $x_n - x_0$ is h_0. Extrapolation to the limit can be applied to Equation 35. If successive interval halving is used—i.e., $h_k = h_0/2^k$—the recursion relation for generating the ever-improving approximation to y_n is

$$Y_m^{(k)} = \frac{2^m Y_{m-1}^{(k+1)} - Y_{m-1}^{(k)}}{2^m - 1} \qquad (36)$$

Convergence of $Y_m^{(k)}$ to $y(x_n)$ is guaranteed if $f(x,y)$ satisfies a Lipschitz condition (Bauer *et al.*, 1963).

In the discussion so far we have considered the extrapolation procedure as applied over the interval x_0 to x_n. What is done in practice is use extrapolation locally at each step in the integration, $x_1, x_2 \ldots$, where $x_{i+1} - x_i = h_0$. The procedure is self-starting and the choice of h_0 is fairly arbitrary, since h is automatically reduced until the required accuracy is achieved at each step of the solution—e.g., $|(Y_m^{(k)} - Y_m^{(k-1)})/$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$|Y_m^{(k-1)}| < \epsilon_y$. A complete description of Euler's method coupled with local extrapolation is given by McCalla (1967).

Gragg (1965) and Bulirsch and Stoer (1966) have proposed using the midpoint rule coupled with local extrapolation. This algorithm converges more quickly than one based on Euler's method because the midpoint rule, $y_{n+1} = y_{n-1} + 2hy'_n$, has an error expansion in powers of h^2 rather than h. The theory behind the algorithm is basically the same as described for Euler's method. Both of these algorithms are based on local extrapolation, where the extrapolation is carried out at each step in the integration. The other alternative is global extrapolation, wherein the particular numerical method is first applied over the entire range of integration from x_0 to x_n for a decreasing sequence of step lengths, and then extrapolation is applied to the values obtained at each original mesh point, $x_n = x_0 + nh_0$.

The algorithm commonly used with global extrapolation is the trapezoidal rule. Global extrapolation with the trapezoidal rule is necessary to preserve A-stability, as we show shortly.

A single-step method is usually expressed in the form, $y_{n+1} = \mu_1(h_0\lambda)y_n$. An extrapolation algorithm applied locally over the interval (x_n, x_{n+1}) can be expressed as

$$y_{n+1} = \beta(h_0\lambda, M, K)y_n \tag{37}$$

where K is the number of step lengths for which the core algorithm is used over the over-all step length h_0, and M is the number of times that extrapolation is carried out. Assuming M and K do not vary from step to step, absolute stability requires that

$$|\beta(h_0\lambda, M, K)| \leq 1 \tag{38}$$

In order to determine the stability bounds of the algorithm it is necessary to determine for fixed M and K the values of $h_0\lambda$ for which Equation 38 is just satisfied. For Euler's method, Equation 36 can be expressed as

$$Y_m^{(k)} = \sum_{i=0}^{m} c_{m,m-i} Y_0^{(k+i)} \tag{39}$$

where the coefficients obey the recursion relation (Bauer *et al.*, 1963)

$$c_{m,m-i} = \frac{2^m c_{m-1,m-i} - c_{m-1,m-1-i}}{2^m - 1} \tag{40}$$

$$c_{m-i,m} = c_{m-i,i-1} = 0$$

If $Y_m^{(k)}$ is taken as y_{n+1}, Equations 37 and 39 yield

$$\beta(h_0\lambda, M, K) = \sum_{i=0}^{M} c_{M,M-i} \left[1 + \frac{h_0\lambda}{2^{k+i}}\right]^{2^{k+i}} \tag{41}$$

In order to determine the stability region corresponding to various values of K and M, it is necessary to determine the values of $h_0\lambda$ corresponding to K and M that Equation 38 is just satisfied. This has been carried out for $K = 0$ (the diagonal elements) and various values of M for real, negative $h_0\lambda$. The bounds are:

M	1	2	3	4	5	6
$(h_0\lambda)_{max}$	−2	−2.785	−4.23	−9.06	−10.88	−13.5

As the number of extrapolations M is increased, the over-all algorithm becomes more stable. In the limit of an infinite number of extrapolations the method approaches A-stability. Because the midpoint rule is only weakly stable (Henrici, 1964), the extrapolation with the midpoint rule will be less stable than with Euler's method and thus not well suited for stiff systems. If the trapezoidal rule were applied locally, the modified characteristic root is

$$\beta(h_0\lambda, M, K) = \sum_{i=0}^{M} c'_{M,M-i} \left\{\frac{1 + \frac{h_0\lambda}{2^{k+1+i}}}{1 - \frac{h_0\lambda}{2^{k+1+i}}}\right\}^{2^{k+i}} \tag{42}$$

It is easy to show that for $M = 1$, $K = 0$, for example, $\lim_{h_0\lambda \to \infty} \beta(h_0\lambda, 1, 0) = 5/3$ and the method loses A-stability.

Single-Step Methods Employing Second Derivatives. Let us now consider the general class of single-step methods employing second derivatives,

$$y_{n+1} = y_n + h\beta_0 y'_{n+1} + h^2\gamma_0 y''_{n+1} + h\beta_1 y'_n + h^2\gamma_1 y''_n \tag{43}$$

for which the characteristic root is

$$\mu_1 = \frac{1 + \beta_1 h\lambda + \gamma_1 h^2\lambda^2}{1 - \beta_0 h\lambda - \gamma_0 h^2\lambda^2} \tag{44}$$

Theorem 2 limits the order of accuracy of A-stable linear multistep methods of the class of Equation 5 to $p \leq 2$. The inclusion of second derivatives in Equation 43 enables us to devise A-stable methods of this class. The same general characteristic root is obtained for the semiimplicit Runge-Kutta methods—e.g., Equation 34. In fact, the characteristic root of Equation 44 is simply a general Padé approximant to exp $(h\lambda)$, the diagonal approximants of which are A-stable (Birkoff and Varga, 1965).

A problem with such A-stable methods is that Equation 44 may still be a poor approximation to exp $(h\lambda_1)$, where λ_1 is the largest eigenvalue. In particular, we not only require A-stability but also want the principal roots $\mu_{1,j}(h\lambda_j)$ to approach exp $(h\lambda_j)$ as $h\lambda_j \to 0$ and $\mu_{1,j}(h\lambda_j)$ to approach zero as $h\lambda_j \to -\infty$. If possible, we want $\mu(h\lambda_j) = \exp(h\lambda_j)$ as $h\lambda_j \to -\infty$. If this were the case, we could say $\mu(h\lambda_j)$ were exponentially fitted to exp $(h\lambda_j)$. This idea was presented by Liniger and Willoughby (1967), who considered two special forms of Equation 43.

$$y_{n+1} = y_n + h[(1 - \beta_1)y'_{n+1} + \beta_1 y'_n] \tag{45}$$

$$y_{n+1} = y_n + \frac{h}{2}[(1 + a)y'_{n+1} + (1 - a)y'_n] - \frac{h^2}{12}[(1 + 3a)y''_{n+1} + (1 - 3a)y''_n] \tag{46}$$

The order of accuracy of Equations 45 and 46 is $p = 1$ and $p = 3$, respectively. A-stability requires that $\beta_1 < 1/2$ for method 1 and $a > 0$ for method 2. Having obtained conditions for A-stability, the next consideration is the accuracy with which the characteristic roots approximate exp $(h\lambda_i)$, $i = 1, 2, \ldots, m$, consistent with these conditions. The conventional procedure is to equate Equation 44 to the Taylor series expansion of exp $(h\lambda)$ about $h\lambda = 0$ and match powers of $h\lambda$ up to the desired order. For the nonstiff eigenvalues this procedure is highly effective. However, for the stiff eigenvalues a Taylor series expansion of exp $(h\lambda_i)$ about $h\lambda_i = 0$ converges very poorly. Thus, as we have noted, conventional methods are inaccurate for stiff systems because they are unable to simulate the stiff components. For stiff systems it is necessary to require accuracy and stability for both large $|h\lambda_i|$ and $h\lambda_i \to 0$.

In Equations 45 and 46 the free parameters β_1 and a can be adjusted to provide accurate representation of the stiff components within the constraint of maintaining A-stability.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In particular, we desire for a certain value of $h\lambda = q_0$ that $\mu_1(q_0) = \exp(q_0)$. We say that the method is exponentially fitted at q_0 if the free parameter is chosen to satisfy this relationship. Requiring that Equations 45 and 46 be exponentially fitted yields

$$\beta_1(q_0) = -q_0^{-1} - (e^{-q_0} - 1)^{-1} \quad (47)$$

$$\alpha(q_0) = \frac{1}{3}[q_0^2 + 6q_0 + 12 - e^{q_0}(q_0^2 - 6q_0 + 12)] \times [e^{q_0}(q_0^2 - 2q_0) + q_0^2 + 2q_0]^{-1} \quad (48)$$

Note from Equation 47 that the trapezoidal rule ($\beta_1 = 1/2$) and the backward Euler method ($\beta_1 = 0$) correspond to exponential fitting at $q_0 = 0$ and $q_0 = -\infty$, respectively. The authors prove that the procedure of exponential fitting is compatible with both A-stability and accuracy in the limit $h\lambda \to 0$ for $0 > q_0 > -\infty$.

Other Methods. Several other methods have been proposed for stiff O.D.E. (Certaine, 1960; Emanuel, 1964; Lomax, 1968a,b; Pope, 1963). Gear (1968) has considered the problem of devising predictor-corrector methods for stiff systems. As in Widlund's case, by requiring less than full left half-plane stability, Gear has devised methods of order as high as 6. Numerical results using Gear's algorithm, reported by Gear (1968) and Ratliff (1968), indicate superiority of the algorithm when compared to classical predictor-corrector methods of comparable order.

Lawson (1967) has devised a class of A-stable explicit Runge-Kutta methods based on a Padé approximation of exp (hA). The method is similar in many respects to the implicit Runge-Kutta methods described previously. Osborne (1968) has considered the problem of designing characteristic roots with the property that $\mu_1 \to (1/h\lambda)$ as $h\lambda \to -\infty$. Dahlquist (1968) has devised an algorithm based on local polynomial approximations. Davison (1968) has considered the problem of the numerical integration of large systems of constant-coefficient linear O.D.E. The poles and zeros of the solution are obtained and the solution constructed in terms of a sum of exponentials. The method is useful in the limited number of cases for which it is applicable. Explicit methods for large stiff systems have been presented by Richards *et al.* (1965) and Fowler and Warten (1967).

Explicit *vs.* Implicit Methods. Ideally a numerical integration method for stiff O.D.E. would possess (1) A-stability, (2) high accuracy, (3) $\mu_1 \to \exp(h\lambda)$ as $h\lambda \to -\infty$, and computational efficiency. Classical explicit methods are highly accurate and computationally efficient, but are not A-stable. Since A-stability is the most important of the above requirements, classical explicit methods are not efficient for stiff O.D.E. because of the rigid conditions on h that must be obeyed. Implicit methods, such as the trapezoidal rule, have two problems associated with the iruse: failure to meet requirement 3 and the necessity to solve a set of nonlinear algebraic equations at each step. If the characteristic root μ_1 of an implicit method is not asymptotic to zero as $h\lambda \to -\infty$, in the initial phase of the solution when the stiff components are nonnegligible, these components will be approximated inaccurately by μ_1. This inaccuracy is reflected as slowly decaying oscillations over the whole range of the solution. For the trapezoidal rule, $\mu_1 \to -1$ as $h\lambda \to -\infty$ and for Calahan's method, Equation 34, $\mu_1 \to -0.735$ as $h\lambda \to -\infty$. Calculations with these methods confirm the existence of oscillations due not to an instability but to an inaccurate simulation of the stiff eigenvalues. There are two ways to circumvent this difficulty. The first is a simple filtering procedure suggested by Dahlquist (1963a) for use with the trapezoidal rule:

Use Equation 12 to compute y_1 and y_2.
Replace y_1 by $(y_0 + 2y_1 + y_2)/4$.
Use Equation 12 to compute y_2 and y_3.
Replace y_2 by $(y_1 + 2y_2 + y_3)/4$.
Use Equation 12 to compute $y_3, y_4, y_5, \ldots$

An alternative to the filtering procedure is to use a small h in the initial phase when the stiff solutions are nonnegligible. Then when these have become negligible, we start using a large h adjusted to the rate of change of the nonstiff components. Even though the stiff components will then be inaccurately simulated, if the stiff components do not reappear, the over-all solution will be accurate.

The second problem associated with implicit methods is that a set of nonlinear algebraic equations must be solved at each step. Even though the method may be A-stable, convergence requirements for the iterative solution of the nonlinear algebraic equations place restrictions on the largest value of h that can be used. These restrictions vary considerably, depending on the particular iterative technique used, but should be far less severe than for conventional explicit methods to retain the advantage of the implicit method. Let us consider this point in more detail.

In particular, we want to solve the implicit form of Equation 5, which can be written

$$y_{n+1} - h\beta_0 f(x_{n+1}, y_{n+1}) - u_n = 0 \quad (49)$$

where u_n includes all the terms independent of y_{n+1}. Let us consider the convergence requirements of four common ways to solve Equation 49: Jacobi iteration, accelerated iteration, Newton-Raphson iteration, and backward iteration.

A solution of Equation 49 by repeated substitutions,

$$y_{n+1}^{(s+1)} - h\beta_0 f(x_{n+1}, y_{n+1}^{(s)}) - u_n = 0 \quad (50)$$

is termed a Jacobi iteration. Let us call y_{n+1}^* the exact solution of Equation 50,

$$y_{n+1}^* - h\beta_0 f(x_{n+1}, y_{n+1}^*) - u_n = 0 \quad (51)$$

Subtracting Equation 51 from Equation 50 and using the mean value theorem,

$$y_{n+1}^{(s+1)} - y_{n+1}^* = h\beta_0[f_y]_{\bar{y}}(y_{n+1}^{(s)} - y_{n+1}^*) \quad (52)$$

where $y_{n+1}^* \leq \bar{y} \leq y_{n+1}^{(s)}$. If we now assume a Lipschitz bound on f_y, $||f_y|| < L$, Equation 52 becomes

$$||y_{n+1}^{(s+1)} - y_{n+1}^*|| \leq h\beta_0 L||y_{n+1}^{(s)} - y_{n+1}^*|| \quad (53)$$

By induction it follows that

$$||y_{n+1}^{(s+1)} - y_{n+1}^*|| \leq (h\beta_0 L)^{s+1}||y_{n+1}^{(0)} - y_{n+1}^*|| \quad (54)$$

A necessary and sufficient condition for convergence of the iterations is then

$$|h\beta_0 L| < 1 \quad (55)$$

or

$$h\beta_0|\lambda_{max}| < 1 \quad (56)$$

This condition is roughly the same as for the classic explicit methods, and thus Jacobi iteration cannot be used efficiently when $|\lambda_{max}|$ is large.

A modification of Jacobi iteration is

$$(1 + \alpha)y_{n+1}^{(s+1)} - h\beta_0 f[x_{n+1}, y_{n+1}^{(s)}] - u_n - \alpha_{n+1}^{(s)} = 0 \quad (57)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

for which α is an acceleration parameter ($\alpha = 0$ is Jacobi iteration). The convergence condition for Equation 57 can be shown to be

$$\left|\frac{h\beta_0\lambda_{max} + \alpha}{1 + \alpha}\right| < 1 \qquad (58)$$

The speed of convergence of Equation 57 can be increased over Equation 50 by proper choice of α.

A well-known method for determining the roots of coupled nonlinear algebraic equations is Newton-Raphson iteration. This method is based on a linearization of f_{n+1} about the previous value of y_{n+1} in the iteration, and is given by

$$y_{n+1}^{(s+1)} = y_{n+1}^{(s)} + [I - h\beta_0 A_{n+1}^{(s)}]^{-1} \times \{h\beta_0 f_{n+1}^{(s)} - y_{n+1}^{(s)} + u_n\} \qquad (59)$$

where $A_{n+1}^{(s)}$ is the Jacobian matrix f_y evaluated at $y_{n+1}^{(s)}$. The necessary condition for convergence of Equation 59 is

$$||[I - h\beta_0 A_{n+1}^{(s)}]^{-1}|| \; || \frac{\partial}{\partial y} \{I - h\beta_0 A_{n+1}^{(s)}\}|| \times ||y_{n+1}^{(s+1)} - y_{n+1}^{(s)}|| \leq 1 \qquad (60)$$

In actual use of Equation 59 it is impractical to recompute A_{n+1} for each iteration, since the time required to invert a large matrix decreases the utility of the method. It is often acceptable to approximate $A_{n+1}^{(s)}$ by $A_{n+1}^{(0)}$. If too many iterations are then required in a given step, h can be reduced and the iteration restarted or A_{n+1} can be re-evaluated.

Finally, we can formulate a backward iteration of the form

$$y_{n+1}^{(s)} = h\beta_0 f[x_{n+1}, y_{n+1}^{(s+1)}] + u_n \qquad (61)$$

the convergence conditions of which can be shown to be

$$(h\beta_0|\lambda_{min}|)^{-1} < 1 \qquad (62)$$

Thus, there is a lower bound on h rather than an upper bound as in the other three methods.

The real question of interest is what technique should be used in conjunction with implicit methods.

Jacobi iteration and accelerated iteration are computationally easy to implement but have convergence requirements depending on the largest eigenvalue of the Jacobian matrix of the O.D.E. Thus, if $|\lambda_{max}|$ is large, an extremely small h is necessary for convergence in these methods. In general, Newton-Raphson iteration has a larger region of convergence than the previous two methods. Backward iteration has a very large region of convergence because of a lower bound on h rather than an upper bound. However, implicit equations still have to be solved in Equation 61.

Thus, we make the following recommendations for the solution of an implicit multistep equation:

1. If the ratio of the largest to the smallest eigenvalue of A is small, say the order of 10, Jacobi iteration or accelerated iteration should be used.
2. If the ratio of the largest to the smallest eigenvalue of A is large, say greater than the order of 10, Newton-Raphson iteration or backward iteration should be used with h selected on the basis of the number of iterations desired per step.

The semiimplicit methods combine the A-stability of implicit methods and the computational efficiency of explicit methods. Because one or more matrix inversions are necessary per step, these methods can be shown to be equivalent to corresponding implicit methods with several applications of Newton-Raphson iteration.

The choice of a method becomes critical for large stiff systems ($m > 20$). Current methods described above usually require evaluation of the Jacobian matrix or its eigenvalues. A method suitable for large stiff systems must not require eigenvalue determination—for example, computing times for standard matrix inversion and eigenvalue determination of an $m \times m$ matrix on an IBM 7094 are:

m	10	20	30	40	50
Inversion, sec.	0.083	0.533	1.684	3.917	7.567
Eigenvalues, sec.	1.5	11.2	60	...	...

Implicit methods usually require at least one matrix inversion per step, if the Jacobian matrix is re-evaluated at each step. Only a few elements of the matrix are likely to change significantly from step to step, so it may be possible to update the matrix at each step using only a small number of derivative evaluations.

Examples

The following three problems will be studied:

1. $y'(x) = -200[y - F(x)] + F'(x)$

   $F(x) = 10 - (10 + x)e^{-x}$

   $y(0) = 10$

   Exact solution $y(x) = 10 - (10 + x)e^{-x} + 10e^{-200x}$

2. Equation 1 with $y(0) = [2, 1, 2]^T$ and

$$A = \begin{bmatrix} -0.1 & -49.9 & 0 \\ 0 & -50 & 0 \\ 0 & 70 & -120 \end{bmatrix}$$

   Exact solution $y_1(x) = e^{-0.1x} + e^{-50x}$

   $y_2(x) = e^{-50x}$

   $y_3(x) = e^{-50x} + e^{-120x}$

3. $y'_1 = -0.04y_1 + 10^4 y_2 y_3$

   $y'_2 = 0.04y_1 - 10^4 y_2 y_3 - 3 \times 10^7 y_2^2$

   $y'_3 = 3 \times 10^7 y_2^2$

   $y_1(0) = 1; y_2(0) = 0; y_3(0) = 0$

System 3 represents a system of reaction rate equations (Robertson, 1967), y_1, y_2, and y_3 representing the mole fractions of the three species. The Jacobian matrix of the system is

$$A = \begin{bmatrix} -0.04 & 10^4 y_3 & 10^4 y_2 \\ 0.04 & -10^4 y_3 - 6 \times 10^7 y_2 & -10^4 y_2 \\ 0 & 6 \times 10^7 y_2 & 0 \end{bmatrix}$$

which is singular. The three eigenvalues of A are given by

$$\lambda_1 = 0$$

$$\lambda^2 + (0.04 + 10^4 y_3 + 6 \times 10^7 y_2)\lambda + (0.24 \times 10^7 y_2 + 6 \times 10^{11} y_2^2) = 0$$

At $x = 0$, the three eigenvalues are $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = -0.04$. The asymptotic behavior of the system for large x is to $y_1 = 0$ and $y_2 = 0$, and $y_3 = 1$, since $\Sigma y_i = 1$ always. For large values of x the eigenvalues approach $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = -10^4$. In fact, from $x = 0$ to $x = 0.02$, λ_3 changes from -0.04 to -2405. The system is thus highly stiff.

The methods used for the three examples are presented in Table II.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table II. Methods Used for Stiff Systems

Method	Designation	Comments
1. 4th-order Runge-Kutta	RK4	
2. Adams 4th-order P-C	DEQ	1 corrector evaluation, RK4 start
3. Treanor's method	TM	Automatic control of h
4. Modified midpoint rule	DIFSYS	
5. Trapezoidal rule	TR	Initial filtering procedure used
6. Trapezoidal rule with extrapolation	TR-EX	Initial filtering. $M = 3$
7. Calahan's method	CAL	Adjustment of h
8. Equation 45	LW1	
9. Equation 46	LW2	

Table III. Computational Results for Example 1

Method	h	R_n at $x = 0.4$	R_n at $x = 1.0$	Time, Sec
RK4	0.01	1.0×10^{-5}	2.0×10^{-9}	11
DEQ	0.005	3.0×10^{-9}	2.0×10^{-9}	18
TM	0.2[a]	6.7×10^{-6}	1.0×10^{-9}	16.5
DIFSYS	[b]	[c]	[c]	[c]
TR	0.2	1.85×10^{-3}	4.3×10^{-6}	2
TR-EX	0.2	1.4×10^{-4}	1.0×10^{-8}	36
CAL	0.01/0.2[d]	1.7×10^{-3}	4.0×10^{-8}	1
LW1	0.2	1.1×10^{-3}	5.0×10^{-8}	3
LW3	0.2	1.8×10^{-3}	9.0×10^{-8}	4

[a] Automatic step size control.
[b] Initial step size 0.1, extrapolations performed until error $<10^{-6}$.
[c] Unstable.
[d] h changed from 0.01 to 0.2 at $x = 0.1$.

Example 1 is a single O.D.E. with a solution containing a rapidly decaying component and a slowly decaying component. The eigenvalue is -200, and the solution is desired from $x = 0$ to $x = 15$. Thus, the solution component $\exp(-200x)$ becomes negligible almost immediately compared to the $\exp(-x)$ component. The results of the numerical integration of Example 1 are presented in Table III. The R_n columns show $\frac{y_n - y(x_n)}{y(x_n)}$ at two values of x, 0.4 and 10. The time column indicates the total computation time in seconds on an IBM 7094. The two points $x = 0.4$ and $x = 10$ were chosen as representative of the errors early and late in the integration. For a stable method we expect that the solution will be accurate at $x = 10$, where the extraneous solutions have become negligible. All computations were performed in a single precision and an entry of zero in the error columns indicates eight-place accuracy.

RK4 and DEQ are both highly accurate but time-consuming. The automatic stepsize selection routine in TM in this case determined an h comparable to RK4 and actually required more time. DIFSYS was unstable for h values comparable to RK4 and DEQ. TR, CAL, LW1, and LW3 were roughly comparable. TR-EX was more accurate, as expected, but required an excessive amount of time.

Example 2 has eigenvalues -120, -50, and -0.1. This system is interesting because it contains two stiff eigenvalues, so that three different characteristic times appear. Results of the integration of Example 2 are shown in Table IV. Again it was desired to integrate from $x = 0$ to $x = 15$ and errors are tabulated at $x = 0.4$ and $x = 10$. TM was somewhat less accurate than RK4, but required only one second compared to 20 seconds for RK4. DIFSYS was again unstable, as evidenced by R_n at $x = 10$. CAL and LW1 were roughly comparable and somewhat more accurate than TR.

Example 3 is a very stiff set of nonlinear O.D.E. As noted, the eigenvalues change from 0, 0, -0.04 to 0, 0, -10^4 over the range $x = 0$ to $x = 40$, and most of this change occurs in the first few instants. Thus, this example represents the severest test of the methods of all the examples. The results of the methods on Example 3 are shown in Table V. All of the explicit methods eventually become unstable. With $h = 0.001$ RK4 became unstable after $x \geq 16$, where $|\lambda_{max}| \cong 2.78 \times 10^3$. On the basis of the time to compute to $x = 10$, RK4 would require 138 seconds to get to $x = 40$ (if it were stable). DEQ with $h = 0.001$ was unstable after $x = 0.012$. DIFSYS with $h = 0.001$ was unstable after $x = 0.358$. TM was strongly influenced by the off-diagonal elements and was completely unstable.

The semiimplicit method CAL was stable with $h = 0.005$ up to $x = 1$ and $h = 0.02$ for $x > 1$. However, slowly occurring oscillations could not be avoided because of the asymptotic root behavior of the method. For $h = 0.05$ for $x > 1$ the numerical solution converged to the wrong values without exhibiting oscillatory behavior.

The full implicit methods, TR, TR-EX, LW1, and LW3, were most applicable. Even though these methods are all A-stable, the necessity to solve nonlinear algebraic equations presented limitations on the size of h (as well, of course, as

Table IV. Computational Results for Example 2

Method	h	R_{1n} at $x = 0.4$	R_{1n} at $x = 10$	R_{2n} at $x = 0.4$	R_{3n} at $x = 0.4$	Time, Sec
RK4	0.01	2.0×10^{-7}	5.4×10^{-7}	3.0×10^{-1}	3.0×10^{-1}	20
DEQ	0.01	2.0×10^{-8}	8.1×10^{-7}	9.5×10^{-1}	7.4×10^{5}	23
TM	0.2[a]	4.0×10^{-4}	1.35×10^{-4}	1.1×10^{5}	1.2×10^{5}	1
DIFSYS	[b]	5.0×10^{-4}	2.16×10^{-4}	9.4×10^{-1}	8.3×10^{3}	22
TR	0.2	1.0×10^{-3}	2.7×10^{-4}	6.5×10^{7}	1.3×10^{5}	1.3
TR-EX	0.2	4.0×10^{-6}	8.1×10^{-7}	5.7×10^{1}	8.0×10^{1}	30
CAL	0.01/0.2[c]	2.0×10^{-3}	2.7×10^{-4}	2.5×10^{5}	1.6×10^{5}	1
LW1	0.2	4.0×10^{-3}	1.1×10^{-3}	5.0×10^{5}	5.0×10^{5}	3

[a] Automatic step size control.
[b] Initial step size 0.1, extrapolations performed until error $<10^{-6}$.
[c] h changed from 0.01 to 0.2 at $x = 0.1$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table V. Computational Results for Example 3

Method	h	R_{1n} at $x = 0.4$	R_{1n} at $x = 10$	R_{2n} at $x = 0.4$	R_{2n} at $x = 10$	R_{3n} at $x = 0.4$	R_{3n} at $x = 10$	Time, Sec[c]
RK4	0.001	0	0	0	0	0	0	[b]
DEQ	0.001	[b]	[b]	[b]	[b]	[b]	[b]	[b]
TM	0.01[c]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
DIFSYS	0.001[d]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
TR	0.2	1.35×10^{-3}	1.05×10^{-3}	2.12×10^{-1}	2.4×10^{-1}	9.0×10^{-2}	1.5×10^{-1}	9.3
TR-EX	0.2	1.72×10^{-6}	3.6×10^{-6}	3.5×10^{-3}	4.3×10^{-6}	6.8×10^{-4}	1.2×10^{-3}	34
CAL	0.005/0.02 at $x = 1.0$	2.4×10^{-3}	1.01×10^{-1}	2.5×10^{0}	6.0×10^{-1}	1.62×10^{-1}	5.4×10^{-1}	10
LW1	0.02	1.6×10^{-4}	4.9×10^{-4}	2.4×10^{-4}	1.3×10^{-4}	3.2×10^{-2}	4.4×10^{-4}	20
LW3	0.02	5.9×10^{-4}	7.1×10^{-5}	2.9×10^{-3}	1.1×10^{-3}	4.0×10^{-3}	1.9×10^{-3}	23.3

[a] IBM 360-75.
[b] Unstable.
[c] Automatic step size control.
[d] Initial step size 0.1, extrapolations performed until error $<10^{-4}$.

accuracy considerations)—for example, convergence of the Newton-Raphson method used with TR required $h \leq 0.25$. The necessary condition for convergence of the Newton-Raphson iteration, Equation 60, predicts that h must be less than $10^{-4}\beta_0^{-1}||y_{n+1}^{(s+1)} - y_{n+1}^{(s)}||$, which is highly conservative. More improved convergence conditions for Newton-Raphson iteration are apparently a topic of current study. The initial filtering procedure was effective in eliminating oscillations due to inaccuracy for $h \leq 0.1$. The total time for $x = 0$ to $x = 40$ was 9.3 seconds. For the increased accuracy of TR-EX, 34 seconds were required. Because of the size of the stiff eigenvalues, the exponential filtering procedure in LW1 and LW3 caused a computer overflow, since the largest exponential argument that can be handled is 174.673. Thus, the allowable maximum h is 0.02 in each method. Computing times were 20 and 23.3 seconds, respectively.

From the examples the following conclusions can be drawn:

1. RK4 and DEQ are both highly accurate, DEQ requiring more computation time because of the small absolute stability bound. Because of the small finite stability bound, it is not recommended that either of these methods be used for stiff O.D.E.
2. DIFSYS has even less desirable stability properties than RK4 and DEQ, confirming our knowledge of the poor stability properties of the midpoint rule. Of all the methods used, DIFSYS is the least desirable for stiff O.D.E.
3. TM with automatic control of h is generally not effective as an all-purpose stiff routine. TM usually decreases the time from that required by RK4 for comparable accuracy; however, in some cases, the automatic step size control may decrease h to values comparable to RK4. In addition, its utility is limited to those O.D.E. with only large diagonal elements in the Jacobian.
4. The four implicit methods studied, TR, CAL, LW1, and LW3, were roughly comparable in terms of accuracy and computing time, and each resulted in significant savings of time over the explicit methods. TR-EX resulted in the highest accuracy in each example but at the expense of considerable computing time. For systems that are only moderately stiff, CAL is slightly more accurate than the other three. However, for highly stiff systems, LW1, with proper scaling to avoid computer overflows, appears to be the best even though it is the least accurate. If exponential fitting is not used, either the initial filtering procedure or the step length adjustment is necessary in TR and CAL to prevent oscillations from inaccurate simulation of the stiff eigenvalues.

Nomenclature

a_i = constants in general root form Equation 20
a = adjustable parameter in Equation 46
$\mathbf{A}$ = Jacobian matrix of O.D.E.
b_i = constants in general root form Equation 20
$\mathbf{B}$ = matrix defined in Equation 21
$c_{i,j}$ = coefficients governed by Equation 40
c_i = constants in Equation 10
d_i = constants in Equation 9
$D_{i,j}$ = parameters in Equation 28
$\mathbf{f}$ = m-dimensional vector function
F_i = parameters defined in Equation 30
h = step length
$\mathbf{I}$ = identity matrix
k = degree of multistep method
$\mathbf{k}_i$ = parameters in Equation 33
K = number of step length sequences used in extrapolation
L = Lipschitz constant
m = dimensionality of O.D.E.
M = number of extrapolations
$\mathbf{M}(h\mathbf{A})$ = characteristic matrix of single-step method
n = step index
N = number of steps in interval of integration
p = order of accuracy of method
P_i, P = parameters in Treanor's method
$\mathbf{P}$ = matrix in similarity transformation, Equation 14
q_i, q_0 = $h\lambda$
u_n = variable in Equation 49
v_i = parameters in Equation 31
w_1, w_2 = weighting coefficients in Equation 33
x = independent variable
$\mathbf{y}$ = dependent variable
$Y_m^{(k)}$ = extrapolation value of y
$\mathbf{z}$ = dependent variable from similarity transform

Greek Letters

α = acceleration parameter
α_i = coefficients in linear multistep method, Equation 5
β = characteristic root in extrapolation
β_i = coefficients in linear multistep method, Equation 5
γ_0, γ_1 = coefficients in Equation 43
ϵ_n = accumulated error at step n
$\epsilon(x_n)$ = magnified error function at step n
θ = angle
λ_i = eigenvalues of O.D.E.
Λ = diagonal matrix of eigenvalues
μ_i = characteristic roots of multistep method

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Literature Cited

Allen, R. H., "Numerically Stable Explicit Integration Techniques Using a Linearized Runge-Kutta Extension," Boeing Scientific Res. Lab. Document D1-82-0929 (October 1969).
Amundson, N. R., *Can. J. Chem. Eng.* 43, 49 (1965).
Amundson, N. R., Luss, D., *Can. J. Chem. Eng.* 46, 425 (1968).
Bauer, F. L., Rutishauser, H., Stiefel, E., "New Aspects in Numerical Quadriture," Proceedings Symposium on Applied Mathematics, Vol. 15, American Mathematical Society, p. 199, 1963.
Birkoff, G., Varga, R. S., *J. Math. Phys.* 44, 1 (1965).
Brayton, R. K., Gustavson, F. G., Liniger, W., *IBM J. Res. Develop.* 10, 4, 292 (1966).
Bulirsch, R., Stoer, J., *Numer. Math.* 8, 1 (1966).
Butcher, J. C., *J. Aust. Math. Soc.* 4, 179 (1964).
Calahan, D. A., *Proc. IEEE (Letters)* 55, 2016 (1967).
Calahan, D. A., *Proc. IEEE (Letters)* 56, 744 (1968).
Calahan, D. S., Abbott, N. E., "Stability Analysis of Numerical Integration," Proceedings of 10th Midwest Symposium on Circuit Theory, Lafayette, Ind., May 1967.
Certaine, T., "Solution of Ordinary Differential Equations with Large Time Constants," "Mathematical Methods for Digital Computers," Chap. 11, p. 128, Wiley, New York, 1960.
Dahlquist, G., *BIT* 3, 27 (1963a,b).
Dahlquist, G., "Numerical Method for Some Ordinary Differential Equations with Large Lipschitz Constants," Proceedings of IFIP Congress, Supplement, Booklet I, p. 32, 1968.
Dahlquist, G., *Math. Scan.* 4, 33 (1956).
Dahlquist, G., "Stability Questions for Some Numerical Methods for Ordinary Differential Equations," Proceedings of Symposium on Applied Mathematics, Vol. 15, p. 147, American Mathematical Society, 1963b.
Davis, P. J., Rabinowitz, R., "Numerical Integration," p. 166, Blaisdell, Waltham, Mass., 1967.
Davison, E. J., *Computer J.* 10, 195 (1967).
Davison, E. J., *A.I.Ch.E.J.* 14, 1, 46 (1968).
Distefano, G. P., *A.I.Ch.E.J.* 14, 946 (1968).
Emanuel, G., "Numerical Analysis of Stiff Equations," Aerospace Corporation, El Segundo, Calif., SSD-TDR-63-380 (1964).
Emanuel, G., "Problems Underlying the Numerical Integration of the Chemical and Vibrational Rate Equations in a Near-Equilibrium Flow," AEDC-TDR-63-82 (1963).
Eschenroeder, A. Q., Boyer, D. W., Hall, G. J., *Phys. Fluids* 5, 5, 615 (1962).
Fowler, M. E., Warten, R. M., *IBM J. Res. Develop.* 11, 537 (1967).
Gear, C. W., "Automatic Integration of Stiff Ordinary Differential Equations," Proceedings of IFIP Congress, Supplement, Booklet A, 81 (1968).
Gragg, W. B., *J. SIAM Numer. Anal.* 2, 384 (1965).
Henrici, P., "Discrete Variable Methods in Ordinary Differential Equations," p. 209, Wiley, New York, 1962.
Henrici, P., "Elements of Numerical Analysis," p. 271, Wiley, New York, 1964.
Hildebrand, F. B., "Introduction to Numerical Analysis," p. 206, McGraw-Hill, New York, 1956.
Kalman, R. E., "Toward a Theory of Difficulty of Computation in Optimal Control," Proceedings of IBM Scientific Comp. Symposium on Control Theory and Applications, 1966.
Kelly, L. G., "Handbook of Numerical Methods and Applications," p. 281, Addison-Wesley, Reading, Mass., 1967.
Lawson, J. D., *J. Soc. Ind. Appl. Math. Numer. Anal.* 4, 372 (1967).
Liniger, W., Willoughby, R. A., "Efficient Numerical Integration of Stiff Systems of Ordinary Differential Equations," IBM Research Report RL-1970 (December 1967).
Lomax, H., "Construction of Highly Stable, Explicit Numerical Methods for Integrating Coupled Ordinary Differential Equations with Parasitic Eigenvalues," NASA Tech. Note NASA TN D-4547 (April 1968a).
Lomax, H., Bailey, H. E., "Critical Analysis of Various Numerical Integration Methods for Computing Flow of a Gas in Chemical Nonequilibrium," NASA TN D-4109 (1967).
Lomax, H., "Stable Implicit and Explicit Numerical Methods for Integrating Quasi-linear Differential Equations with Parasitic-Stiff and Parasitic-Saddle Eigenvalues," NASA TN D-4903 (1968b).
Mah, R. S. H., Michaelson, S., Sargent, R. W. H., *Chem. Eng. Sci.* 17, 619 (1962).
Makinson, G. J., *Computer J.* 11, 305 (1968).
McCalla, T. R., "Introduction to Numerical Methods and FORTRAN Programming," p. 341, Wiley, New York, 1967.
Osborne, M. R., "New Method for Integration of Stiff Systems of Ordinary Differential Equations," Proceedings of IFIP Congress, Mathematical Booklet A, pp. 86-90, 1968.
Pope, D. A., *Comm. ACM* 6, 8, 491 (1963).
Ratliff, K., "Comparison of Techniques for the Numerical Integration of Ordinary Differential Equations," University of Illinois, Dept. Computer Science, Rept. 274 (July 8, 1968).
Richards, P. I., Lanning, W. D., Torrey, M. D., *SIAM Rev. (Soc. Ind. Appl. Math.)* 7, 3, 376 (1965).
Robertson, H. H., "Solution of a Set of Reaction Rate Equations," in Numerical Analysis, J. Walsh, ed., p. 178, Thompson Book Co., Washington, 1967.
Rosenbrock, H. H., *Computer J.* 5, 320-330 (1963).
Thompson, W. E., *Computer J.* 10, 417 (1967).
Treanor, C. E., *Math. Comp.* 20, 39-45 (1966).
Widlund, O. E., *BIT* 7, 65 (1967).

RECEIVED for review April 25, 1969
ACCEPTED February 24, 1970

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

P-II

Control of Plug-Flow Tubular Reactors by Variation of Flow Rate

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reprinted from I&EC FUNDAMENTALS, Vol. 9 Page 651, November 1970
Copyright 1970 by the American Chemical Society and reprinted by permission of the copyright owner

Control of Plug-Flow Tubular Reactors by Variation of Flow Rate

John H. Seinfeld,[1] George R. Gavalas, and Myungkyu Hwang
Department of Chemical Engineering, California Institute of Technology, Pasadena, Calif. 91109

The control of isothermal and adiabatic plug-flow tubular reactors by variation of flow rate was studied. Proportional feedback, feedforward, and optimal control responses were compared for the regulation of reactor conversion in the presence of inlet disturbances. The optimal control, consisting of a singular solution in each case, produces a considerably improved response over both feedforward and proportional feedback control.

The control of tubular reactors is a problem of considerable importance in chemical processing. Based on the mode of operation—e.g., isothermal, adiabatic, etc.—control can be exercised in a variety of ways—e.g., flow rate variation, inlet condition variation, heating or cooling rate variation, etc. From the standpoint of control, a convenient method of classification is by the form of the mathematical model used to describe the reactor. Reactor models can generally be placed in two categories: *hyperbolic systems, in which axial and* radial diffusion effects are neglected (plug-flow); and parabolic systems, in which diffusion effects are included.

In the present study we consider both isothermal and adiabatic plug-flow reactors for which the control objective is to maintain the outlet composition at a desired value in the presence of inlet concentration and temperature fluctuations. In the isothermal case, control can be exercised by variation of the flow rate and the temperature. In the adiabatic case, control can be exercised by variation of the flow rate and the inlet temperature, assuming that the inlet concentration is not available for adjustment. Ogunye and Ray (1970) determined the optimal temperature control policy in both the isothermal and adiabatic plug-flow cases in the presence of catalyst decay. We consider the other alternative for plug-flow reactor control—namely, control of the flow rate. This mode of control is of practical importance, since flow rate is an easily manipulated variable.

Manipulation of the flow rate of an isothermal plug-flow reactor to control the exit composition was considered by Koppel (1966a,b). Since Koppel based his feedback proportional law on a transformed variable rather than directly on the outlet concentration, his results are not generally applicable. In fact, as a result of using a transformed variable, the control no longer has a linear relationship to the error and is not proportional as stated.

The objectives of this work are the following. We wish to solve directly the nonlinear problem for the dynamic response of the isothermal reactor with proportional control of the flow rate. Then, we wish to determine the optimal flow rate control policy for both the isothermal and adiabatic cases that minimizes the integral square error of the outlet concentration for a given inlet disturbance. Finally, the optimal response is compared to the proportional feedback and simple feedforward responses to determine the degree of improvement achieved by optimal control.

[1] To whom correspondence should be sent.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

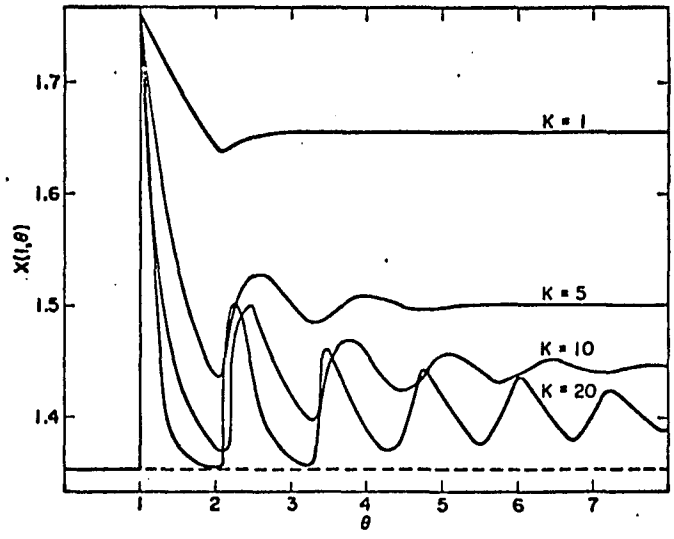

Figure 1. Dynamic response of outlet concentration for $A = 0.3$, $\beta = 2$, and $n = 1$

Proportional control

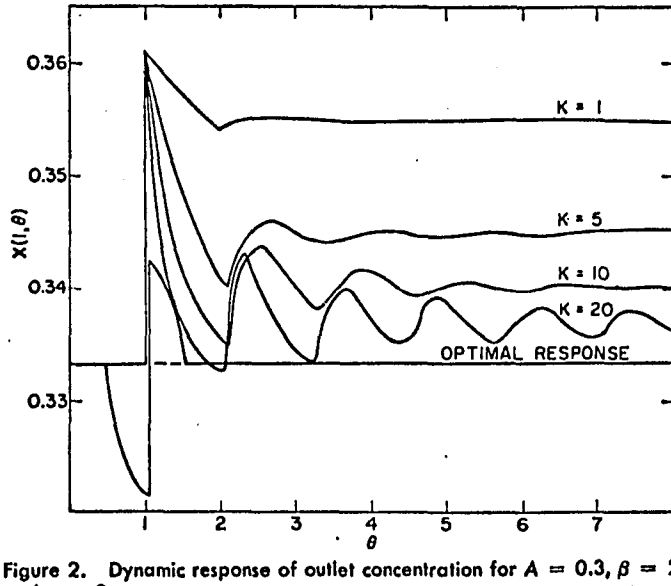

Figure 2. Dynamic response of outlet concentration for $A = 0.3$, $\beta = 2$, and $n = 2$

Proportional and optimal control

Isothermal Case

Proportional Control. The dynamics of an isothermal plug-flow, tubular reactor with an nth-order irreversible reaction and proportional control of flow rate is described in dimensionless terms by

$$\frac{\partial x(\theta,\eta)}{\partial \theta} + \{1 - K[x(\theta,1) - x^d]\}\frac{\partial x(\theta,\eta)}{\partial \eta} = -\beta x(\theta,\eta)^n \quad (1)$$

where the concentration sensing takes place at the reactor outlet, $\eta = 1$, and the desired outlet concentration, $x(\theta, 1)$, is x^d. We assume that for $\theta < 0$ the reactor is in a steady state for which $x(1) = x^d$, so that the control is shut off. The inlet concentration is assumed to undergo a step change of magnitude A at $\theta = 0$ from its steady-state value of 1,

$$x(\theta,0) = 1 + A \qquad \theta > 0 \quad (2)$$

The object of this section is to determine the exact dynamical response of the reactor for fixed values of β, n, A, and x^d and different values of the gain, K. The numerical technique based on the method of characteristics for obtaining the solution of Equation 1 is described by Hwang (1968).

The exit response $x(\theta,1)$ is shown in Figure 1 for $\beta = 2$, $A = 0.3$, $n = 1$, and $K = 1, 5, 10, 20$. Similar responses are shown in Figure 2 for $\beta = 2$, $A = 0.3$, $n = 2$, and $K = 1, 5, 10$, and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

20. The magnitude of the offset, defined as the difference between the asymptotic outlet concentration $x(\infty,1)$ and x^d, can be determined from

$$\{1 - K[x(1,\infty) - x^d]\} \ln\left[\frac{x(\infty,1)}{1+A}\right] + \beta = 0 \quad (n = 1) \quad (3)$$

$$x(\infty,1) - x^d = -\frac{-\alpha_1 + (\alpha_1^2 - 4\alpha_2)^{1/2}}{K} \quad (n = 2) \quad (4)$$

where

$$\alpha_1 = 1 + \beta(1 + A) - K(x^d - A - 1) \quad (5)$$

$$\alpha_2 = -K[x^d(1 + \beta + \beta A) - (1 + A)] \quad (6)$$

In each case, as K is increased the offset is decreased. With larger K, the system undergoes more rapid oscillations before reaching the asymptotic value. Gain K cannot be chosen arbitrarily large, because the total velocity must be greater than zero. The maximum allowable value of K can be determined from this requirement as

$$K_{max} = \frac{e^\beta}{A} \quad (n = 1) \quad (7)$$

$$K_{max} = \left[\frac{1+A}{1+\beta(1+A)} - \frac{1}{1+\beta}\right]^{-1} \quad (n = 2) \quad (8)$$

since the maximum deviation $x(\theta,1) - x^d$ occurs when $\theta = 1$—for example, for $n = 2$, $\beta = 2$, and $A = 0.3$, $K_{max} = 36$.

If a pure time delay of magnitude τ exists in the control loop, $x(\theta,1)$ in Equation 1 is replaced by $x(\theta - \tau, 1)$. The numerical technique used can be extended to include this case; however, these results are not reported here.

Feedforward Control. An alternative to feedback proportional control is simple feedforward control, in which as soon as the step change in inlet concentration is sensed, the flow rate is changed to the steady-state value corresponding to the new inlet concentration which will produce the same outlet concentration. The response of $x(\theta,1)$ in this case is shown in Figure 3 for $n = 2$. By comparison to Figure 2, we see that the speed of response has been improved considerably over proportional control, mainly because the residence time lag in the reactor has been avoided.

Optimal Control. It is of interest to determine the optimal open-loop flow rate policy and response and compare to the closed-loop proportional and the simple feedforward responses. Let us rewrite Equation 1 as

$$\frac{\partial x(\theta,\eta)}{\partial\theta} + v(\theta)\frac{\partial x(\theta,\eta)}{\partial\eta} = -\beta x(\theta,\eta)^n \quad (9)$$

We formulate the optimal control problem as follows: It is desired to determine $v(\theta)$ over the given time interval $(0,\theta_f)$ subject to $v^* > v(\theta) > v_*$, the maximum and minimum allowable flow rates, such that the integral square error

$$P = \int_0^{\theta_f} [x(\theta,1) - x^d]^2\, d\theta \quad (10)$$

is minimized. By application of the necessary conditions for optimality for distributed parameter systems (Katz, 1964; Koppel *et al.*, 1968; Seinfeld and Lapidus, 1968a), the optimal policy is found to be

$$v(\theta) = \begin{cases} v^* & G(\theta) < 0 \\ v_* & G(\theta) > 0 \end{cases} \qquad G(\theta) = \int_0^1 p(\theta,\eta)\frac{\partial x(\theta,\eta)}{\partial\eta}\, d\eta \quad (11)$$

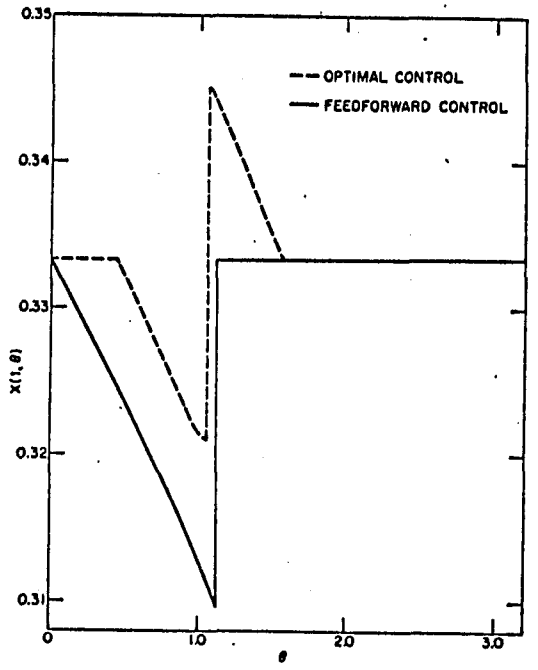

Figure 3. Comparison of outlet concentration responses with feedforward and optimal control

Isothermal reactor

where the adjoint variable $p(\theta,\eta)$ is governed by

$$\frac{\partial p(\theta,\eta)}{\partial\theta} + v(\theta)\frac{\partial p(\theta,\eta)}{\partial\eta} = 2\delta(\eta - 1)[x(\theta,1) - x^d] + n\beta p(\theta,\eta)x(\theta,\eta)^{n-1} \quad (12)$$

$$p(\theta,1) = 0 \qquad p(\theta_f,\eta) = 0 \quad (13)$$

The two-point boundary value problem represented by Equations 1, 2, 11, 12, and 13 cannot be solved analytically. In cases when the optimal control is given by a bang-bang law, the switching times can be determined most easily by the method of direct search on the performance index (Seinfeld and Lapidus, 1968a).

One complication can arise in optimal bang-bang control—that is, if $G(\theta) = 0$ on a finite time interval, a singular arc results and $v(\theta)$ may be undefined. Previous work (Seinfeld and Lapidus, 1968b) has shown that the direct search method is particularly effective for determining optimal singular controls, especially when the value of the control on the singular arc can be determined.

The direct search is simply a systematic search over a number M of preselected control values until the value of P can no longer be decreased.

If the existence of a singular solution can be ruled out *a priori*, then $M = 2$ with the two choices as v^* and v_*. In the present case, however, the possibility of a singular solution cannot be ruled out, since the two-point boundary value problem of Equations 1, 2, 11, 12, and 13 cannot be solved analytically. In fact, hyperbolic optimal control problems of this type have been shown to involve terminal singular arcs (Koppel, 1967; Seinfeld and Lapidus, 1968a). The terminal singular control $v(\theta)$ in these cases corresponds to the simple

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

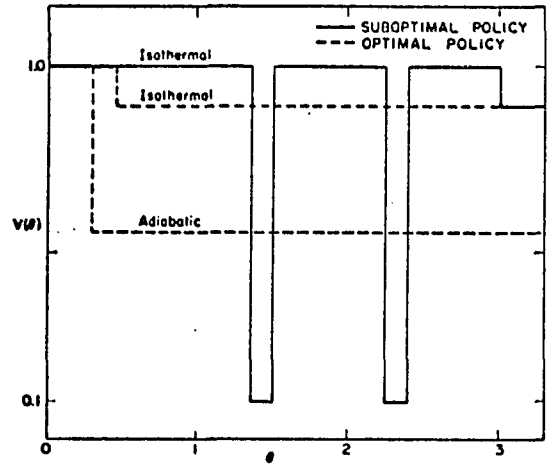

Figure 4. Velocity policies in suboptimal and optimal cases

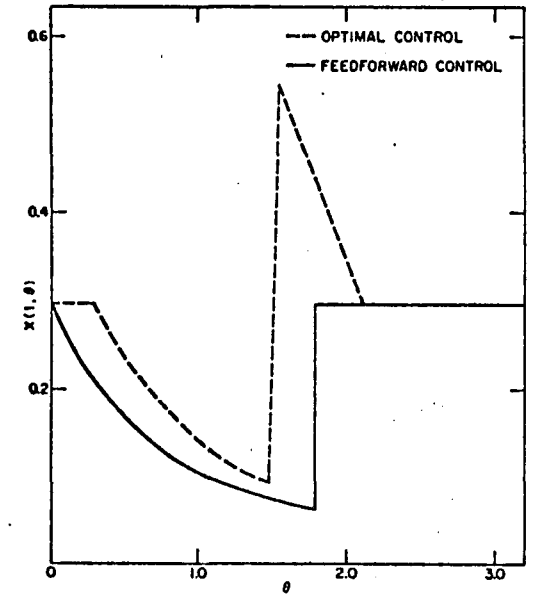

Figure 5. Comparison of outlet concentration responses with feedforward and optimal control

Adiabatic case

feedforward value obtained by setting $x(\theta,1) = x^d$. In the present example, $v(\theta) = 0.895$ for $n = 2$.

The following computations were performed for $n = 2$, $\beta = 2$, $A = 0.3$, $v^* = 1.0$, $v_* = 0.1$, and 20 time increments.

1. $M = 2$ $v = 0.1, 1.0$
2. $M = 5$ $v = 0.1, 0.3, 0.6, 0.895, 1.0$

Case 1 was carried out without regard to the existence of a singular arc. Case 2 included several control values, one of which was the feedforward flow rate value of 0.895. Figure 4 presents the results from cases 1 and 2. The minimum value of P was achieved for the policy labeled optimal, indicating the existence of a terminal singular arc for $\theta > 0.45$. The value of the direct search in handling singular solutions is evident, since the singular control value can be used directly in the search. The outlet response $x(\theta,1)$ corresponding to the optimal flow rate policy is shown in Figures 2 and 3 for comparison to the two previous modes of control.

The value of the performance index, P, in the simple feedforward case with $v(\theta) = 0.895$, $\theta > 0$, is 2.187×10^{-4}, whereas the value of P for optimal control is 6.309×10^{-5}. This provides a quantitative measure of the improvement gained by optimal control over simple feedforward control, each of which is decidedly superior to proportional feedback control.

Adiabatic Case

Feedforward and Optimal Control. The time-dependent behavior of an adiabatic plug-flow reactor with an nth-order irreversible reaction is described in dimensionless terms by

$$\frac{\partial x(\theta,\eta)}{\partial \theta} + v(\theta)\frac{\partial x(\theta,\eta)}{\partial \eta} = -\phi \exp\left[\psi - \frac{\gamma}{T(\theta,\eta)}\right] x(\theta,\eta)^n \quad (14)$$

$$\frac{\partial T(\theta,\eta)}{\partial \theta} + v(\theta)\frac{\partial T(\theta,\eta)}{\partial \eta} = \omega \exp\left[\psi - \frac{\gamma}{T(\theta,\eta)}\right] x(\theta,\eta)^n \quad (15)$$

For $\theta < 0$ the reactor is in a steady state for which $x(1) = x^d$. The inlet concentration is assumed to undergo a step change of magnitude A at $\theta = 0$ as in Equation 2. The inlet temperature is assumed to undergo a step change of magnitude B at $\theta = 0$,

$$T(\theta,0) = 1 + B \quad \theta > 0 \quad (16)$$

The following values of the parameters were chosen: $\phi = 3$, $\psi = 38$, $\gamma = 40$, $\omega = 0.35$, $n = 2$, $A = 0.3$, and $B = -0.03$. With these values, $x^d = 0.02962$.

Since it was shown in the isothermal case that proportional control compares poorly with even simple feedforward control, only feedforward and optimal control were examined in the adiabatic case. Feedforward control consists of setting v equal to the new steady-state value corresponding to x^d as soon as the inlet disturbances are sensed. The value of $v(\theta)$ corresponding to the parameters used is 0.5556. The response to feedforward control is shown in Figure 5. The value of P in this case is 5.372×10^{-4}.

The optimal $v(\theta)$ policy was determined by the direct search on the performance index (Figure 4). Again, there is a terminal singular solution for $\theta > 0.3$ with $v(\theta)$ equal to the new steady-state value. The response to the optimal flow rate policy is shown in Figure 5, and the value of P in this case is 3.775×10^{-4}. The advantage in using optimal rather than simple feedforward can be seen by comparing the responses in Figure 5 as well as the values of P obtained. This improvement is not as pronounced as in the previously examined isothermal case.

Summary

A direct comparison of feedback, feedforward, and optimal flow rate control has been presented for isothermal and adiabatic plug-flow reactors with a single reaction. Applications of this work would be important in the control of liquid and gas phase reactions carried out in flow reactors—e.g., nitration of aromatic compounds and pyrolysis of lower paraffins.

In both isothermal and adiabatic operation, optimal control produced a considerably better response than simple feedforward control, and both modes were far superior to feedback

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

control. The optimal flow rate policy in each case had a terminal singular arc corresponding to the new steady value of $v(\theta)$.

Nomenclature

A = dimensionless inlet concentration change
B = dimensionless inlet temperature change
$G(\theta)$ = switching function
K = proportional gain
M = number of control values in direct search
n = reaction order
P = performance index
$p(\theta,\eta)$ = adjoint variable
$T(\theta,\eta)$ = dimensionless temperature
$v(\theta)$ = dimensionless velocity
$x(\theta,\eta)$ = dimensionless concentration

Greek Letters

α_1,α_2 = constants
β = dimensionless reaction group
γ = dimensionless activation energy
$\delta(\cdot)$ = Dirac delta function
η = dimensionless spatial variable
θ = dimensionless time
ϕ = dimensionless frequency factor
ψ = dimensionless constant
ω = dimensionless heat generation constant

Superscripts

d = desired
$*$ = maximum

Subscript

$*$ = minimum

Literature Cited

Hwang, M., M. S. report, Department of Chemical Engineering, California Institute of Technology, June 1968.
Katz, S., *J. Elect. Control* 16, 189 (1964).
Koppel, L. B., Ind. Eng. Chem. Fundam. 5, 403 (1966a).
Koppel, L. B., Ind. Eng. Chem. Fundam. 5, 413 (1966b).
Koppel, L. B., Ind. Eng. Chem. Fundam. 6, 209 (1967).
Koppel, L. B., Shih, Y. P., Coughanowr, D. R., Ind. Eng. Chem. Fundam. 7, 286 (1968).
Ogunye, A. F., Ray, W. H., *A.I.Ch.E. J.*, in press, 1970.
Seinfeld, J. H., Lapidus, L., *Chem. Eng. Sci.* 23 (12), 1461 (1968a).
Seinfeld, J. H., Lapidus, L., *Chem. Eng. Sci.* 23 (12), 1485 (1968b).

Received for review February 28, 1969
Accepted July 21, 1970

Work supported in part by National Science Foundation Grant GK-3342.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

P-III

Some Results on Estimation of Parameters in Ordinary Differential Equations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

SOME RESULTS ON ESTIMATION OF PARAMETERS IN ORDINARY DIFFERENTIAL EQUATIONS

Abstract

A new computational algorithm for the estimation of parameters in ordinary differential equations is presented. The algorithm suggested does not require either the particular solution of the linearized system equation for discrete measurements or the solution of the adjoint equation for continuous observations. Through consideration of the properties of common methods such as quasilinearization at a local minimum of the objective function, manipulation of the weighting matrix combined with the proposed scheme is suggested for convergence to the global minimum. An additional modification of the scheme is presented to remove ill-posedness. The new algorithm is demonstrated on a simple example which exhibits a local minimum of the objective function.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. INTRODUCTION

The estimation of parameters in a mathematical model from actual observations is an important problem in process analysis and control. We consider the case in which the model consists of a set of ordinary differential equations (O.D.E.'s). The estimation of parameters in O.D.E.'s has received much attention[1-3]. Various numerical techniques, such as quasilinearization[3-5] and steepest descent[6], have been suggested and demonstrated in the literature for this problem.

All of the techniques are plagued with two common difficulties. First, convergence to the global minimum of the objective function is never guaranteed if local minima exist. Second, if the system is ill-posed, that is, when large changes in parameter values cause only small changes in the objective function, numerical convergence is extremely unstable. In addition, when the observation process is continuous in time, we have to solve a two point boundary value problem as the necessary condition for minimization of the objective function. The numerical solution of the two point boundary value problem is often unstable and sensitive to the initial guess of the unspecified boundary condition of the adjoint variables.

The objective of the present study is to present a new computational algorithm for the estimation of parameters in O.D.E.'s which: (1) does not require the solution of the adjoint equations when the observations are continuous in time, as a modification of quasilinearization, (2) converges to the global minimum of the objective

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

function, and (3) removes the ill-posedness of a problem. The only requirement that must be satisfied for the present study is that for the given set of error-free measurements there exists a unique set of parameters, i.e., observability with respect to the parameter vector.

2. PROBLEM STATEMENT

We consider a dynamical system described by the O.D.E.

$$\dot{x}(t) = f(t,x,p) \tag{1}$$

$$x(0) = x_o \tag{2}$$

where x is an n-vector and p is an ℓ-vector of constant parameters. The observations of the system are related to the state by

$$y(t) = h(t,x) + (\text{errors}) , \qquad 0 \leq t \leq T \tag{3}$$

where y is an m-vector of observations. The problem is to find the value of p which minimizes the least square objective function. If the observations are taken continuously in time t, the objective function is

$$I(p) = \int_0^T \| y(t) - h(t,x(t;p)) \|^2_{Q(t)} \, dt \tag{4}$$

where the norm,

$$\| y(t) - h(t,x(t;p)) \|^2_{Q(t)} = [y(t) - h(t,x(t;p))]^T$$

$$Q(t)[y(t) - h(t,x(t;p))] \tag{5}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and $x(t;p)$ denotes the solution of equations (1) and (2). If the observations are made only at discrete times, $t_1, t_2, \cdots, t_s$, then I is given by

$$I(p) = \sum_{i=1}^{s} \| y(t_i) - h(t_i, x(t_i;p)) \|^2_{Q(t_i)} \tag{6}$$

We assume that f and $h \in C^1$, and f and h have continuous first order partial derivatives with regard to their arguments. $Q(t)$ is a symmetric, positive-definite weighting matrix. Equation (4) can be reduced to equation (6) by the special choice of $Q(t)$ as $Q(t)\ s(t-t_i)$, so that in the following part the system with discrete observations is considered as a special case of the continuous measurement system. Finally, we assume that the error free system of equations (1) and (2) is observable with regard to $\bar{p}$, i.e., there exists a unique value of the parameter p^* at which $I(p^*) = 0$.

3. A NEW ALGORITHM

Assume we have an initial guess $p(0)$ which generates a trajectory of equations (1) - (3) denoted by $x^{(o)}$. To describe the trajectory of equations (1) - (3) generated by $p^{(o)} + \delta p^{(o)}$ we can linearize equation (1) about $x^{(o)}$ if $\delta p^{(o)}$ is chosen such that for some $\eta > 0$,

$$\| x^{(1)}(t) - x^{(o)}(t) \| < \eta \tag{7}$$

Then we can write the following unique perturbation equations[7,8]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\delta\dot{x}^{(o)} = f_x^{(o)}\delta x^{(o)} + f_p^{(o)}\delta p^{(o)} \tag{8}$$

$$\delta x^{(o)}(0) = 0 \tag{9}$$

$$\delta y^{(o)} = h_x^{(o)}\delta x^{(o)} \tag{10}$$

where $f_x^{(o)}$ denotes $(\partial f/\partial x)_{x^{(o)},p^{(o)}}$.

Substituting the solution of equations (8) and (9) into equation (10), we obtain

$$\delta y^{(o)}(t) = \theta(t,x^{(o)})\delta p^{(o)} \tag{11}$$

where

$$\theta(t,x^{(o)}) = h_x^{(o)}\, D(t,x^{(o)}) \tag{12}$$

$$D(t,x^{(o)}) = \int_0^t \phi^{(o)}(t,\tau)\, f_p^{(o)}(\tau)\,d\tau \tag{13}$$

and the fundamental matrix satisfies

$$\frac{\partial\phi^{(o)}(t,\tau)}{t} = f_x^{(o)}\phi^{(o)}(t,\tau) \tag{14}$$

$$\phi^{(o)}(t,t) = I \text{ (identity matrix)} \tag{15}$$

It can be easily shown that $D(t,x^{(o)}) = (\partial x/\partial p)^{(o)}$, the matrix of sensitivity coefficients, which satisfies

$$\dot{D}^{(o)} = f_x^{(o)}D^{(o)} + f_p^{(o)} \tag{16}$$

$$D^{(o)}(0) = 0 \tag{17}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

If $\delta y^{(o)}(t) = y^{(1)} - y^{(o)}$ does not contain measurement errors, and $\theta(t,x^{(o)})$ is not singular for all admissible $p^{(o)}$ and all t, then from equation (11) we can determine $\delta p^{(o)}$ uniquely with one measurement $\delta y^{(o)}(t)$, $t \in [0,T]$. Furthermore, we can obtain p^* uniquely by using equation (11) only. This point will be considered again. In practice, observations are noisy, $\theta(t,x^{(o)})$ is singular if $\ell \neq m$, and it will be necessary to consider all given measurements in order to generate $\delta p^{(o)}$. Then $\delta p^{(o)}$ can be evaluated by using the pseudo-inverse matrix,

$$\delta p^{(o)} = K(T,p^{(o)})^{-1} \int_0^T \theta(t,x^{(o)})^T Q(t)\, \delta y^{(o)} dt \qquad (18)$$

where

$$K(T,p^{(o)}) = \int_0^T \theta(t,x^{(o)})\, dt \qquad (19)$$

If $K(T,p^{(o)})$ is nonsingular, there exists a unique perturbation $\delta p^{(o)}$ corresponding to the perturbation $\delta y^{(o)}(t)$, $0 \leq t \leq T$. This implies that the system of equations (1)-(3) is locally observable with regard to $p^{(o)}$ at $(x^{(o)},p^{(o)})$ [8]. In addition, equation (18) can be applied for the case where $\delta y^{(o)}$ contains measurement errors, i.e., equation (10) is replaced by

$$\delta y^{(o)}(t) = h_x^{(o)}\, \delta x^{(o)} + (\text{errors}) \qquad (20)$$

because $\delta p^{(o)}$ given by equation (18) minimizes

$$\int_0^T \| \delta y^{(o)} - \theta(t,x^{(o)})\, \delta p^{(o)} \|^2_{Q(t)}\, dt$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Equation (18) can be used to update $p^{(o)}$. However, in order to use equation (18) we must give some special attention to $\delta y^{(o)}$. Theoretically $\delta y^{(o)}$ is given by $y^{(1)} - y^{(o)}$, but $y^{(1)}$ is unknown because $p^{(1)}$ is unknown. Thus, to use equation (18) we need to assume $\delta y^{(i)}$. We therefore replace $\delta y^{(i)}$ by

$$\delta y^{(i)} = \varepsilon'[y(t) - h(t,x^{(i)})] \tag{21}$$

for some $\varepsilon' > 0$ such that equation (7) is satisfied. Combining equations (18) and (21) we obtain for the general iteration i

$$\delta p^{(i)} = \varepsilon' K(T,p^{(i)})^{-1} \int_0^T \theta(t,x^{(i)})^T Q(t)(y(t) - h(t,x^{(i)}))dt \tag{22}$$

Equation (22) provides the basic algorithm to update $p^{(i)}$.

To explore the meaning and numerical convergence of equation (22) we consider another derivation of equation (22). At any step in the iteration, equation (4) can be written as

$$I(p) = \int_0^T \| y - h^{(i)} - (h - h^{(i)}) \|^2_{Q(t)} \, dt \tag{23}$$

Linearizing h about $h^{(i)}$ we obtain

$$h - h^{(i)} = h_x^{(i)} D^{(i)} \delta p^{(i)} + O(\delta p^{(i)2}) \tag{24}$$

Substituting equation (24) into equation (23) and neglecting second order terms, we obtain

$$I(\delta p^{(i)}) = \int_0^T \| y - h^{(i)} - h_x^{(i)} D^{(i)} \, \delta p^{(i)} \|^2_{Q(t)} dt \tag{25}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Using the stationarity condition

$$\frac{\partial I}{\partial(\delta p^{(i)})} = 0 \tag{26}$$

and solving for $\delta p^{(i)}$ we obtain equation (22) with $\varepsilon' = 1$.

In fact, we can show that the quasilinearization scheme commonly applied will yield equation (22) with proper changes in the boundary conditions for the homogeneous and the particular solutions of the linearized state equation. This is demonstrated as follows. If we adjoin to equation (1) the ℓ relations $\dot{p} = 0$ and define the $(n+\ell)$-vector $z = (x^T, p^T)^T$, then $z(t)$ satisfies

$$\dot{z}(t) = g(t,z) \tag{27}$$

$$z(o) = (x_o^T, p^T)^T \tag{28}$$

Linearizing equation (27) about the i^{th} iterate of z

$$\dot{z}^{(i+1)}(t) = g_z^{(i)} z^{(i+1)} + g^{(i)} - g_z^{(i)} z^{(i)} \tag{29}$$

The solution of this equation is

$$z^{(i+1)}(t) = \phi^{(i)} \alpha^{(i+1)} + \psi^{(i)}(t) \tag{30}$$

where the $(n+\ell) \times (n+\ell)$ matrix $\phi^{(i)}$ satisfies

$$\dot{\phi}^{(i)}(t) = g_z^{(i)} \phi^{(i)}(t) \tag{31}$$

$$\phi^{(i)}(0) = I \quad \text{(identity matrix)} \tag{32}$$

and the $(n+\ell)$-vector $\psi^{(i)}$ satisfies

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\dot{\psi}^{(i)}(t) = g_z^{(i)}\psi^{(i)} + g^{(i)} - g_z^{(i)}\, z^{(i)} \tag{33}$$

$$\psi^{(i)}(o) = 0 \tag{34}$$

The objective function is

$$I = \int_0^T \| y - h(\phi^{(i)}\alpha^{(i+1)} + \psi^{(i)}) \|^2_{Q(t)}\, dt \tag{35}$$

and is to be minimized with regard to $\alpha_k^{(i+1)}$, $k = n+1, \cdots, n+\ell$. If the initial conditions of equations (31) and (33) are taken as

$$\phi_{kk}^{(i)}(0) = \begin{cases} 0 & k \leq n \\ 1 & k \geq n+1 \end{cases} \tag{36}$$

$$\psi_k^{(i)}(0) = \begin{cases} x_{ko} & k \leq n \\ 0 & k \geq n+1 \end{cases} \tag{37}$$

Then equation (30) can be rewritten as

$$x^{(i+1)} = D^{(i)}\, p^{(i+1)} + \psi^{(i)} \tag{38}$$

and the solution of equation (33) can be shown to be

$$\psi^{(i)} = x^{(i)} - D^{(i)}\, p^{(i)} \tag{39}$$

Substituting equations (38) and (39) into equation (35), expanding h about $h^{(i)}$, and minimizing with regard to $p^{(i+1)}$, we obtain equation (22) for $\varepsilon' = 1$. The important point is that the present algorithm is computationally much faster than quasilinearization with the same accuracy because integration of the particular solution governed by

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

equation (33) is avoided for discrete measurements.

4. NUMERICAL CONVERGENCE OF THE PROPOSED SCHEME

Let us first examine the properties of the proposed scheme based on equation (22) in the vicinity of a minimum of I. For this purpose we rewrite equation (25) as

$$I(\delta p^{(i)}) = \int_0^T \| y - h^{(i)} \|^2_{Q(t)} dt - 2 \int_0^T \theta(t,x^{(i)})^T Q(t)(y - h^{(i)}) dt \; \delta p^{(i)} + \| \delta p^{(i)} \|^2_{K(T,p^{(i)})} \tag{40}$$

or, equivalently,

$$I(\delta p^{(i)}) = \| \delta p^{(i)} - K(T,p^{(i)})^{-1} \int_0^T \theta(t,x^{(i)})^T Q(t)(y-h^{(i)}) dt \|^2_{K(T,p^{(i)})} + \int_0^T \| y - h^{(i)} \|^2_{Q(t)} dt - \left\| \int_0^T \theta(t,x^{(i)})^T Q(t)(y-h^{(i)}) dt \right\|^2_{K(T,p^{(i)})} \tag{41}$$

Equations (40) and (41) represent an approximation to the I surface by a quadratic function of $\delta p^{(i)}$ in the neighborhood of $p^{(i)}$. The value of I corresponding to $p^{(i)}$ is the second term on the R.H.S. of equation (41) and can be denoted $I^{(i)}$. The minimum of $I(\delta p^{(i)})$ occurs at $\delta p^{(i)}$ given by equation (22) with $\varepsilon' = 1$. Since the third term on the R.H.S. of equation (41) is positive, the value of I at the minimum of the quadratic approximation is decreased from $I^{(i)}$ by that amount. Normally, $p^{(i)}$ will not be in the neighborhood of a true minimum of I, so that equation (22) provides a small step in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the direction of decreasing I by approximating the actual I surface by the quadratic form of equation (25). If, however, the minimum of $I(\delta p^{(i)})$ occurs at an actual minimum of I, then equation (41) is exact to $0(\delta p^{(i)2})$ and the condition which holds at the minimum is

$$\int_0^T \theta(t,x^{(i)})^T Q(t)(y - h^{(i)})\, dt = 0 \qquad (42)$$

The essential difficulty with convergence to a local minimum is embodied in equation (42). Since equation (42) holds at a local minimum, $\delta p^{(i)}$ becomes zero from equation (22). Thus, in order to avoid local minima we need another relationship by which $\delta p^{(i)}$ evaluated does not become zero at local minima.

Let us return to equation (11) to develop a scheme which can only converge to the global minimum. We rewrite equation (11) in the form

$$y(t) - h(t,x(t;p^{(i)})) = \theta(t,x^{(i)})\, \delta p^{(i)} \qquad (43)$$

Since the L.H.S. of equation (43) is the difference between the given measurements and the predicted measurements with $p^{(i)}$, in the absence of errors, the L.H.S. is equal to zero for all t only when $p^{(i)} = p^*$, the true value. Let us define an ms-vector Y by choosing $y(t_j)$, $j=1,\cdots,s$ from continuous observation $y(t)$, $0 \le t \le T$, as

$$\delta Y^{(i)} = [(y(t_1) - h^{(i)}(t_1))^T, \cdots, (y(t_s) - h^{(i)}(t_s))^T]^T \qquad (44)$$

Similarly, an $(sm \times \ell)$ matrix $\Theta^{(i)}$ is given as

$$\Theta^{(i)} = [\theta^{(i)}(t_1)^T, \cdots, \theta^{(i)}(t_s)^T]^T \qquad (45)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Therefore we have

$$\delta Y^{(i)} = \Theta^{(i)} \delta p^{(i)} \tag{46}$$

We can choose, in principle, a nonsingular $(\ell \times \ell)$ matrix $\tilde{\theta}^{(i)}$ from $\Theta^{(i)}$ for which the corresponding ℓ-vector measurements $\delta\tilde{y}^{(i)}$ from $\delta Y^{(i)}$ are not identically zero. $\delta p^{(i)}$ is, then, determined from

$$\delta\tilde{y}^{(i)} = \tilde{\theta}^{(i)} \delta p^{(i)} \tag{47}$$

Computing $\delta p^{(i)}$ in this way, we can avoid convergence to local minima of I. However, $\delta p^{(i)}$ generated by the above method may not be in the direction of the global minimum of I. This can be illustrated easily. Let us consider a scalar p and $y(t) = x_1(t)$. Then from equation (47), $\delta p^{(i)}$ is given by (with $\tilde{y} = y(t^*)$, $t^* \in [0,T]$)

$$\delta p^{(i)} = \frac{\delta y^{(i)}(t^*)}{D_{11}(t^*,x^{(i)})} \tag{48}$$

Suppose $\delta y(t) \geq 0$ for $p_* < p^{(i)} < p^*$ and all time, but $D_{11}(t,x^{(i)})$ changes its sign in some interval of time and in the range of $p^{(i)}$, then $\delta p^{(i)}$ may have a wrong direction, depending on the choice of t^* and $p^{(o)}$. Furthermore, $D_{11}(t^*,x^{(i)})$ must be zero at some value of $p^{(i)}$ in the range. For true convergence with $\tilde{y} = y(t^*)$, therefore, it is necessary that $D_{11}(t^*,x^{(i)})$ does not become zero for all admissible $p^{(i)}$. In general, it is required that $\tilde{\theta}$ is not singular for all admissible values of $p^{(i)}$ which is the observability condition of the system with $\tilde{y}$ for p. In practice, $y(t)$ contains measurement errors, and the observability condition for the given system with the above discrete observation

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

raises another question in addition to the difficulty in choosing the measurement locations in the time axis. On the basis of the above discussions, we can propose a modification of the computational scheme as a trial and error procedure for the initial period of iteration. Instead of discretizing the continuous observation, we can manipulate the weighting matrix $Q(t)$, knowing the fact that the sign of the sensitivity matrix $D(t,x^{(i)})$ is usually fixed near the given boundary condition. Therefore, we can start the computation by using equation (22) with $Q(t) = I$ (identity matrix). When $\delta p^{(i)}$ becomes zero or less than a preset value, then we can change the weighting matrix $Q(t)$ such that a special weight is given to some time interval near the boundary where the state variables are specified originally. Hence, we can avoid convergence to local minima without changing the scheme. When $I^{(i)}/I^{(o)} < \lambda$, another preset value, then $Q(t)$ can be so changed again as to give even weight to all data for the faster convergence. This idea is shown in the example. The algorithm suggested can be summarized as follows:

1) Select $\varepsilon', \varepsilon_*, \lambda$ and $Q(t)$.
2) Make an initial guess $p^{(o)}$.
3) Solve the system equations (1)-(3) with $p^{(o)}$ to generate $x^{(o)}$. Evaluate $\theta^{(o)}$, $K(T,p^{(o)})$, and $I^{(o)}$.
4) Compute $\delta p^{(o)}$ by equation (22), then determine $p^{(1)} = p^{(o)} + \delta p^{(o)}$.
5) Repeat step 3, replacing $\delta p^{(o)}$ by $\delta p^{(1)}$, etc.
6) When $\delta p^{(i)}$ and $I^{(i)}/I^{(o)} > \lambda$ change $Q(t)$ such that a special weight is given to the data near $t = 0$. Then compute $\delta p^{(i)}$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

7) When $I^{(i)}/I^{(o)} < \lambda$, change $Q(t)$ again so that even weight is given to all data. Or, without changing $Q(t)$, proceed until $I^{(i)}$ is less than the desired final convergence accuracy.

5. TREATMENT OF ILL-POSED PARAMETER ESTIMATION PROBLEMS

Up to this point we have discussed the convergence problems when local minima exist. In this section we consider briefly the case when the problem is ill-posed, that is, when large changes in the parameter values cause only small changes in the objective function.

To remove any ill-posedness we will employ the results of Klinger[10] and Franklin[11]. We quote the following theorem without proof.

Theorem (10). If A is normal, i.e., $AA^* = A^*A$, where A^* is the conjugate transpose of the matrix A , then for all $\sigma > 0$

$$[A + \sigma(A^*)^{-1}]x = b \tag{49}$$

is better conditioned than $Ax = b$ in terms of the P-condition number, unless $P(A) = 1$, where

$$P(A) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|} \tag{50}$$

and λ_i are the eigenvalues of A .

Applying this theorem to equation (22) we obtain

$$\delta p^{(i)} = (K^T K + \sigma I)^{-1} K^T \int_0^T \theta(t,x^{(i)})^T Q(t) \delta y^{(i)}(t) \, dt \tag{51}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

When the problem is ill-posed $K(T,p^{(i)})$ becomes nearly singular (note that K cannot be singular because of the observability assumption). The use of equation (51) instead of equation (22) will remove the inaccuracy in computing $\delta p^{(i)}$ as a result of K being almost singular. If we choose a large value for σ, compared to K^TK, the magnitude of $\delta p^{(i)}$ will be decreased. A more complete treatment of ill-posed linear problems is given by Franklin[11] and it can be shown that the theorem above, due to Klinger, is a special case of the more general theory developed by Franklin.

6. EXAMPLE

We wish to estimate p for the system (true value is $\pi^2 = 9.8696044$)

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -px_1$$

$$x_1(0) = 0 \; ; \; x_2(0) = \pi$$

$$y(t) = \sin \pi t \qquad 0 \leq t \leq 1 \qquad (53)$$

The curve of $I = \int_0^T (y(t) - x_1(t))^2 dt$ vs. p is shown in Fig. 1. There are a local minimum at about $p = 117$ and other local minima not shown at larger values of p.

First, using only equation (22) which, as we have noted, is equivalent to quasilinearization, the iterations converged to the local minimum at $p = 116.1$ from initial guesses of 58 and 100. This confirms the inability of the quasilinearization-type algorithm to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

avoid a local minimum.

Next, the globally convergent scheme suggested is applied with $p^{(o)} = 220$, $\varepsilon' = 0.2$ for $i \leq 7$, and $\varepsilon' = 1$ for $i > 7$ and $Q(t) = \delta(t-t^*)$. Thus, the numerical scheme is given by equation (48). Consequently, its convergence becomes slow. With $t^* = 0.1$, $D_{11}(0.1,x^{(8)}) < 0$ was maintained for all iterations. But with $t^* = 0.5$, $D_{11}(0.5,x^{(i)})$ changed its sign during iteration. The resulting state variables oscillated widely, even though the correct convergence was obtained eventually. The results of the iteration are shown in Tables 1 and 2. For noisy observations simulated as

$$y(t) = (1 + 0.2 \text{ Gauss } (0,1))\sin \pi t \tag{54}$$

where Gauss (0,1) indicates the Gaussian distribution with zero mean and the standard deviation of 1, the suggested scheme is applied with $p^{(o)} = 220$ and

$$Q(t) = \begin{cases} 10 & t \leq 0.3 \\ 0.1 & t > 0.3 \end{cases} \tag{55}$$

The results obtained are shown in Table 3. Because of high noise level and the fixed $Q(t)$, the final numerical value has a relative error of about 1% from the true value π^2.

7. SUMMARY

We have considered three aspects of the estimation of parameters in ordinary differential equations. First, we presented a computational method, embodied in equation (22). Second, using the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

properties of equation (22) in the region of a minimum of I, we suggested a new technique for computing an iteration $\delta p^{(i)}$ which would avoid a local minimum of I. Third, we employed a result of Klinger and Franklin to remove ill-posedness in a particular problem. The present study can be directly extended to more general cases where the initial conditions and parameters are unknown. Also it can be extended to nonlinear distributed parameter systems. The algorithm suggested was illustrated on an example exhibiting a local minimum of the objective function.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

8. NOTATION

A = arbitrary matrix

D = sensitivity matrix

f = n-dimensional vector function

g = $(n+\ell)$-dimensional vector function

h = m-dimensional vector function

I = identity matrix, or performance index

$I(p)$ = performance index

$K(T,p)$ = observability matrix defined by equation (19)

p = ℓ-dimensional constant parameter

$p(A)$ = condition-number defined by equation (50)

Q = weighting matrix

s = constant

t,T = time variables

x = n-dimensional state vector

y = m-dimensional observation vector

Y = ms-dimensional vector defined by equstion (44)

z = $(n+\ell)$-dimensional vector

Greek Letters

α = $(n+\ell)$-dimensional constant

ε' = constant

η = constant

θ = $(m \times \ell)$-matrix defined by equation (12)

Θ = $(ms \times \ell)$-matrix defined by equation (45)

λ = eigenvalues

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

σ = constant

$\phi(t,\tau)$ = transition matrix

ψ = particular solution defined by equation (33) or (39)

Superscripts

i = at the i^{th} iteration

$*$ = particular value, or conjugate transpose of a matrix, or upper bound

Subscripts

o = initial

$*$ = low bound

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

REFERENCES

1. Seinfeld, J. H.,"Nonlinear Estimation Theory", Industrial and Engineering Chemistry, 62, 32-42 (1970).

2. Nieman, R. E., Fisher, D. G. and Seborg, D. E., "A Review of Process Identification and Parameter Estimation Techniques", International Journal of Control, in press.

3. Lee, E. S., Quasilinearization and Invariant Imbedding, Academic Press, New York, 1968.

4. Leondes, C. T. and Paine, G., "Extensions in Quasilinearization Techniques for Optimal Control", Journal of Optimization Theory and Applications 2, 316-330 (1968).

5. Leondes, C. T. and Paine, G.,"Computational Results in Extensions in Quasilinearization for Optimal Control", Journal of Optimization Theory and Applications 2, 395-410 (1968).

6. Hestenes, M. R., "Multiplier and Gradient Methods", Journal of Optimization Theory and Applications 4, 303-320 (1969).

7. Hestenes, M. R. and Guinn, T., "An Embedding Theorem for Differential Equations", Journal of Optimization Theory and Applications 2, 87-101 (1968).

8. Hwang, M. and Seinfeld, J. H., "Observability of Nonlinear Systems", Journal of Optimization Theory and Applications, submitted for publication.

9. Franklin, J. N., Matrix Theory, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

10. Klinger, A., "Approximate Pseudoinverse Solutions to Ill-Conditioned Linear Systems", Journal of Optimization Theory and Applications 2, 117-124 (1968).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

11. Franklin, J. N., "Well-Posed Stochastic Extensions of Ill-Posed Linear Problems", Willis H. Booth Computing Center, Technical Report No. 135, California Institute of Technology, 1969.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

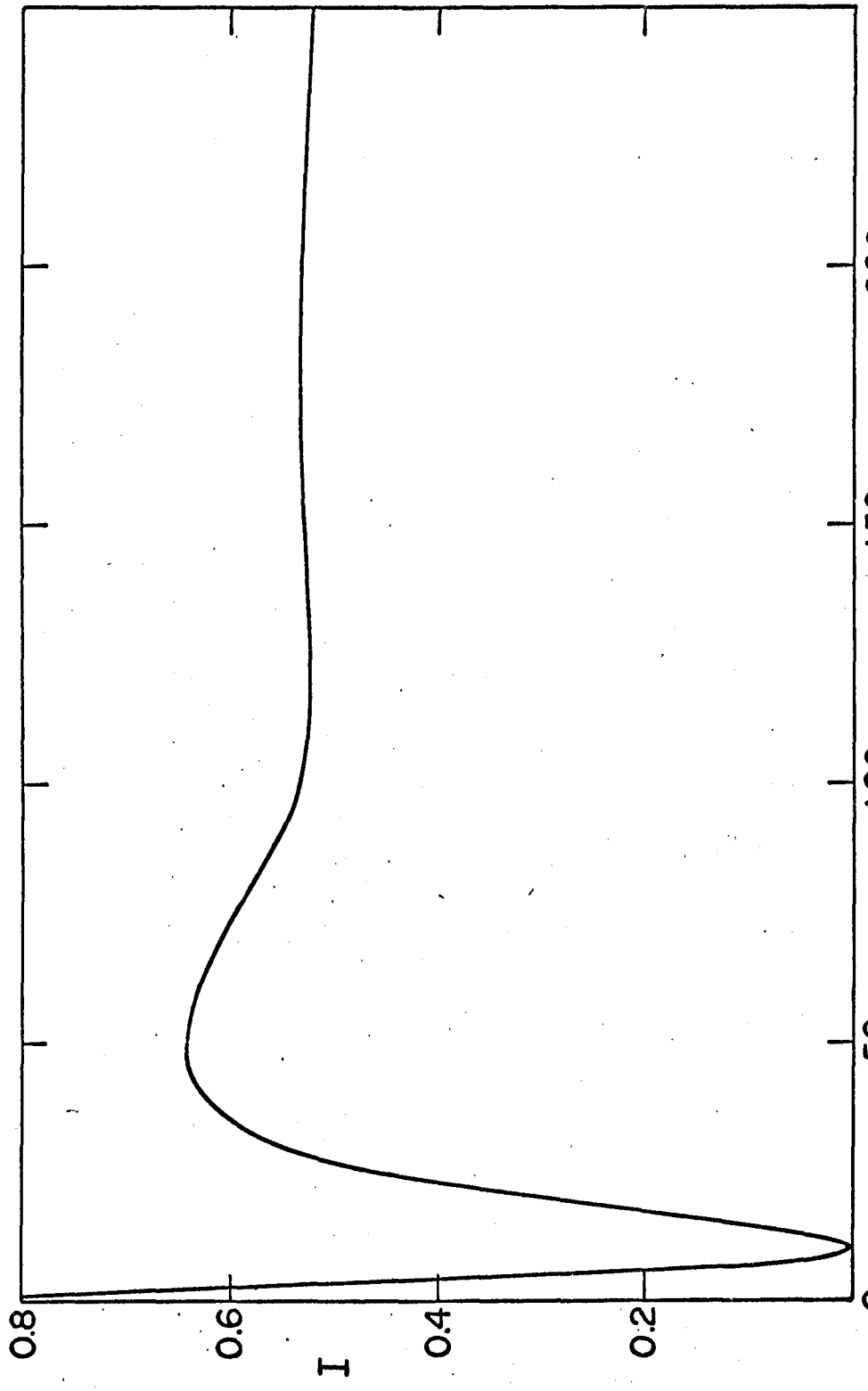

FIG. 1

The objective function I as a function of p for the example. No observation noise.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 1. Progress of Iterations with $t^* = 0.1$

Iteration Number	$p^{(i)}$	$y^{(i)}(0.1)$	$D_{11}(0.1, x^{(i)})$
1	220	9.8029×10^{-2}	-3.1323×10^{-4}
2	157.40900	7.1033×10^{-2}	-3.3053×10^{4}
3	114.42744	5.1439×10^{-2}	-3.4279×10^{-4}
4	84.415085	3.7231×10^{-2}	-3.5153×10^{-4}
5	63.233170	2.6936×10^{-2}	-3.5780×10^{-4}
6	48.176544	1.9482×10^{-2}	-3.6231×10^{-4}
7	37.421631	1.4089×10^{-2}	-3.6555×10^{-4}
8	29.713409	1.0186×10^{-2}	-3.6788×10^{-4}
9	4.7934265	-2.6386×10^{-3}	-3.7551×10^{-4}
10	11.181425	6.7967×10^{-4}	-3.7355×10^{-4}
11	9.5135345	-1.8471×10^{-4}	-3.7406×10^{-4}
12	10.007348	7.1347×10^{-5}	-3.7391×10^{-4}
13	9.8165331	-2.7478×10^{-5}	-3.7396×10^{-4}
14	9.8900099	1.0669×10^{-5}	-3.7394×10^{-4}
15	9.8614779	-4.4107×10^{-6}	-3.7395×10^{-4}
16	9.8732719	1.7881×10^{-6}	-3.7395×10^{-4}
17	9.8684893	-5.9605×10^{-7}	-3.7395×10^{-4}
18	9.8700829	1.7881×10^{-7}	-3.7395×10^{-4}
19	9.8696041	0	-3.7395×10^{-4}

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2. Progress of Iterations with $t^* = 0.5$

Iteration Number	$p^{(i)}$	$\delta y^{(i)}(0.5)$	$D_{11}(0.5, x^{(i)})$
1	220	8.0823×10^{-1}	1.5325×10^{-3}
2	325.47607	9.3133×10^{-1}	-2.1042×10^{-3}
3	236.95714	7.9849×10^{-1}	5.8717×10^{-4}
4	508.93408	1.1336	2.1553×10^{-4}
5	1560.8730	9.4235×10^{-1}	4.0292×10^{-4}
6	2028.6382	1.0273	-3.3674×10^{-4}
7.	1418.5103	1.0047	4.7748×10^{-4}
8	1839.3652	9.6004×10^{-1}	-2.0574×10^{-4}
9	-2360.1895	-1.0562×10^{9}	-3.2463×10^{6}
10	-2064.3965	-2.3899×10^{8}	-8.0455×10^{5}
11	-1792.0991	-5.5254×10^{7}	-2.0441×10^{5}
12	-1521.7881	-1.1537×10^{7}	-4.7491×10^{4}
13	-1278.8391	-2.5043×10^{6}	-1.1517×10^{4}
14	-1061.3914	-5.6417×10^{5}	-2.9134×10^{3}
15	-867.74243	-1.3169×10^{5}	-7.6834×10^{2}
16	-696.33740	-3.1799×10^{4}	-2.1119×10^{2}
17	-545.76123	-7.9265×10^{3}	-6.0497×10^{1}
18	-414.73682	-2.0348×10^{3}	-1.8070×10^{1}
19	-302.13477	-5.3619×10^{2}	-5.6368
20	-207.01096	-1.4430×10^{2}	-1.8429
21	-128.70990	-3.9254×10^{1}	-6.3723×10^{-1}
22	-67.109558	-1.0521×10^{1}	-2.3865×10^{-1}
23	-23.025330	-2.5760	-1.0310×10^{-1}
24	1.9609528	-4.4556×10^{-1}	-5.8723×10^{-2}
25	9.5483904	-1.6343×10^{-2}	-4.8562×10^{-2}
26	9.8849701	7.7552×10^{-4}	$-4.8142 \times 10-2$
27	9.8688107	-3.9518×10^{-5}	-4.8162×10^{-2}
28	9.8696308	1.6093×10^{-6}	-4.8161×10^{-2}
29	9.8695974	-4.1723×10^{-7}	-4.8161×10^{-2}

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3. Progress of Iterations with Noise Observations

Number i of Iteration	$p^{(i)}$
1	220
2	111.52681
3	74.752411
4	55.308624
5	44.019943
6	36.247849
7	16.143158
8	8.4432821
9	9.6022644
10	9.7483568
11	9.7580004
12	9.7585392
13	9.7585135

N.B.

(1) Relative error at 13^{th} iteration $= 7.53 \times 10^{-7}$

(2)
$$\varepsilon' = \begin{cases} 0.3 & , \ i \leq 5 \\ 1 - \frac{1}{(i+2)} & , \ 5 < i \leq 10 \\ 1 & , \ i > 10 \end{cases}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.