14:45:55 OCA PAD INITIATION - PROJECT HEADER INFORMATION 08/19/87

Project # E-21-676 MOD # REV # 0
Contract # AFOSR-87-0308 OCA file # Status A
Contract entity GTRC Prime contract #
PDPI HADDAD A H (DR.
SSN - - Unit EE Phone ( ) -
Project unit EE Unit code 02.010.118
Sponsor/Division AIR FORCE / BOLLING AFB, DC
Sponsor#/division # 104 / 001
Type of document GRANT
Award period: from 87 / 08 / 01 to 88 / 07 / 30 (perf) 88 / 08 / 30 (rpts)

Sponsor amount	New this change	Total to date
Contract value	$ 100997	100997
Funded	$ 100997	100997

Cost sharing # Cost sharing $
Does subcontracting plan apply? (Y/N) N
Title -
ESTIMATION AND CONTROL OF NONLINEAR AND HYBRID SYSTEMSWITH APPLICATIONS......

CTR project # R6376-0A0 CTR cost sharing #

Are there existing subprojects? (Y/N) N
Is this a subproject? (Y/N) N Main project #
Continuation of project # Type of research RES

Coproject director name

SSN - - Unit

Coproject director name

SSN - - Unit

PROJECT ADMINISTRATION DATA

Administrative data OCA contact BRIAN J. LINDBERG PAD CO BJL 894-4820

Sponsor technical contact	Sponsor issuing office
JAMES M CROWLEY, MAJOR, USAF/NM	HUGH M MCELROY/PKZ
(202) 767 - 5025	(202) 767 - 4952
AFOSR/NM	AFORS/PKZ
BUILDING 410	BUILDING 410
BOLLING AFB DC 20332-6448	BOLLING AFB DC 20332-6448

Security class (U,C,S,TS) U ONR resident rep. is ACO (Y/N) N
Defense priority rating NA
NA supplemental sheet
Equipment title vests with Sponsor GIT X Comment follows -
TITLE TO EQUIP ACQUIRED AT COST OF < $5000 SHALL VEST IN GIT UPON ACQUISITION

Admin comments -
INITIATION OF AFOSR PROJECT E-21-676/HADDAD.

12/88

2-N-7
5R2/9

GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION

NOTICE OF PROJECT CLOSEOUT

Date 9/20/89

Project No. E-21-676 Center No. R6376-0A0

Project Director A. H. Haddad School/Lab EE

Sponsor Air Force

Contract/Grant No. AFOSR-87-0308 GTRC XX GIT ____

Prime Contract No. N/A

Title Estimation and Control of Nonlinear and Hybrid Systems with Applications to Air-to-Air Guidance

Effective Completion Date 3/31/89 (Performance) 5/30/89 (Reports)

Closeout Actions Required:

____ None
____ Final Invoice or Copy of Last Invoice - Already submitted.
X Final Report of Inventions and/or Subcontracts -Patent questionnaire sent to PI.
X Government Property Inventory & Related Certificate
X Classified Material Certificate
____ Release and Assignment- Already submitted.
____ Other ____

Includes Subproject No(s). ____

Subproject Under Main Project No. ____

Continues Project No. ____ Continued by Project No. ____

Distribution:

X Project Director
X Administrative Network
X Accounting
X Procurement/GTRI Supply Services
X Research Property Management
____ Research Security Services
X Reports Coordinator (OCA)
X GTRC
X Project File
2 Contract Support Division (OCA)
____ Other ____

E-21-676

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30332

TELEPHONE: (404) 894-3930

February 5, 1988

Major James M. Crowley
AFOSR/NM
Bldg 410
Bolling AFB, DC 20332-6448

Dear Major Crowley:

Enclosed is an advance copy of the progress report for the research project No. AFOSR-87-0308 covering the period 1 August 1987 to 31 January 1988. The formal report with the attachments will be forwarded by the Georgia Tech contract office. We'll be glad to provide any additional information you may need in the future.

Thank you very much for your continued support of our research.

Sincerely yours,

A. H. Haddad
Professor

cc: Mr. Johnny Evers
AFATL/DLMA
Eglin AFB, FL 32542

Progress Report

Estimation and Control of Nonlinear and Hybrid Systems with Applications to Air-to-Air Guidance

by

A. H. Haddad
E. I. Verriest

submitted to

Major James M. Crowley
AFOSR/NM
U. S. Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332-6448

Grant No. AFOSR-87-0308

August 1, 1987

to

January 31, 1988

Estimation and Control of Nonlinear and Hybrid Systems with Applications to Air-to-Air Guidance

Progress Report

During the past six months the research continued in the directions established and developed under the earlier project supported by the US Air Force Armament Laboratories at Eglin Air Force Base. Four major areas of research were undertaken. While we continued the effort to develop models for an air-to-air encounter that will be suitable for simulating the filtering and control schemes developed for nonlinear and hybrid models, the primary focus during the period was the derivation of the theoretical basis for these methods. The four areas are involved with the nonlinear filtering approximation and implementation problem, the control and stabilization of hybrid stochastic system models, the approximate analysis and implementation of controllers for quantized and piecewise linear systems subject to fast and slow dynamics, and realization theory for minimum sensitivity sensor and actuator placement for the guidance of uncertain systems. These topics are directly applicable to the basic approximation involved in the switched Markov approximation used for the nonlinear model in the air-to-air scenario.

The nonlinear filtering scheme developed earlier (see Reference 1) is continued to be evaluated for a more complex assumptions such as a higher dimensionality and longer memory for the number of multiple models involved. In addition an alternative approach to the modeling of target maneuvers via self-exciting Poisson inputs into the system model has yielded a new filtering scheme that is still under evaluation (see Reference 2). A preliminary simulation indicates the potential benefit of the scheme, and a more rigorous bound on the resulting error is being derived. Future studies will focus on relaxing the assumptions on the model of the Poisson input and in combining the resulting scheme with the switched Markov model filter for the slow switching case since it is based on a combined detection and estimation approach.

The hybrid system models considered in this research assume a model that is switching among several linear models. During the early phase we concentrated on the properties of the deterministic hybrid models where the transitions are not random. We derived controllability, observability, and stability properties for such systems (see Reference 3). These properties were then used to design control algorithms for stabilizing such systems and for feedback control of such systems. The second phase was concerned with the stochastic case which the appropriate model for the switched Markov approximation used for our original nonlinear system. In this case an average system has been defined and is used to derive stochastic stability and controllability properties for the model(see Reference 4) and obtain conditions for the design of optimal control algorithms. Our final objective is to relate the results to the guidance and control of the original problem when combined with the appropriate nonlinear filter.

The third area of research was involved with the two-time scale design and implementation of nonlinear control systems with fast and slow dynamics. Such models may occur in the slow and fast switching in the switched Markov model. In considering nonlinearities we first concentrated on deterministic cases that involve both quantized inputs and piecewise linear nonlinearities. The approximate reduced order two-time scales design was shown to be simpler and requires one third the computing time as the exact nonlinear model without affecting the desired behavior of the controlled system (see References 5 and

6). In order to apply the results to the uncertain case the emphasis in the next phase has been on such models with stochastic inputs (see Reference 7).

Other efforts concentrated on modelling (the Willems' formalism), parameterization, and realization schemes (see References 8 through 14). This included an application of our minimal sensitivity design problem to the optimal sensor/actuator placement or design in multivariable systems. A case study for the linearization scheme of the Fokker-Planck equation was studied but the results concerning the validity of the approach were rather discouraging. Two major subproblems in the sensor placement area were considered. The actuator related problems will not be discussed because of an obvious "duality". The first part of the sensor placement problem is deterministic in nature and deals with whether or not a physical system with some specific arrangement of sensors is simply observable. That is, if all the parameters of the system are known exactly and the measurement is perfect, can we determine the state of the system at some point in the past given enough data? Although this problem has been addressed, the picture is not yet complete. A related question is whether some observable sensor configurations have certain optimality properties not universal to the whole class of observable configurations. Our work to date seems to indicate that such is the case, especially in view of the next part of the sensor location problem: the introduction of plant and measurement uncertainty. In the stochastic framework one talks about optimal sensor location in the sense of minimizing some type of error in the state estimation. It seems that the results in the deterministic problems shed light on the stochastic case. Another interesting aspect of the sensor problem that was pursued with success is its relation to design sensitivity and robustness. The optimal sensor problem can be cast in to a geometric framework that was researched earlier (see Reference 15).

The future directions in all four areas are as outlined above with an emphasis on relating the primary approaches to the solution of the general problem of guidance and control of nonlinear stochastic systems with applications to scenarios that reflect an air-to-air encounter.

REFERENCES

[1] A. H. Haddad, E. I. Verriest, and P. D. West, "Approximate Nonlinear Filtering for Piecewise Linear Systems," NATO/AGARD Guidance and Control Panel's 44th Symposium, Athens, Greece, 5-8 May 1987.

[2] M. A. Ingram and A. H. Haddad, "Optimal and Suboptimal Filtering for Linear Systems Driven by Self-Excited Poisson Processes", Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, October 1987.

[3] J. Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems", Proc. 1988 American Control Conference, Atlanta, June 1988.

[4] J. Ezzine and A. H. Haddad, "On the Stabilizability of Two-Form Hybrid Systems via Averaging," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

[5] B. S. Heck and A. H. Haddad, "On Linear Singularly Perturbed Systems with Quantized Control," Automatica, to appear.

[6] B. S. Heck and A. H. Haddad, "Singular Perturbation in Piecewise Linear Systems", Proc. 1988 American Control Conference, Atlanta, June 1988. (to appear also in IEEE Transactions on Automatic Control).

[7] B. S. Heck and A. H. Haddad, "Extensions of Singular Perturbation Analysis in Piecewise-Linear Systems," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

[8] J. A. Ramos, and E. I. Verriest, "A Note on the Cross Riccatian and Related Properties for Symmetric Stochastic Realizations", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, pp. 1171-1173, December 1987.

[9] E. I. Verriest, "A Unified Theory of Model Reduction via Gleason Measures" in Mathematics in Signal Processing, (T. S. Durrani, Ed.), Oxford University Press, 1987.

[10] T. K. Gaylord and E. I. Verriest, "Matrix Triangularization using Arrays of Integrated Optical Givens Rotation Devices," IEEE Computer, pp. 59-66, December 1987.

[11] E. I. Verriest, "Stochastic Modelling and Model Reduction in a Measure Theoretic Framework," Proc. 12th IMACS Congress on Scientific Computation, Paris, France, July 1988.

[12] E. I. Verriest, "On Three Dimensional Rotations, Coordinate Frames, and Canonical Forms for it all", Proceedings of the IEEE, 1988.

[13] E. I. Verriest, "Alternating Discrete Time Systems: Invariants, Parametrization and Realization," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

[14] E.I. Verriest, "Maximal Tori and Unique Canonical Singular Value Decompositions," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

[15] W. S. Gray and E. I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, pp. 124-128, December 1987.

Extensions of Singular Perturbation Analysis in Piecewise-Linear Systems

B.S. Heck and A.H. Haddad
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

Summary

This paper addresses problems in piecewise-linear systems which are singularly perturbed. Such systems are found in many applications including electrical circuits and in flight controls. The piecewise-linearity may be due to nonlinear elements such as saturation or may result from a linearization about various operating points of a nonlinear plant. Singular perturbation theory is used to separate the system into reduced-order models, one containing the slow dynamics and one containing the fast dynamics. Standard singular perturbation techniques, however, are limited to systems which are smooth [1]. Recently, it has been extended by these authors to certain types of piecewise-linear systems [2,3].

This paper extends the previous work done in singularly perturbed piecewise-linear systems. Specifically, new theorems are presented allowing for the application of this theory to a broader range of systems. The previous work was based on geometric ideas and the resulting theorems difficult to use.

The two types of systems analyzed are those which are continuous and those which are the result of a quantized control. Both types of systems may be expressed in the following form.

$$\dot{x} = f_1(x,z)$$

$$\mu\dot{z} = f_2(x,z)$$

where μ is a small positive parameter and f_1 and f_2 are piecewise-linear functions mapping from R^{r+p} to R^r and R^p, respectively. The state space is partitioned into nonintersecting regions of the form $S_i = \{(x,z): d_i \le K_x x + K_z z < d_{i+1}\}$ (where K_x and K_z are row vectors) so that both f_1 and f_2 are affine in the interior of each region.

Reduced-order models are given for the quantized control case in [2] and for the continuous piecewise-linear case in [3]. Conditions guaranteeing that the solutions of the reduced-order models match the solution of the actual system with an error on the order of $O(\mu)$ are given in these papers. The conditions given are restrictive in their application and are difficult use.

A discussion of the limitations in applying these previous conditions to physical systems will be included in the full paper. Also, the results are extended to include the occurance of a sliding mode in the quantized control case. A new nongeometric criterion is introduced for using singular perturbation in the continuous piecewise-linear case. This criterion is easy to use and the proof is straightforward. Finally, the effect of a random input will be examined.

References

[1] P.V. Kokotovic, R.E. O'Malley and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--An Overview," Automatica, Vol. 12, 1976, pp. 123-132.

[2] B.S. Heck and A.H. Haddad, "On Linear Singularly Perturbed Systems with Quantized Control," Proceedings of the Twenty-first Annual Conference on Information Sciences and Systems, March, 1987, Johns Hopkins University, Baltimore, Maryland, pp. 24-29.

[3] B.S. Heck and A.H. Haddad, "Singular Perturbation in Piecewise-Linear Systems," to appear in the Proceedings of the 1988 American Control Conference.

ON THE STABILIZABILITY OF TWO-FORM HYBRID SYSTEMS VIA AVERAGING

Jelel Ezzine and A. H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia; 30332-0250

SUMMARY

The paper discusses some practical methods of analysis and control of two-form hybrid systems. The main tools in simplifying the analysis and stabilizability of such systems will be some basic ideas from Lie algebras, linear algebra and a tool from stability theory of ordinary differential equations. These systems are called hybrid systems because the set of linear time-invariant systems among which the system is switching is finite. This kind of model can be used to represent systems subject to known abrupt parameter variations such as commutated networks or to approximate certain time-varying systems. This can be done by imposing a "deterministic" switching rule on the time behavior of the form index. However, to model unknown abrupt phenomena such as component and interconnection failures the form index can be modeled, for example, as a finite-state Markov chain (FSMC).

The class of hybrid systems considered in this paper are assumed to have the form

$$dx/dt = A(r(t))x(t) + B(r(t))u(t) \tag{1}$$

$$y(t) = C(r(t))x(t); \tag{2}$$

where x is the plant state vector of dimension n, u is the plant control input vector of dimension p, y is the plant output vector of dimension m, and r(t) is the "form index" which is a scalar sequence taking values in the finite index set $\mathbf{N}=\{1, 2, ..., N\}$.

In the first part of the paper a practical averaging technique is introduced which will be the key step in the analysis and control (stabilization) of two-form hybrid systems. The proposed averaging method applies to multi-form hybrid systems as well.

The averaging methodology used in the paper is based on a formula from Lie algebras known as the Baker-Campbell-Hausdorff Formula [1]: Given two real matrices A and B there is no guarantee that there exists a real matrix C such that

$$\mathrm{Exp}(A)\mathrm{Exp}(B) = \mathrm{Exp}(C). \tag{3}$$

This will be the case, however, if $\|A\|+\|B\|\leq\ln(2)$ [2], and then C will be given by a *convergent* infinite expression

$$C = A + B + (1/12)[[A,B],B] + (1/12)[[B,A],A] + \dots \qquad (4)$$

where the symbol $[A,B]\equiv AB-BA$ (i.e., the commutator product.) This expression is the Baker-Campbell-Hausdorff formula (BCH).

There are two very important issues in using such a formula and averages derived from it. The first one is the error introduced by only using few terms in the BCH expression while computing the average matrix. The second one is the difference between the average system and the actual system. In the paper we present what we beleive is the first treatment of such problems related to the accuracy of the usage of a truncated BCH formula. In the first part upper bounds of the errors introduced by using a truncated BCH formula are derived. The analysis of the difference between the F2-system and its average is delayed to the section which deals with the stability of hybrid systems.

Using the BCH formula, one can obtain an approximation $\hat{C}$ of C, to any order he wishes. Consequently, C can be written as $C=\hat{C}+\tilde{C}$, where $\tilde{C}$ is the unknown error due to the approximation. Therefore, the induced error in computing Exp(C) by using the approximate matrix $\hat{C}$ is

$$E \equiv \mathrm{Exp}(CT)-\mathrm{Exp}(\hat{C}T) \qquad (5)$$

The known formula for the solution of inhomogenious differential equations can be elegantly used to derive an *exact* expression of E. In the same section a useful approximate expression for E is derived using perturbation techniques. The results of the section is summarized in the following two propositions:

Proposition 1 [3]: Let E_1 denote the first order approximation in ε to E, then E_1 satisfies the following matric D.E.

$$dY/dt = \hat{C}Y + \varepsilon\tilde{C}_p\mathrm{Exp}(\hat{C}t), \quad Y(0) = 0. \qquad (6)$$

where $\tilde{C}\equiv\varepsilon\tilde{C}_p$, and ε is a scalar.

Proposition 2: Assume that $\|\mathrm{Exp}(\hat{C}t)\|\leq M(t)\mathrm{Exp}(\beta(t)t)$, with $\beta(t)$ a scalar function, then

$$\|E_1\| \leq (\|\tilde{C}\|/2\|\hat{C}\|)M(t)\mathrm{Exp}\{\beta(t)t\}(\mathrm{Exp}\{2\|\hat{C}\|t\}-1). \qquad (7)$$

Using other means [3], and the added assumption that M(t) is monotone one gets

$$\|E\| \leq t\|\tilde{C}\|M^2(t)\mathrm{Exp}\{(\beta(t) + M(t)\|\tilde{C}\|)t\}. \qquad (8)$$

It is also shown that it is possible, sometimes, to avoid the computation of A-"average". That is, under certain conditions, it is possible, *via state feedback*, to make the F2 system time-invariant in A. Necessary and sufficient conditions are derived for the existence of K and a compact computation recipe based on the Kronecker-product and the generelized-inverse techniques is given.

One of the key assumptions made to design the regulator via averaging is the controllability of the average system. This assumption is not unreasonable since the controllability property of linear time invariant systems is generic. However, one can construct hybrid systems such that their averages are not controllable.

The following theorem singles out a class of hybrid systems for which the average system is controllable too.

THEOREM C1: The average system of a hybrid system is controllable if
1-rank $[C_1, C_2, \ldots, C_N] = n$; C_i is the controllability matrix of Σ_i, i in N,
2-All forms are simultaneously diagonalizable,

The first condition in the theorem is a very interesting one, it was shown in [4] that it is a necessary condition for a hybrid system to be controllable, moreover it is very close to be a sufficient condition too!

Moreover, Theorem C1 is interesting in its owne right. It says that given a set of simultaniously diagonalizable systems then any element in the convex hull (or cover) of these systems is controllable.

The last part of the paper is based on the previous parts, it deals with the stabilizability of two-form hybrid systems. A stabilizability theorem based on the proposed averaging technique is given. The theorem requires that the average system be stabilizable and that the dynamics of the error between the actual system and the average one be stable. To check for the stability of the error a simple sufficient stability test based on the mathematical notion of Logarithmic norm [3] is derived.

Discussions, proofs and examples supporting the above summary are given in the paper.

REFERENCES:

[1] R. W. BROCKETT AND J. R. WOOD, Electrical networks containing controlled switches, in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, April 1974, San Francisco, pp. 1-11.

[2] J. G. F. BELINFANTE, Explicit version of the Campbell-Baker-Hausdorff formula: Integral representation for ln $e^x e^y$, unpublished.

[3] J. EZZINE AND C. D. JOHNSON, Analysis of continuous/discrete model parameter sensitivity via a perturbation technique, Proceedings of The Eighteenth Southeastern Symposium on System Theory, 1986, pp. 545-550.

[4] J. EZZINE AND A. H. HADDAD, On the controllability and observability of hybrid systems, Accepted for the ACC 1988.

Poston, T. and Stewart, I., (1978) "Catastrophe Theory and its Applications". Pitman, London.

A UNIFIED THEORY OF MODEL REDUCTION VIA GLEASON MEASURES

E.I. Verriest
*(School of Electrical Engineering,
Georgia Institute of Technology, Atlanta, Georgia)*

ABSTRACT

Earlier work has cast the stochastic realization and approximation problem in the framework of the RV-coefficient. This allowed the introduction of a common measure for the "goodness of fit" for the different realization algorithms. This paper explores the deeper geometrical basis for this common measure in a unified theory for the data driven and exact covariance approaches.

1. INTRODUCTION

1.1 Scope of the Paper

In the theory of identification, signal processing, and digital filtering, a problem of fundamental importance is that of finding a finite dimensional Markovian representation of a stochastic process from the covariance information. This problem is known as the Stochastic Realization Problem, and has received a great deal of attention. Whenever a finite set of real data is gathered, all processing is done over finite sets, and an underlying probabilistic description is absent in most cases. As a result, covariances must be estimated by sample covariances, and a "degradation" of the theoretical realization solutions results. A more direct, data driven approach is needed. Moreover, for many applications, the Markovian representation or state space model may be too complex, due to its high dimensionality, thus barring efficient computational management. This motivates the quest for approximate lower order models, and the need for common measures to evaluate and compare different approaches.

1.2 *Historical Background*

Akaike (1975), Faurre (1976), and Baram (1981) developed a stochastic realization theory based on the information interface between the past and the future of a time series and the concepts of canonical correlation analysis (CCA). Desai and Pal (1982) extended the results. They obtain forward-backward dual models with state covariances which are equal and diagonal. They are the stochastic counterpart of the deterministic balanced realizations. See Moore (1981) and Verriest and Kailath (1983).

Arun and Kung (1983) proposed the Karhunen-Loeve method (KLM) as a basis for the stochastic realization.

Ramos and Verriest (1984) unified the theory by showing that both CRA and KLM, given the exact covariances, are special cases of a more general optimization problem, using the RV-coefficient introduced by Escoufier (1973). Verriest (1985) explored the connection of the exact covariance and real data case further by relating the RV-coefficient to certain operators in a tensor product space.

2. STOCHASTIC REALIZATION PROBLEM

2.1 *Problem Formulation*

Given the covariance sequence $\Lambda(k)$ of a rational stationary, zero mean, discrete time vector sequence $\{y_k\}$, the stochastic realization problem consists in finding a Markovian representation of the form

$$x_{k+1} = Fx_k + w_k \qquad (2.1.1)$$

$$\bar{y}_k = Hx_k + v_k \qquad (2.1.2)$$

where $\{w_k\}$ and $\{v_k\}$ are White Gaussian noises with

$$E\begin{bmatrix} w_k \\ v_k \end{bmatrix} [w_e' \quad v_e'] = \begin{bmatrix} Q & S \\ S' & R \end{bmatrix} \delta_{k,\ell} \; ; \forall k,\ell \qquad (2.1.3)$$

such that $E(\bar{y}_{k+n}\bar{y}_k') = \Lambda(k)$. $\delta(k,\ell)$ is the Kronecker delta.

The solution to this problem is described by Faurre (1976).

2.2 *Information Interface Between Past and Future*

Assume that the stochastic time-series $\{y_k\}$ is Gaussian (with zero mean). The relevant random variables are then in the Hilbert space $L_2(\Omega,B,P)$ and conditional expectations can be interpreted as orthogonal projections onto subspaces $L_2(\Omega, F_k^{\{y_k\}}, p)$.

For the time-series $\{y_k\}$ define the infinite vectors

$$Y_k^+ = \begin{bmatrix} y_k \\ y_{k+1} \\ \vdots \end{bmatrix}, \text{ the future} \qquad Y_k^- = \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ \vdots \end{bmatrix}, \text{ the past} \qquad (2.2.1)$$

and define the semi-infinite covariance matrices

$$\hat{H} = E\{Y_k^+(Y_k^-)'\},\ R^+ = E\{Y_k^+(Y_k^+)'\}\ ,\ R^- = E\{Y_k^-(Y_k^-)'\}\ . \qquad (2.2.2)$$

Within this representation, the forward and backward predictor subspaces are

$$X_k = \mathrm{Span}(Y_k^+ \mid Y_k^-),\ Z_{k-1} = \mathrm{Span}(Y_k^- \mid Y_k^+) \qquad (2.2.3)$$

$(A|B)$ denotes the projection of span (A) onto the Hilbert space spanned by the components of B. These two spaces form the information interface between R^+ and R^-. Either one can be used to define a minimal Markovian representation (forwards or backwards). The canonical correlations lead to a natural distance measure between the past and the future, which for the Gaussian case is exactly the Kullback-Leibler information.

Alternatively, the past can be treated as the instrumental variables for predicting the future. See Arun and Kung (1983).

Ramos and Verriest (1984) and Ramos (1985) resolved the two methods by putting them in a common framework, optimizing Escoufier's RV-coefficient (1973) under different constraints. If for random vectors X and Y

$$\mathrm{cov}(X,Y)=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}, \text{ then } \mathrm{RV}(X,Y) = \frac{\mathrm{Tr}(\Sigma_{12}\Sigma_{21})}{\sqrt{(\Sigma_{11}^2)\ \mathrm{Tr}(\Sigma_{22}^2)}} \geqslant 0 \tag{2.2.4}$$

This measure also allows the computation of a "figure of merit" for each algorithm in a consistent way. Our new Gleason approach gives a natural interpretation for (2.2.4) (and other) measures.

3. THE LATTICE OF SUBSPACES AND GLEASON MEASURES

Let H be a Hilbert space. The set of all closed subspaces of H has the structure of an orthocomplemented complete lattice, also called a logic. A one-to-one correspondence exists between the lattice of all closed subspaces of H and the lattice Proj H of all orthoprojectors on H.

In his study of the mathematical foundations of quantum mechanics, Mackey posed the problem of finding all positive measures on the closed subspaces of a Hilbert space. Such a measure must have the property that for any countable collection $\{S_i\}$ of mutually orthogonal closed subspaces the mapping is σ-additive, i.e.,

$$\sum_i \mu(S_i) = \mu(\sum_i S_i) \tag{3.1}$$

A measure satisfying the above property is for instance obtained by selecting a vector v in the Hilbert space H, and for each subspace A of H defining

$$\mu_v(A) = \| P^A(v) \|^2 \tag{3.2}$$

where P^A is the projection operation on A. Clearly, finite convex conbinations of such measures also satisfy the conditions for such measures, and passing to the limit, any positive semidefinite trace class operator T also defines such a measure via

$$\mu(A) = \mathrm{Tr}(TP^A) \tag{3.3}$$

Gleason (1957) has shown that in a separable Hilbert space of dimension at least three, every measure on the closed subspaces can be represented as above, with T a positive definite operator of trace class. Further extensions of Gleason measures which are vector and operator valued, have been surveyed by Jajte (1979).

Let H be a separable Hilbert space, and Proj H the lattice of all orthogonal projectors in H. Let also E be some Banach space, then:

<u>Definition</u>: A mapping ξ : Proj H → E is said to be an E-valued Gleason measure if

(1) For any sequence of pairwise orthogonal projectors $P_1, P_2, \ldots$ from Proj H

$$\sum_i \xi P_i = \xi(\sum_i P_i) \tag{3.4}$$

the series on the left hand side being weakly convergent,

(2) $$\sup\{\| \xi P_i \| : P_i \in \text{Proj H}\} < \infty \tag{3.5}$$

An important class of Gleason measures taking values in a Hilbert space are the Orthogonally Scattered Measures (OSG). Let H and K be two Hilbert spaces, with dimension at least three.

<u>Definition</u>: A Gleason measure ξ : Proj H → K is said to be an orthogonally scattered measure (OSG-measure) if for any orthogonal projectors P,Q in Proj H the following implication holds.

$$P \perp Q \longrightarrow \xi p \perp \xi Q \tag{3.6}$$

Note that automatically for all P $\|\xi P\| \leq \|\xi I\|$ is implied, where I is the identity operator, corresponding with the (sub) space H.

Any OSG-measure defines a positive Gleason measure by

$$\mu P = \|\xi P\|^2, \ P \in \text{Proj H} \tag{3.7}$$

By Gleason's theorem, there exists then a nonnegative self-adjoint trace-class operator T such that

$$\mu P = \mathrm{Tr}\ TP\ ,\ P \in \text{Proj H} \tag{3.8}$$

The above can be interpreted as a "variance"; if we similarly define a "covariance" by $\mathrm{COV}\{P,Q\} = (\xi P, \xi Q)_K$, then for any commuting projectors P, Q ε Proj H one has

$$(\xi P, \xi Q)_K = \mathrm{Tr}\ TPQ \qquad (3.9)$$

This formula is not true in general for arbitrary projectors in a complex Hilbert space. However, if H and K are real Hilbert spaces, then it can be shown that (for T given by Gleason's theorem (3.9))

$$(\xi P, \xi Q)_K = \mathrm{Tr}\ TPQ = \mathrm{Tr}\ TQP, \text{ all } P, Q \in \mathrm{Proj}\ H \qquad (3.10)$$

4. APPLICATIONS TO REALIZATION THEORY

4.1 A Correlation Measure for Subspaces

As shown by Verriest (1985), the (exact) stochastic realization and the (real) signal modelling benefit from the use of the RV-coefficient. In the first case, the formalism is used in the comparison of random variables, while in the second it compares data matrices. We present here a geometric point of view, motivated by the observation that for the stochastic realization problem, the underlying space $L_2^p(\Omega,B,m)$ and in the real data case, the space R^{pxN} are isomorphic with the tensor product spaces, respectively

$$L_2^p(\Omega,B,m) \sim R^p \otimes L_2(\Omega,B,m) \qquad (4.1.1)$$

$$R^{pxN} \sim R^p \otimes R^N \qquad (4.1.2)$$

In general, let G and H be separable Hilbert spaces. Let $\{\phi_i\}$ be a Complete Orthonormal Set (CONS) in G, and $\{\psi_i\}$ a CONS in H. Any vector x in the tensor product space $G \otimes H$ has a decomposition

$$x = \sum_i \mid x_i > < \psi_i \mid \qquad (4.1.3)$$

where x_i ε G. The vector x in the tensor product space will be referred to as a "prior." Introduce now the superposition of measures on Proj G induced by the prior.

$$\mu_x = \sum_i \mu_i \qquad (4.1.4)$$

The μ_i are the Gleason measures corresponding to x_i. For all

subspaces of G, it follows that

$$\mu_x(A) = \mathrm{Tr}\ T_x P^A \qquad (4.1.5)$$

where

$$T_x = \sum_i \mid x_i > < x_i \mid = xx' \qquad (4.1.6)$$

is interpreted as a gramian or covariance operator.

The measure $\mu_x(A)$ gives a numeric value to the closeness of A to G, given the prior x.

The problem of finding the subspace of fixed dimension which "looks most like H from the point of view of x" is then solved by letting P^A be the projector on the eigenspace of T_x with the largest principal components. See Aragon and Couot (1976), who also stated several equivalent problems relating to the principal component analysis. Note that $\mu_x(H) = \mathrm{Tr}\ T_x$.

However, this measure does not lead to a useful definition of the correlation between subspaces. Indeed, consistent with the above "variance" μ_x we have the covariance (using 3.10)

$$(\xi_x(A),\ \xi_x(B)) = \mathrm{Tr}\ T_x P^A P^B \qquad (4.1.7)$$

But for A ⊥ B, we get $(\xi_x(A),\ \xi_x(B)) = 0$. There is no interface between A and B. This situation is displeasing, but can be resolved. The operator $T_x : G \to G$ is a characteristic for the given x in $G \otimes H$ (in fact, a "sufficient statistic"), and one can think of T (or μ) as conditioned by the vector x ε $G \otimes H$. In this sense, the extended projectors $\tilde{P}^B$ ε Proj $G \otimes H$ defined by

$$\tilde{P}^B x = \sum_i P^B \mid x_i > < \phi_i \mid = \sum_i \xi_{x_i}(B) < \phi_i \mid \qquad (4.1.8)$$

yield a "coherent" addition of OSG measures, conditioned on x (i.e., a posterior measure). The posterior variance of A ε Proj G, given x is then the operator from G → G

$$(\tilde{P}^A x)(\tilde{P}^A x)' = \sum_i P^A |x_i><x_i| P^A = P^A T_x P^A \qquad (4.1.9)$$

and the covariance

$$(\tilde{P}^B x)(\tilde{P}^A x)' = \sum_i P^B |x_i><x_i| P^A = P^B T_x P^A \qquad (4.1.10)$$

This displays the coupling or interface between A and B given x. In order to attach a numerical value to this interface, any norm on the various restrictions $P^B T_x P^A$ can be chosen. The following natural definition follows.

<u>Definition</u>: The "Correlation" between subspaces A and B in Proj G is

$$\rho(A,B|x) = \frac{\| P^A T_x P^B \|}{\sqrt{\| P^A T_x P^A \| \, \| P^A T_x P^A \|}} \qquad (4.1.11)$$

Let K be the fixed subspace span $(\phi_1,\dots,\phi_k)$ of G, then the principal component analysis and canonical correlation analysis are respectively (in Frobenius norm) (O(K) is the orthogonal group on K)

$$\text{PCA} : \max_{M \in O(K)} \frac{\text{Tr}\,(MT_{12}T_{21}M')}{\sqrt{\text{Tr}\,(MT_{11}M')^2 \; \text{Tr}\,(T_{22})^2}} \qquad (4.1.12)$$

$$\text{CCA} : \max_{\substack{M \in O(K) \\ N \in O(K^{\perp})}} \frac{\text{Tr}\,(MT_{12}NN'T_{21}M')}{\sqrt{\text{Tr}\,(MT_{11}M')^2 \; \text{Tr}\,(NT_{22}N')^2}} \qquad (4.1.13)$$

which are the formulas obtained by Robert and Escoufier (1976). If $G \otimes H = L^2_p(\Omega,B,m)$ or R^{pN} then T is respectively the covariance Σ or sample covariance S.

4.2 Data Driven Stochastic Realization Solution

Assuming that a data stream $\{y_k, |k| \leq N\}$ is observed, a data matrix Y can be formed by considering $(Y_{-N},0,\dots,0)'$, $(Y_{-N+1},Y_{-N},0,\dots,0)'$ ... $((Y_N,Y_{N-1},\dots,Y_{-N})'$, ..., $(0,\dots,0,Y_N)'$ as consecutive samples of the vector in $G = R^{2N+1}$ (the "pure states"). In order to avoid the nasty end effects due to the substitution of zeros where data is missing, a linear superposition of these states, weighted by the sequence $\{q_j \geq 0;\ j=1,\dots,4N+1\}$ may be used. Let $Q = \text{diag}\{q_i\}$. The Gleason operator is $T_{(Y,q)} = YQY'$. The "past" is span $\{|\phi_{-N}>\dots|\phi_{-1}>\}$, and the future span $\{|\phi_0>\dots|\phi_N>\}$. A recursive realization algorithm, which optimally uses all the data (in real time) would necessarily involve the update of $T_{(Y,q)_N}$.

5. CONCLUSIONS

By determining well motivated measures for the correlation between subspaces of a Hilbert space, based on the available prior information, it was possible to unify several tools from multivariate analysis, and introduce common measures for their evaluation. We have only discussed the principal component and the canonical correlation analyses. Discriminant analysis and a rational way for discarding variables can also be treated. Our inspiration for this work came from the desire to better motivate and explain the use of the RV-statistic problems. In particular, exact realization theory and its signal processing counterpart (i.e., the real data case) are unified. The fact that the data should come first looks natural from this viewpoint. Deterministic modelling, cluster analysis (in pattern recognition), and quantization of random fields are other applications of our abstract framework. The variation lies in the choice of the spaces G and H, and the constraints that are natural for the problem.

6. REFERENCES

Akaike, A., (1975) Markovian Representation of Stochastic Processes by Canonical Variables, *SIAM J. Control* **13**, No 4, 162-173.

Aragon, Y. and Couot, J., (1976) Une Definition de l'Operateur d'Escoufier, *Comptes rendus*, 283, series A, 867-869.

Arun, K.S., Bhaskar Rao, D.V. and Kung, S.Y., (1983) A New Prediction Efficiency Criterion for Approximate Stochastic Realization, Proc. 22nd Conf. on Decision and Control, 1353-1365.

Baram, Y., (1981) Realization and Reduction of Markovian Models from Nonstationary Data, *IEEE Trans. Automatic Control AC-26*, no. 6, 1225-1231.

Desai, U.B. and Pal, D., (1982) A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations, Proc. 21st Conf. on Decision and Control, 1105-1111.

Escoufier, Y., (1973) Le Traitement des Variables Vectorielles, *Biometrics* 29, 751-760.

Faurre, P.L., (1976) Stochastic Realization Algorithms, in: Mehra, R.K. and Lainiotis, D.G., (eds.), System Identification: Advances and Case Studies, (Academic Press).

Gleason, A.M., (1957) Measures on the Closed Subspaces of a Hilbert Space, *J. Math. Mech.* 6, 885-893.

Jajte, R., (1979) Gleason Measure, in: Bharucha-Reid (ed.), Probabilistic Analysis and Related Topics, Vol. 2 (Academic Press).

Moore, B.C., (1981) Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, *IEEE Trans. Automatic Control*, AC-26, No. 5 (1) 17-32.

Ramos, J.A., (1985) A Stochastic Approach to Streamflow Modelling, Ph.D. Dissertation, School of Civil Engineering, Georgia Institute of Technology.

Ramos, J.A. and Verriest, E.I., (1984) A Unifying Tool for Comparing Stochastic Realization Algorithms and Model Reduction Techniques, Proc. 1984 ACC, San Diego.

Robert, P. and Escoufier, Y., (1976) A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient, *Appl. Statist.* 25, no. 3, 257-265.

Verriest, E.I., (1985) Projection Techniques for Model Reduction, in Lindquist, A., and Byrnes, C., (eds.) Modelling, Identification and Robust Control, (North Holland).

Verriest, E.I. and Kailath, T., (1983) On Generalized Balanced Realizations, *IEEE Trans. Automatic Control, AC-28*, no. 8, 833-844.

A FRESH APPROACH TO THE DERIVATIVE SAMPLING THEOREM

J.R. Higgins
*(Department of Science,
Cambridgeshire College of Arts and Technology, Cambridge)*

ABSTRACT

Convergence properties of the derivative sampling series are addressed. Two forms of the series are discussed, one which allows for over-sampling, the other in which sampling is at Nyquist rate. Some attention is given to the way in which the two parts of the series partition the signal to be reconstructed, and to which subsets of the Hilbert space of finite energy band-limited signals these parts belong.

1. INTRODUCTION

The reconstruction of a band-limited signal from knowledge of its samples together with samples of its derivatives was suggested by C.E. Shannon (see, e.g., Higgins (1985) p. 60 and the references cited there). The formula

$$f(t) = \sum_{n=-\infty}^{\infty} \left\{ \frac{f'(n)}{\pi} \, \frac{\sin^2 \pi(t-n)/2}{\pi(t-n)/2} + \frac{f(n)}{2} \left[\frac{\sin \pi(t-n)/2}{\pi(t-n)/2} \right]^2 \right\} \tag{1.1}$$

is known to hold for every f belonging to the Hilbert space PW of Paley-Wiener functions, that is, the class of finite energy signals band-limited to $[-\pi,\pi]$. The convergence is in the norm of PW and also pointwise, uniformly over all of $\mathbb{R}$.

For the sake of brevity let

$$S_p^q(t) \triangleq \frac{\sin^q(\pi t/q)}{(\pi t/q)^p}, \quad 0 < p \leq q.$$

Matrix Triangularization Using Arrays of Integrated Optical Givens Rotation Devices

Thomas K. Gaylord and Erik I. Verriest
Georgia Institute of Technology

Integrated optical chips, implemented using waveguides and voltage-tunable diffraction gratings, can be used to solve sets of linear equations.

Solving linear equations is centrally important in many large-scale data processing problems. For example, problems such as weather prediction and the aerodynamic design of aircraft require repeated solution of the Navier-Stokes equation. The describing nonlinear systems of equations can be linearized at each time step to produce a linear system of equations that simplifies the problem-solving process.

In general, there are two approaches to solving sets of linear equations—direct methods and iterative methods. Perhaps the best known of the direct methods is Gaussian elimination. This method is generally sequential in nature. It also has the potential for being unstable: small errors in the intermediate steps may produce large errors in the final results. Iterative methods, on the other hand, are generally parallel and stable. However, they require an approximate matrix inverse as a starting point.

Optics-based devices and systems offer one means of solving sets of linear equations using a stable, direct method. These devices can employ Givens rotations[1] to perform matrix triangularization. In this article, we describe how to solve linear equations using an electro-optical system that employs arrays of optical Givens rotation devices. We examine how to implement this system using two different configurations, parallel and pipelined, and how to calibrate the system to minimize errors.

Attributes of optics

In telecommunications, fiber optics has played a dramatically increasing role in recent years. In addition, the role of guided wave optics (consisting of both fiber optics and integrated optical circuits) is expected to continue to grow at a rapid rate. This activity is largely oriented to switching networks. However, it will have a direct impact on signal and data processing as well.

The use of optics in computation is an exciting field offering great potential for large-scale, high-speed computing power. However, the nature of this potential must be well understood before it can be successfully utilized. The favorable and unfavorable characteristics of optics must be understood in relation to those of electronics so that overall optoelectronic systems can be designed to maximize the desirable features of each. Optics inherently has four powerful attributes:

- **Large bandwidth.** The high carrier frequency ($\approx 10^{14}$ Hz) offers the potential for very high speed operation. This attribute is primarily responsible for the success of fiber optics.
- **Parallelism.** Integrated optical (two-dimensional) and bulk optical (three-dimensional) systems are capable of handling and processing many channels of data simultaneously.
- **Interconnectivity.** In optical form, channels of data can physically pass

0018-9162/87/1200-0059$01.00 ©1987 IEEE

through each other without altering the data. This property distinguishes optical signals significantly from the charge-based signals in metallic conductors, which must remain separate from each other. Interconnectivity allows the switching (interchanging or broadcasting) of data channels in any arbitrary pattern.

- Special functions. Numerous analytic functions can be implemented directly with optics. The best known of these is the Fourier transform, which gave rise to the field of "Fourier optics."[2] Other transforms (the Hadamard, the Hartley, the Mellin, the Radon, etc.) can also have direct optical implementations. Similarly, the sine and cosine can be implemented in optical form and are central to the type of processing described in this article.

A primary disadvantage of analog optics is low accuracy. This shortcoming makes these systems appropriate for fast, first-pass processors used in applications that do not require high accuracy. The attributes of optics thus differ dramatically from those of electronics. In general, a one-for-one substitution of optical devices for electronic devices in computing architectures should not be attempted. Such a strategy can be a fundamental mistake, as evidenced by some notable failures in the past.

Hybrid optical-electronic processing systems must efficiently use the inherent attributes of both optics and electronics in order to provide the advantage needed for large-scale complex processing problems. In this article, we describe a hybrid system that requires optical-electronic phase-sensitive detection and electronic feedback. Our discussion concentrates on the new technology involved in the optical component of this system.

The Givens rotation

The plane rotation operation for a rotation can be expressed as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad (1)$$

This equation can be written compactly as $\bar{y} = \mathbf{R}\bar{x}$, where the output vector $\bar{y} = [y_1, y_2]^T$, **R** is the rotation matrix, the input vector $\bar{x} = [x_1, x_2]^T$, and T denotes "transpose." (Throughout this article, a symbol in boldface type denotes a matrix.) The Givens orthogonalization[1] is obtained when $\sin\psi$ and $\cos\psi$ are found such that $y_1 = 0$. This operation can be used to make any specific element of a vector a zero. In fact, all but one entry of a vector can be zeroed by successive Givens rotations (involving different entries of the vector). For an $N \times N$ matrix, the triangularization process zeros all the elements above the diagonal (producing a lower triangular matrix **L**) or zeros all the elements below the diagonal (producing an upper triangular matrix **U**). The process of producing a lower triangular matrix for the case of $N = 5$ can be represented schematically as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \rightarrow \qquad (2)$$

$$\begin{bmatrix} b_{11} & 0 & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 & 0 \\ b_{31} & b_{32} & b_{33} & 0 & 0 \\ b_{41} & b_{42} & b_{43} & b_{44} & 0 \\ b_{51} & b_{52} & b_{53} & b_{54} & b_{55} \end{bmatrix}$$

To start the process of lower matrix triangularization, $N-1$ rotations involving entries $i = 1, \ldots, N$ and $j = N$ are applied to transform the Nth column into the vector $[0, \ldots, 0, b_{NN}]^T$. The same rotations (in the same order) are used to transform columns 1 to $N-1$. The triangularization of the $N \times N$ matrix is accomplished recursively by zeroing the $N-1$ column of the resulting upper left $N-1 \times N-1$ submatrix, and so forth. Subsequent operations do not change the values in previously zeroed columns.

This algorithm lends itself naturally to cascaded or pipelined hardware implementations. Due to the nonlinear $\sin\psi$ and $\cos\psi$ functions, the Givens rotation operation consumes a significant amount of time and/or semiconductor material when implemented in digital electronics, even though efficient bit-recursive methods using simple shift and add operations known as coordinate rotation digital computing (Cordic) have been developed.[3]

Integrated optical Givens rotation device

The Givens rotation operation simulates a form of wave propagation and can be modeled as a lossless transmission line structure. Thus, wave propagation effects are a relevant design factor in the construction of this device. In a recent article, we reported on a coherent integrated optical implementation of an elementary rotation matrix device that operates on optical amplitude.[4] This device uses electro-optic grating diffraction and phase shifting to achieve the required sine evaluation, cosine evaluation, multiplications, addition, and subtraction in the Givens rotation operation.

The evaluation of sine and cosine is accomplished naturally and straightforwardly via diffraction by a thick transmission phase grating[5] induced by a voltage applied to periodic metallic electrodes on the surface of the device. The multiplication of the input amplitudes by the sine and cosine is accomplished as part of the diffraction process.

The summations in the Givens rotation are achieved by coherently combining the output waves from the grating. The phases of the waves are adjusted with electro-optic phase shifters to achieve the required addition and subtraction indicated in Equation 1. The subtraction process (for the y_1 output) is equivalent to coherent image subtraction. This operation may be performed using thick holograms and has been analyzed and experimentally demonstrated by Guest, Mirsalehi, and Gaylord.[6] The coherent addition process (for the y_2 output), likewise, is well established.

All of these functions can be combined into a single Givens rotation device as illustrated schematically in Figure 1a. A top view of an integrated optical implementation of this device is shown in Figure 1b. The optic axis is perpendicular to the surface. The crystalline material is Z-cut lithium niobate. The input light signals of amplitudes x_1 and x_2 are guided as transverse magnetic (TM) modes in channel waveguides (≈ 8 μm wide). The interdigitated electrodes on this electro-optic material have a period, Λ, and an orientation such that the Bragg condition for diffraction is satisfied for both input waves for the freespace optical wavelength, λ. The angle of rotation, ψ, in Equation 1 is the grating strength parameter. It is proportional to the voltage, V_g, applied to the interdigitated electrodes and is given approximately by

$$\psi = [\pi d_g n_E^3 r_{33} V_g]/[\lambda\Lambda\cos(\alpha/2)] \qquad (3)$$

where d_g is the thickness of the grating, n_E is the index of refraction for the TM mode polarization, r_{33} is the electro-optic coefficient for this configuration, and α is the angle between the waveguides. One arm of the device contains electro-optic phase shifters to which static voltages are applied to produce the correct phase relationships between the input waves and between the output waves.[4]

This Givens rotation device is potentially simple to fabricate. It can be constructed by (1) fabricating (by diffusion or proton exchange) the channel waveguides, (2) growing a SiO_2 buffer layer over the surface, and (3) depositing the metal electrodes. In fact, this general type of device has been developed and fabricated for intensity modulation and switching applications[7,8] and recently analyzed for Givens rotation-type applications.[9]

Arrays of Givens rotation devices for matrix triangularization

In 1958, Givens[1] pointed out that plane rotations could be used to triangularize a matrix by applying elementary rotation operations repeatedly over pairs of ele-

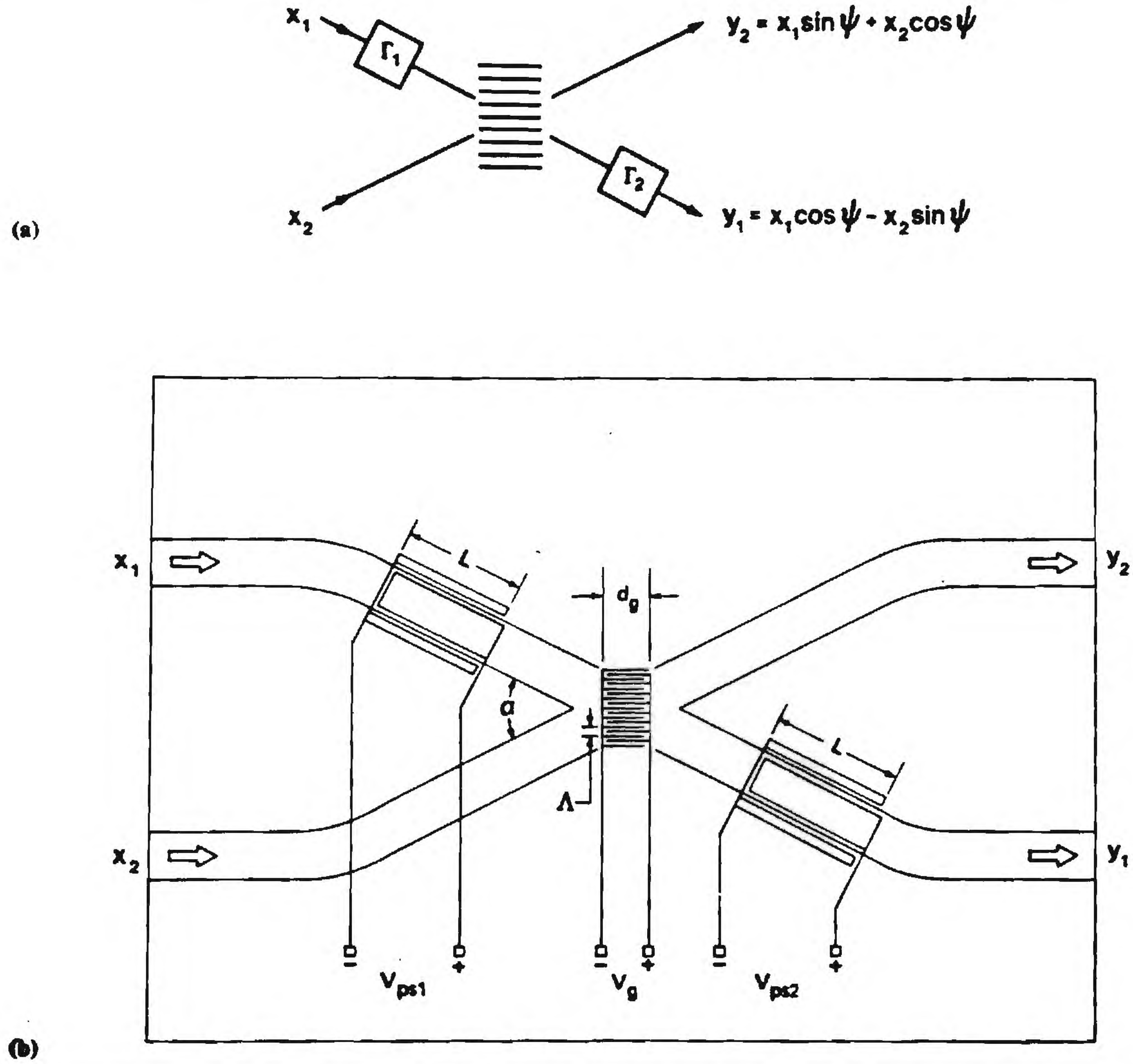

Figure 1. (a) The implementation of the elementary rotation operation. The optical input amplitudes x_1 and x_2 are diffracted by the thick grating and coherently combined to produce the output amplitudes y_1 and y_2. The external phase shifters (Γ_1 and Γ_2) have fixed values so that the transmitted and diffracted waves combine in phase (addition) for the y_2 output and combine 180 degrees out of phase (subtraction) for the y_1 output. (b) Schematic physical configuration of integrated optical elementary rotation device. A refractive index grating is formed at the intersection of the channel waveguides through the electro-optic effect by applying a voltage (V_g) to the interdigitated electrodes.

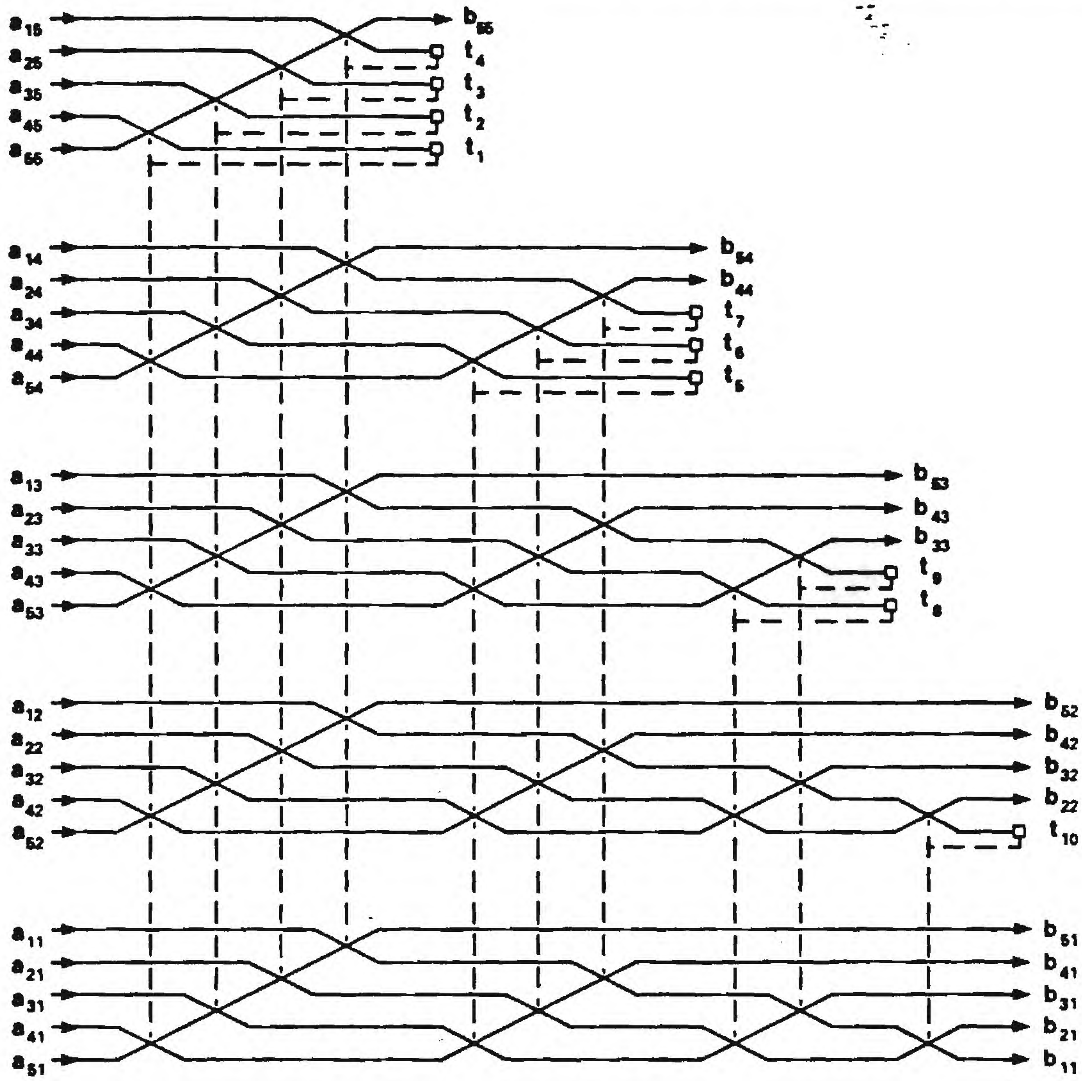

Figure 2. Parallel architecture for matrix triangularization. The solid lines are optical channel waveguides. A rotation device is located at each intersection of the channel waveguides. All rotation devices in a vertical column are electrically connected in parallel as indicated by the dashed lines. The squares represent photodetectors used to detect the nulling of matrix elements.

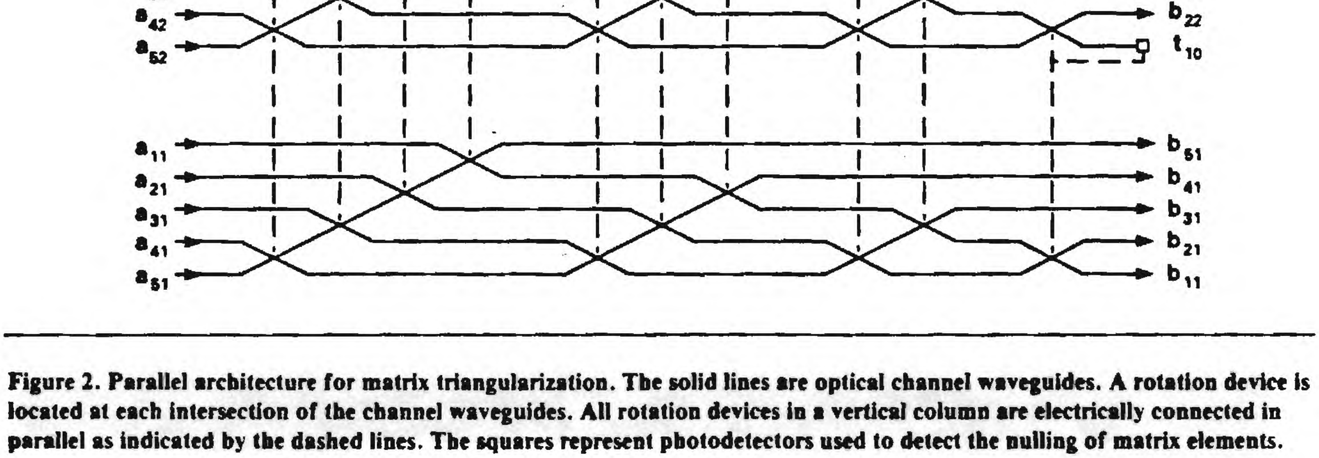
a_{11} a_{12} a_{13} a_{14} a_{15}
a_{21} a_{22} a_{23} a_{24} a_{25}
a_{31} a_{32} a_{33} a_{34} a_{35}
a_{41} a_{42} a_{43} a_{44} a_{45}
a_{51} a_{52} a_{53} a_{54} a_{55}
t_4
t_3
t_7
t_2
t_6
t_9
t_1
t_5
t_8
t_{10}
b_{51} b_{52} b_{53} b_{54} b_{55}
b_{41} b_{42} b_{43} b_{44} 0
b_{31} b_{32} b_{33} 0 0
b_{21} b_{22} 0 0 0
b_{11} 0 0 0 0

Figure 3. Pipelined architecture for matrix triangularization. The solid lines represent optical channel waveguides. A rotation device is located at each intersection of the channel waveguides. The squares represent photodetectors.

ments within the matrix. The integrated optical Givens rotation device described in the previous section has the capability of performing this operation. In principle, a single device could be used many times to accomplish the matrix triangularization. However, a more practical application is to integrate many Givens rotation devices on a single dielectric substrate in a manner analogous to integrating transistors on a single semiconductor substrate. Matrix triangularization can be accomplished by a parallel architecture or a pipelined architecture.

Figure 2 illustrates a parallel architecture implementation for the $N=5$ matrix triangularization that we used as an example above (see Equation 2). In this configuration, all N^2 matrix elements are entered into the system simultaneously. The solid lines in Figure 2 represent channel waveguides. At each intersection of channel waveguides is a rotation device as described in the previous section. Each dashed line represents a pair of metal conductors. All Givens rotation devices in a vertical column are connected together in parallel as shown. The optical amplitudes corresponding to the elements of the Nth column of the matrix enter the N channel waveguides at the top of the diagram. Detectors are represented by squares at the right. A detector at the end of each channel waveguide corresponds to matrix elements to be zeroed in the triangularization process.

The time steps associated with the triangularization process are labeled t_i (in this case, t_1 to t_{10}). In the first time interval, t_1, the value of $a_{4\,5}$ is initially detected (by the detector labeled t_1). The voltage applied to the leftmost column of rotators is swept until the amplitude detected is zero. This step zeros the $a_{4\,5}$ element and changes the $a_{5\,5}$ element according to Equation 1. Simultaneously, the elements a_{4j} and a_{5j} ($j=1$ to 4) are changed by the same rotation. The voltage on this column of rotators remains at this value for the remainder of the triangularization process. In the second time interval, t_2, the value of $a_{3\,5}$ is detected (by the detector labeled t_2). The voltage applied to the second column of rotators is swept until the detected amplitude of $a_{3\,5}$ is zero. The $a_{5\,5}$ element is correspondingly changed. At the same time, the elements a_{3j} and a_{5j} ($j=1$ to 4) are similarly transformed. After the time interval t_4, the fifth column has been entirely zeroed except for one element. This element has been transformed four times and is now $b_{5\,5}$, an element of the final triangularized matrix. All elements of the other columns experience the same transformations that zeroed the fifth column since the rotators in a vertical column are connected in parallel. In the fifth time interval, t_5, the new value of $a_{3\,4}$ is detected. The voltage applied to the fifth column of rotators is set so as to zero this amplitude. This step correspondingly changes the value of the $a_{4\,4}$ element and the other elements in this column of rotators. These operations continue through all time intervals. The output amplitudes shown in Figure 2 ($b_{1\,1}$ through $b_{5\,5}$) are the element values of the triangularized matrix. The total number of time steps and number of detectors is $(N^2-N)/2$. The total number of rotators required is $(N^3-N)/3$.

The $a_{i\,j}$ input elements and the $b_{i\,j}$ output elements may be positive or negative real numbers. The detected optical wave associated with a negative number is shifted in phase by 180 degrees relative to the phase of a positive number. Therefore, the values of the triangularized matrix must be detected using phase-sensitive techniques such as heterodyne or homodyne detection.[10] Thus the output amplitudes are detected in both magnitude and phase.

A pipelined architecture implementation for matrix triangularization using integrated optical Givens rotation devices is shown in Figure 3. Again, this architecture implements a triangular matrix of $N=5$. In this configuration, one column of N matrix elements is entered in parallel into the system. The optical amplitudes corresponding to the elements of the Nth column of the matrix enter first. This pipelined architecture includes a detector for each rotation device.

In the first time interval, t_1, the value of $a_{4\,5}$ is detected (by the detector labeled t_1) and the voltage applied to the first rotator is swept until the amplitude detected is zero. This step transforms the $a_{5\,5}$ element according to Equation 1. The voltage on this rotator remains at this value for the remainder of the triangularization process. Later, when the $N-1$ column arrives, the elements $a_{4\,4}$ and $a_{5\,4}$ are transformed by the same rotation. In the second time interval, t_2, the value of $a_{3\,5}$ is detected (by the detector labeled t_2). The voltage applied to this rotator is swept until the detected amplitude is zero. This step changes the $a_{5\,5}$ element correspondingly. After the time interval t_4, the fifth column has been entirely zeroed except for the $b_{5\,5}$ element of the final triangularized matrix. One after another, the elements of the other columns experience the same transformations that zeroed the fifth column. It experiences no further changes as shown by the straight section of waveguide in Figure 3. In the fifth time interval, t_5, the new value of $a_{3\,4}$ is detected. The voltage applied to this rotator is set so as to zero this amplitude. The value of $a_{4\,4}$ is correspondingly transformed. This mode of operation continues. Finally, the first column enters and experiences all of the previously set rotations. The output amplitudes shown ($b_{1\,1}$ through $b_{5\,5}$) are the element values of the triangularized matrix.

The total number of time steps is $(N^2+N-2)/2$. The number of detectors is again $(N^2-N)/2$. The total number of rotators required is $(N^2-N)/2$, making the pipelined architecture a more practical configuration than the parallel architecture. However, the stricter timing and the required delays create more complexity for signal flow control.

Inaccuracies, detection, and calibration

In the above discussion, we have described the transformation produced by the ideal Givens rotation. In practice, a variety of errors are possible during device operation. First, errors may be induced by inaccuracies in the construction of the device. Second, errors may be induced by inaccurate control settings (the voltages applied to the interdigitated electrodes and to the phase shifters). Third, at the detection stage, inaccuracies may occur due to the shot noise in the optical signal and thermal noise in the electronic amplifiers following the detectors. The first two types of errors can be minimized by proper calibration before operation. In fact, the values of the needed rotation angles are "hidden variables": their values are not explicitly required if the devices are used in an adaptive mode such as that of the architectures described in the previous section.

An analysis of the device physics reveals several possible deviations from the desired rotation transformation matrix given in Equation 1. The bends in the waveguides may produce small losses. Any deviation from the design value in the angle between the crossing waveguides may induce crosstalk between the channels. Such losses result in an additional rotation, small losses, and phase shifts in the beams. Inaccuracies in the device parameters and small drifts in the operating frequencies induce further deviations in the grating strength as given by Equation 3. These produce a further additive component in the rotation angle. As a result, the actual transformation matrix before calibration is not given by Equation 1, but in phasor notation by

$$\exp(j\Gamma_0)\begin{bmatrix}\sigma_1\exp(j\Gamma_2) & 0\\ 0 & \sigma_2\end{bmatrix}\cdot$$
$$\begin{bmatrix}\cos\psi & \exp(j\zeta_a)\sin\psi\\ \exp(j\zeta_b)\sin\psi & \cos\psi\end{bmatrix}\cdot \qquad (4)$$
$$\begin{bmatrix}\sigma_3\exp(j\Gamma_1) & 0\\ 0 & \sigma_4\end{bmatrix}$$

where Γ_0 is the overall phase shift of the output with respect to the local oscillator (in the detection process), σ's are attenuation coefficients, Γ_1 and Γ_2 are each a tunable phase shift plus a phase shift due to inaccuracies, ζ_a and ζ_b are phase shifts inherent in the diffraction process,[9] and ψ is the grating strength parameter. The attenuations are typically very small, so the attenuation coefficients can be given as $\sigma_i = 1 - \gamma_i^2$ with $\gamma_i \ll 1$. The grating strength parameter is $\psi = KV_g$, where V_g is the voltage applied to the interdigitated electrodes and K is an effective gain. The controllable parameters are V_g and Γ_i. As shown below, the calibration procedure can eliminate most of the errors by a suitable choice of V_g and Γ_i. However, small inaccuracies may still exist. These persistent errors occur because the calibration procedure (in the same manner as the system operation) involves detection of the output.

The coherent detection in the present case is like that commonly used in coherent fiber-optic communication systems. A received signal amplitude is coherently added with a stronger signal from a local oscillator (a reference beam). The combined optical signal impinges on a photodiode. In homodyne detection, the signal and the reference beam have the same wavelength. Two photodiodes are used to establish a balanced detection scheme. One receives the sum signal and the other the difference signal. To obtain the sum and difference signals, the beams (typically with a 90 degree angle between them) intersect at a beamsplitter (each beam having a 45 degree angle of incidence). For amplitudes of the transmitted reference and signal beams measuring A_R and A_S respectively and a relative phase shift between the beams of Γ_0, the difference of the photo detector currents Δi is[11]

$$\Delta i = (2\eta q\lambda/\mathrm{hc})A_R A_S\cos\Gamma_0 \qquad (5)$$

where η is the photodetector quantum efficiency, q is the electronic charge, λ is the wavelength, h is Planck's constant, and c is the speed of light.

Inaccuracies may occur at the detection stage due to the quantization of the optical field. These inaccuracies are equivalent to the presence of noise usually called "shot noise," which is a manifestation of the discrete nature of photons.[11] Noise also occurs as thermal noise in the electronic amplifiers. Cooling the amplifiers minimizes the latter. Integration of the signal in time also decreases these uncertainties, but at the expense of slower system operation. The shot noise increases with local oscillator power level, so that if the reference amplitude is made sufficiently large, the thermal component becomes negligible. The shot-noise-limited signal-to-noise ratio (achievable, for instance, with a low capacitance p-i-n diode followed by a microwave field-effect transistor amplifier) is[11]

$$SNR = (2\eta\lambda/B\mathrm{hc})A_S{}^2\cos^2\Gamma_0 \qquad (6)$$

where B is the bandwidth of the detector. The overall effect is that a signal amplitude A (in phase with the reference beam) is detected as $A + \sigma\varepsilon$, where ε can be modeled as a standard normally distributed error with variance $\sigma^2 = 2[Bk T + (\eta\lambda/B\mathrm{hc})]/\tau$, where T is the absolute temperature, k is Boltzmann's constant, and τ is the integration time of the detector amplifier.

Calibration before operation eliminates most of these errors. This procedure uses relatively long integration times and slowly ramped voltages to eliminate noise effects. The calibration procedure has three steps. For the first two steps, the x_2 input amplitude (see Equation 4) is set to unity and the x_1 amplitude to zero. In this case, the detectable output signals $\hat{y}_1$ and $\hat{y}_2$ are

$$\hat{y}_1 = \sigma_1\sigma_4\cos(\Gamma_0 + \Gamma_2 + \zeta_a)\sin\psi$$
$$\text{and} \qquad (7)$$
$$\hat{y}_2 = \sigma_2\sigma_4\cos\Gamma_0\cos\psi \quad .$$

First, the grating voltage V_g is applied so that the detected signal $\hat{y}_1$ is zero. The phase shift Γ_0 is adjusted so that the detected signal $\hat{y}_2$ is maximum. The product $\sigma_2\sigma_4$ is measured. Second, using the same configuration, V_g is tuned so that $\hat{y}_2$ is zero, and Γ_2 is adjusted to minimize $\hat{y}_1$. The product $\sigma_1\sigma_4$ is measured. Third, keeping a fixed voltage applied to the grating, the first input beam is set to unity and the second input beam to zero. Now the detectable signals are

$$\hat{y}_1 = 0$$
$$\text{and} \qquad (8)$$
$$\hat{y}_2 = \sigma_2\sigma_3(\Gamma_1 + \zeta_b)$$

The phase shift Γ_1 is adjusted to maximize the detected signal $\hat{y}_2$. The system is now calibrated. A small error remains due to the residual individually unadjustable attenuation factors σ_i. Therefore, with system calibration, the rotation transformation matrix is

$$\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}\cos\psi & -\sin\psi\\ \sin\psi & \cos\psi\end{bmatrix}\begin{bmatrix}\sigma_3 & 0\\ 0 & \sigma_4\end{bmatrix} \qquad (9)$$

Tuning the gains in the detection amplifiers compensates for the effects of some of the products of σ_i. For one device, this procedure only gives two degrees of freedom. Fortunately, with proper fabrication these attenuations can be very small. (For example, Becker and Johnson recently reported a loss of 0.08 decibel per one degree bend.[12])

Processor accuracy

The accuracy of this type of matrix triangularization processor is an issue of fundamental importance. In this section, we evaluate accuracy in three stages. First, we examine the accuracy in one elementary rotation transformation. Second, we examine the process of nulling all but one element of a given N-dimensional vector. Third, we assess the accuracy of the solution of a set of linear equations using the pipelined matrix triangularization architecture in a hybrid configuration that employs backsubstitution.

Accuracy of the elementary rotation transformation. In operation, the integration time of the detector amplifier is not as large as during calibration. If a rotation is to null the element y_1 in Equation 1, the actual detection will terminate when the signal plus noise are zero ($\hat{y}_1 = 0$). The nonzero y_2 output also contains noise. The detected output $\hat{y}_2$ is described by

$$\begin{bmatrix}0\\ \hat{y}_2\end{bmatrix} = \begin{bmatrix}\cos\psi & -\sin\psi\\ \sin\psi & \cos\psi\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}e_1\\ e_2\end{bmatrix} \qquad (10)$$

The two detector noises, e_1 and e_2, are independent and identically distributed. The correct rotation angle ψ must satisfy $\tan\psi = x_1/x_2$. However, the actual rotation angle produced in the presence of the noise is $\psi' = \psi - \sin^{-1}[e_1/(x_1^2+x_2^2)^{1/2}]$, where it is assumed that the total signal power in the vector $\bar{x}$ far exceeds the noise power. Letting $r_{i,j}$ equal $(x_i^2+x_j^2)^{1/2}$ and using $1/\varrho$ for the ratio of the standard deviation σ of the independent identically distributed errors to $r_{1,2}$, we can write the actual angle of rotation as $\psi' = \psi - \sin^{-1}(\varepsilon_1/\varrho)$, where $e_1 = \sigma\varepsilon_1$, and ε_1 is modeled by a standard Gaussian distribution. The small noise assumption corresponds to $\varrho \gg 1$. As noted earlier, this rotation angle is not explicitly needed.

The detected output signal $\hat{y}_2$ is also a random variable given by $\hat{y}_2 = e_2 + (x_1^2+x_2^2-e_1^2)^{1/2} \approx \sigma\varepsilon_2 + r_{1,2}(1-\varepsilon_1^2/2\varrho^2)$. Using the same low noise approximation, the average value of the detected output is

$$<\hat{y}_2> = r_{1,2}(1-1/2\varrho^2) \qquad (11)$$

The average output $<\hat{y}_2>$ is the norm of the vector $[x_1, x_2]^T$ as computed optically. It contains bias. The normalized variance in this quantity is $\sigma^2(1 + 1/2\varrho^2)$, assuming a Gaussian distribution. Note that these error statistics are the same if the $\hat{y}_2$ is zeroed, and the norm of $\hat{y}_1$ is detected. The attenuations left after calibration are negligible compared to the detector-induced errors.

Accuracy of zeroing all of the elements but one in an *N*-dimensional vector. In the pipelined implementation shown in Figure 3, zeroing all of the elements but one in a column is equivalent to computing the norm of an N-dimensional vector. Detection of the output signal $\hat{b}_N$ (the norm of the Nth column vector) occurs after $N-1$ rotation steps. An error is introduced at detector t_1 as the a_{N-1} component of the vector is zeroed. The first resultant signal, $\hat{b}_{N,1}$ (a "partial" norm), at the upper output branch of this rotation device is then $\hat{b}_{N,1} = r_{N,N-1}(1-e_1^2/r_{N,N-1}^2)^{1/2}$. Since this signal does not need to be detected, there is no additional noise except for that induced by the nulling method. This relation can be rewritten $\hat{b}_{N,1}^2 = a_{N-1}^2 + a_N^2 - e_1^2$. Proceeding diagonally along the pipeline, an additional error e_2 is introduced, giving a second resultant signal of $\hat{b}_{N,2}^2 = a_{N-2}^2 + a_{N-1}^2 + a_N^2 - e_1^2 - e_2^2$. Iterating through $N-1$ steps and adding the final detection error e_N yields the detected norm $\hat{b}_N$ for the entire Nth column of

$$\hat{b}_N = b_N[1 - \sum_{i=1}^{N-1}(e_i/\varrho)^2]^{1/2} + e_N \qquad (12)$$

The quantity ϱ for this case is b_N/σ, where b_N is the exact norm of the vector. The average value is $<\hat{b}_N> = b_N[1 - (N-1)/2\sigma^2]$ with variance $\sigma^2[1+(N-1)/2\varrho^2]$. In the presence of noise, the process of zeroing all of the elements but one in a column vector of length N can be expressed as

$$\begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \hat{b}_N \end{bmatrix} = \begin{bmatrix} & & \\ & \mathbf{R}_N & \\ & & \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{bmatrix} \qquad (13)$$

where R_N is the overall orthogonal column transformation that has produced the detected output vector $[0, \ldots, 0, \hat{b}_N]^T$.

Accuracy of linear equation solving via triangularization. A nested series of the orthogonal column transformations described by Equation 13 is needed to analyze the accuracy in triangularizing a matrix. After $N-1$ sets of rotation devices, a lower triangular matrix of processed elements is produced together with independent and identically distributed error elements in the upper triangular portion of the matrix. This process can be represented by

$$\{\mathbf{R}_2 \cdots R_{N-1}R_N\}\mathbf{A} = \begin{bmatrix} \diagdown & \mathbf{E}_U \\ \mathbf{L} & \diagdown \end{bmatrix} \qquad (14)$$

where $\mathbf{L}$ represents the processed elements in the triangularized matrix and $\mathbf{E}_U$ represents the errors in the upper portion of the matrix. Detection of the processed elements in $\mathbf{L}$ adds another lower triangular matrix $\mathbf{E}_L$ of independent and identically distributed detection errors.

The solution of the system of equations $\bar{v} = \mathbf{A}\bar{u}$ proceeds by entering the columns of the $\mathbf{A}$ matrix followed by $\bar{v}$, the vector of constants. If the resulting matrix were perfectly triangular, the solution for $\bar{u}$ would follow easily by backsubstitution. However, only a perturbed version of the lower triangular matrix $\{\mathbf{R}_2 \ldots \mathbf{R}_{N-1}\mathbf{R}_N\}\mathbf{A}\bar{u}$ and the processed $\bar{v}$ elements, namely $\{\mathbf{R}_2 \ldots \mathbf{R}_{N-1}\mathbf{R}_N\}\bar{v}$, are available for detection. From these values, electronic backsubstitution can proceed to obtain a solution. The solution vector $\hat{u}$ is described by $(\mathbf{L}+\mathbf{E}_L)\hat{u} = \{\mathbf{R}_2 \ldots \mathbf{R}_{N-1}\mathbf{R}_N\}\bar{v} + \bar{e}_v$, where $\bar{e}_v$ and $\mathbf{E}_L$ are respectively a vector and an upper triangular matrix of independent and identically distributed errors introduced in the detection process. The computed solution vector $\hat{u}$ can be expressed to first-order in terms of the exact solution vector $\bar{u}$ as

$$\hat{u} = \bar{u} + \mathbf{L}^{-1}(\bar{e}_v + \mathbf{E}\bar{u}) \qquad (15)$$

where $\mathbf{E} = \mathbf{E}_U - \mathbf{E}_L$. For a single pipeline, the errors $\mathbf{E}_L$, $\mathbf{E}_U$, and $\mathbf{E}$ are all fixed by the initial triangularization process. Therefore, averaging over $\bar{e}$ reveals a bias. The ensemble average over numerous pipelines solving the same problem is zero with a corresponding covariance matrix of $\sigma^2(1+\|\bar{u}\|^2)\mathbf{A}^{-1}\mathbf{A}^{-T}$, where $\|\bar{u}\|$ is the norm of $\bar{u}$ and $\mathbf{L}^{-1}\mathbf{L}^{-T}$ is approximated by $\mathbf{A}^{-1}\mathbf{A}^{-T}$. It is possible to eliminate this bias by iterating alternately between an optical pipeline and an electronic processor (performing only additions and multiplications). Thus a processing system solving the same problem repetitively converges to a statistical steady state whose average is the exact solution $\bar{u}$.

The arrays of Givens rotation devices that we have described utilize several of the favorable attributes of optics. They utilize parallelism (in one dimension) by allowing a simultaneous input into arrays of devices. They utilize interconnectivity by the intersection and modification of the channels of data. Finally, they utilize the special functions capability of optics by evaluating sine and cosine directly (without the use of a sequential algorithm).

Beyond solving sets of linear equations, matrix triangularization can be used to implement various square-root algorithms (in Kalman filtering and solving the Lyapunov and Riccati equations). Furthermore, a form of the lattice (or ladder) filter structure, described by the square-root-normalized lattice equations, has a natural interpretation in terms of rotations. These structures can also be implemented with arrays of other types of integrated optical devices.[13] Lattice filters are widely used for prediction and filtering in the areas of speech processing, channel equalization, seismic data interpretation, and electroencephalogram (EEG) analysis. This range of possible uses suggests that arrays of integrated optical Givens rotation devices may have many applications in the future, not only in solving sets of linear equations, but also in performing other critical signal-processing functions. □

Acknowledgments

This work was supported in part by a grant from the Joint Services Electronics Program under contract no. DAAG29-84-K-0024, in part by a grant from the Innovative Sciences Office of the Strategic Defense Initiative administered through the Office of Naval Research under contract no. N00014-86-K-0626, and in part by a grant from the US Air Force under contract no. F08635-84-C-0273.

References

1. W. Givens, "Computation of Plane Unitary Rotations Transforming a General Matrix to Triangular Form," *SIAM J. Applied Math.*, Mar. 1958, pp. 26-50.
2. J.W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, San Francisco, 1968.
3. J.E. Volder, "The CORDIC Trigonometric Computing Technique," *IRE Trans. Electronic Computation*, Sept. 1959, pp. 330-334.
4. M.M. Mirsalehi, T.K. Gaylord, and E.I. Verriest, "Integrated-Optical Givens Rotation Device," *Applied Optics*, May 15, 1986, pp. 1608-1614.
5. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell System Tech. J.*, Nov. 1969, pp. 2909-2947.
6. C.C. Guest, M.M. Mirsalehi, and T.K. Gaylord, "EXCLUSIVE OR Processing (Binary Image Subtraction) Using Thick Fourier Holograms," *Applied Optics*, Oct. 1, 1984, pp. 3444-3454.
7. C.M. Verber et al., "Large-Angle Optical Switching in Waveguides in Lithium Niobate," *Ferroelectrics*, May 1976, pp. 253-256.
8. E.M. Philipp-Rutz, R. Linares, and M. Fakuda, "Electro-optic Bragg Diffraction Switches in Low Cross-Talk Integrated-Optics Switching Matrix," *Applied Optics*, June 15, 1982, pp. 2189-2194.
9. E.N. Glytsis and T.K. Gaylord, "Rigorous Three-Dimensional Coupled-Wave Diffraction and Analysis of Single and Cascaded Anisotropic Gratings," *J. Optical Soc. Amer. A*, Nov. 1987, pp. 2061-2080.
10. B. Glance, "Performance of Homodyne Detection of Binary PSK Optical Signals," *J. Lightwave Technology*, Feb. 1986, pp. 228-235.
11. D. Marcuse, *Principles of Quantum Electronics*, Academic Press, New York, 1980.
12. R.A. Becker and L.M. Johnson, "Low-Loss Multiple-Branching Circuit in Ti-indiffused LiNbo$_3$ Channel Waveguides," *Optics Letters*, June 1984, pp. 246-248.
13. B. Moslehi et al., "Fiber-Optic Lattice Signal Processing," *Proc. IEEE*, July 1984, pp. 909-930.

Thomas K. Gaylord is Julius Brown Chair Regents' Professor of Electrical Engineering at Georgia Institute of Technology. He is the author of some 130 technical journal publications in the areas of optical data processing, electro-optics, grating diffraction, integrated optics, and semiconductors. He is the recipient of six Sigma Xi research awards, two teaching awards, the ASEE "Outstanding Contribution in Research" medal, and an honorary professional degree from the University of Missouri-Rolla. He is a Fellow of the Optical Society of America and of the IEEE.

Gaylord received the BS degree in physics and the MS degree in electrical engineering from the University of Missouri-Rolla, and the PhD degree in electrical engineering from Rice University.

Erik I. Verriest is an associate professor in the School of Electrical Engineering at the Georgia Institute of Technology. His research interests include mathematical system theory, stochastics, and optical computing.

Verriest received the degree of "Burgerlijk Electrotechnisch Ingenieur" from the State University of Ghent, Belgium, in 1973, and the MS and PhD degrees from Stanford University in 1975 and 1980, all in electrical engineering. He is a Francqui fellow (Belgian-American Educational Foundation).

Readers may write to the authors at the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

Progress Report
Optimal Sensor Placement Problem

W. Steven Gray
School of Electrical Engineering
Georgia Institute of Technology

February 1, 1988

1 Introduction

The purpose of this report is to summarize my research progress on the optimal sensor placement problem at the Georgia Institute of Technology during 1987 and to outline the future research to be conducted in this area during 1988.

The specific research topic being investigated is the optimal placement of sensors or transducers in physical systems which are best modelled by a set of partial differential equations. The most common example of such a system cited in the engineering literature is the process of heat conduction in a one-dimensional metal beam [2,3,6,7,8,11]. This example is popular because of the relatively simple dynamics of the heat equation and the utility of the results in industry.

Thermal conduction is often modelled by a spatially distributed system with a forcing function on one end representing the thermal energy input from the furnace and a disturbance input representing the beam's thermal interaction with the environment. The optimal sensor placement problem herein is to select a finite number of locations along the length of the beam where temperature sensors are to be placed so that the temperature profile of the beam can be accurately measured and controlled. The major difficultly in the practical problem is that both the behavior of the furnace and the environment are uncertain. Thus, the system must be modelled stochastically.

2 Research Progress for 1987

To put my research efforts in the proper perspective it helps to survey the history of the optimal sensor placement problem. The problem was first studied in the mid-1960's with relatively little success [10]. Among the major obstacles were the lack of addressing the observability question, the need for solving partial differential equations in real-time, and the inability to establish error bounds on the resulting solutions [2].

With the success of the Kalman filtering method in the early 1970's, a vast number of optimal sensor placement algorithms began to appear based on approximating the true solution to the state equation by various eigenfunction expansions. The idea was then to place the system's sensors by minimizing the error between the approximation (whose parameters were determined through measurement) and the true solution (see for example [1,2,11]). Such an approach had a major weakness in that an eigenfunction approximation of the solution could only be justified for a limited class of systems. Furthermore, it had not been demonstrated rigorously that such an optimality problem had a theoretical solution. The solution existence problem was later addressed by Omatu et al. [8] where the sufficient condition derived gave the intuitive meaning that *"the optimal sensor location should be allocated at the points where the maximum value of the amplitude of the state with respect to the spatial coordinate is attained..."*.

Over the past ten years some reseachers have abandoned the optimal filtering approach to the sensor placement problem and have introduced new mathematical tools with which to attack the problem. For example, Nakamori et-al. [7] have extended some concepts from information theory which give a new perspective on the problem, as have the function-theoretic methods of Jai and Pritchard [5]. My research concerning this problem can also be classified in the 'new perspective' catagory. In the paragraphs that follow I will summarize the results of my efforts.

Part of my research work has been in the problem formulation stage. It is not apparent to me that the *general* optimal sensor placement problem is well posed. Most researchers circumvent this fault by introducing various restrictions on the problem which in the final analysis are either over-restrictive or artificial. I have taken the perspective that the sensor placement problem can be divided into two parts.

The first part of the sensor placement problem is deterministic in nature and deals with whether or not a physical system with some specific arrangement of sensors is simply observable. That is, if all the parameters of the system are known exactly and the measurements are perfect, can we determine the state of the system at some point in the past given enough data? Although such a question has been addressed to some degree in the literature (see for example [11]), the picture is not complete. A related question is whether some observable sensor configurations have certain optimality properties not universal to the whole class of observable configurations. My research to date indicates that this is the case, especially in view of the second part of the sensor problem: the introduction of plant and measurement uncertainty. In the stochastic framework one talks about optimal sensor location in the sense of minimizing some type of error in the state estimation. My investigation indicates that one can use the results in the deterministic problem to get insight into the stochastic version.

Another interesting aspect of the sensor problem that I have pursued with some success is linking it to the seemingly unrelated area of design sensitivity and robustness. Preliminary indications are that the optimal sensor problem, at least a lumped system approximation of it, can be cast into a geometric framework that has been used successfully to study the parametric sensitivity of state space realizations [4,9]. It is believed that such a link could lead to results complementary to both areas of study.

3 Future Research

Much of my future research work on the optimal sensor location problem will be focused on three subprojects. First, I will continue to investigate the link between optimal sensor placement and the geometric approach to robust design. Second, I will derive a better understanding of how to best use the theoretical results of my research in the practical design problem (i.e. design algorithms). Finally, I plan to explore the relationship between my approach to the sensor problem and those of other investigators. Success on any of these three projects will lead to an execellent research paper.

4 Conclusions

In this report I have briefly presented my research progress on the optimal sensor location problem during 1987. The goal was to present the results in a form appropriate for the non-specialist in the area and to provide a historical context in which to better understand the problem and the research results. Furthermore, future research work was briefly outlined.

References

[1] Aidarous, S.E., Gevers, M.R., and Installe, M.J., "Optimal Sensors' Allocation Strategies for a Class of Stochastic Distributed Systems, "*Inter. Journal of Control*, Vol. 22, No.2, pp. 197-213, 1975.

[2] Cannon, J.R. and Klein, R.E., "Optimal Selection of Measurement Locations in a Conductor for Approximate Determination of Temperature Distributions," *Proc. Joint Automatic Control Conference*, 1970, pp. 750-756.

[3] Curtain, R.F., "Finite-Dimensional Compensator Design for Parabolic Distributed Systems with Point Sensors and Boundary Input, "*IEEE Trans. Automatic Control*, AC-27, No.1, pp. 98- 104, February 1982.

[4] Gray, W.S. and Verriest, E.I., "Optimality Properties of Balanced Realizations: Minimum Sensitivity," *Proc. IEEE Conference on Decision and Control*, 1987, pp. 125-128.

[5] Jai, A.E. and Pritchard, A.J., "Sensors and Actuators in Distributed Systems, "*Inter. Journal of Control*, Vol.46, No.4, pp. 1139-1153, 1987.

[6] Kumar, S., and Seinfeld, J.H., "Optimal Location of Measurements for Distributed Parameter Estimation," *IEEE Trans. Automatic Control*, AC-23, No.4,pp. 690-698, August 1978.

[7] Nakamori, Y., Miyamoto, S., Ikeda, S., and Sawaragi, Y., "Measurement Optimization with Sensitivity Criteria for Distributed Parameter Systems, "*IEEE Trans. Automatic Control*, AC-25, No. 5, pp. 889-900, October 1980.

[8] Omatu, S., Koide, S., and Soeda, T., "Optimal Sensor Location Problem for a Linear Distributed Parameter System, "*IEEE Trans. Automatic Control*, AC-23, No.4, pp. 665-673, August 1978.

[9] Verriest, E.I. and Gray, W.S., "Robust Design Problems: A Geometric Approach," *Proc. Inter. Symp. on the Mathematical Theory of Networks and Systems*, 1987 (to appear).

[10] Wang, P.K.C., "Control of Distributed Parameter Systems, "*Advances in Control Systems*, ed. C. T. Leondes, Academic Press, pp. 75-172, 1964.

[11] Yu, T.K. and Seinfeld, J.H., "Observability and Optimal Measurement Location in Linear Distributed Parameter Systems", *Inter. Journal of Control*, Vol.18, No.4, pp. 785-799, 1973.

Biography

W. Steven Gray was born in Muncie, Indiana on February 22, 1961. He received the B.S. degree in electrical engineering from Purdue University in 1983 and the M.S. degree in electrical engineering from the Georgia Institute of Technology in 1985.

While an undergraduate he worked summers on VLSI design software for IBM in East Fishkill, New York. After graduation he took the position of design engineer at Harris Semiconductor in Melbourne, Florida and worked primarily on integrated circuit design. In the fall of 1984 he began graduate studies at Georgia Tech in the areas of linear system theory, multivariable systems, and stochastic modelling and control. He is currently working toward the Ph.D. degree in electrical engineering.

E-21-676

FINAL REPORT

Grant AFOSR-87-0308

August 1, 1987 - March 31, 1989

Estimation and Control of Nonlinear and Hybrid Systems with Applications to Air-to-Air Guidance

by

A. H. Haddad
E. I. Verriest

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

Prepared for:

Air Force Office of Scientific Research
Major James M. Crowley/NM
Bolling AFB, Washington DC 20332

SUMMARY

This is the final report of Grant AFOSR-87-0308 with the Air Force Office of Scientific Research, which is also the continuation of Contract FO8635-84-C-0273 with the U.S. Air Force Armament Laboratory at Eglin Air Force Base. The work was performed at the School of Electrical Engineering at the Georgia Institute of Technology. The work proposed under this Grant is continuing under a new Grant to Northwestern University jointly with Georgia Tech.

The research covered several aspects of the basic issues that are needed to develop and implement nonlinear and hybrid systems schemes for the filtering, tracking, and control of maneuvering vehicles in an uncertain and nonlinear geometry. It is based on the approximation of the original nonlinear problem by a switched Markov linear models which in turn lead to hybrid model formulation or to piecewise linear approximations. Four aspects are considered: 1. Approaches to handling hybrid systems models; 2. Fast and slow decomposition for piecewise linear systems; 3. Estimation in the presence of impulsive inputs that can serve as either models for the switching behavior or the changes in maneuvers; 4. Modeling, parameterization, and realization issues for hybrid systems. Applications to nonlinear filtering and tracking schemes and their implementation is also addressed.

The research was performed under the direction of A. H. Haddad (now with Northwestern University) and E. I. Verriest. Dr. James M. Crowley of AFOSR and Dr. James Cloutier of AFATL monitored the Grant's progress.

Five Ph. D. students were supported under the Grant: B. S. Heck, J. Ezzine, M. A. Ingram, P. West, and S. Gray. The first three have completed their dissertation in August 1988, May 1989, and August 1989, respectively.

SECTION I

INTRODUCTION

The objective of this research was to develop nonlinear filtering and tracking algorithms for systems subject to complex geometries and uncertainties. These attributes characterize the air-to-air engagement scenario. The approach was based on the approximation of the original nonlinear stochastic model with a piecewise linear model. Then the resulting model was further approximated by a switched Markov linear model. The result is a dynamic system of the form

$$\dot{X}(t) = A[r(t)]\, X(t) + B[(r(t)]\, U(t) \tag{1a}$$

$$Y(t) = C[r(t)]\, X(t) + V(t) \tag{1b}$$

where the sate vector is X(t), the observation vector is Y(t), U(t) can serve as the control vector when the control problem is considered, or can serve as the process noise model for the filtering problem, V(t) is the observation noise vector. The noise processes are assumed to be white and Gaussian. The process r(t) is called either as the form index, the switching process, or the macro-state process, and is assumed to be a finite state Markov process taking the values in {1,2,...,N}. The approximation is via what is known as either switched Markov models or hybrid systems. The linear system in such a case switches among the forms (A[i],B[i],C[i]) according to the value of r(t), i.e., when the macro-state is equal to i.

In earlier reports the validity of the approximation has been analyzed as discussed in reference 1, and its applications to nonlinear filtering have been investigated as provided in References 2 and 3. This report addresses several aspects of the resulting approximate model and general approaches to its estimation, realization, and control. The main report is subdivided into four major sections. Section II addresses the general properties of hybrid systems from the point of view of control and stabilization. Section III addresses fast and slow decomposition of the original piecewise linear approximation with the view of simplifying the resulting algorithms. Section IV addresses an alternative model for the jumps representing the maneuvers and develops approximate nonlinear filtering algorithms for these models. Section V discusses several issues resulting from the realization of such systems as they affect sensitivity, robustness, and identification. The body of each section will be relatively short, as the results are provided in appropriate appendices.

SECTION II

HYBRID SYSTEM MODELS

Since the approximation to the original nonlinear model is represented by the hybrid model (1), a major part of the study dealt with control and stabilization properties of hybrid systems. General properties of controllability and observability of such models are given in Reference 4 and provided also in Appendix A. These properties carry over from the linear time-invariant case and stress the simplification of the algorithms used for controllability, observability, and stability. Usually, these system models switch among several realization. An important issue to consider is the ability to represent such models by an average model. Such an average model may be suitable under certain condition, or under cases where the switches may be fast. The use of such averaging methods can simplify the resulting control and filtering algorithm. Several averaging procedures for the stabilization of hybrid systems are reported in Reference 5 and provided in Appendix B. Two properties of the average system are investigated in References 6 and 7, and are given in Appendix C and D. The first considers the error that results from averaging and how to determine the validity of the use of the average model. The second considers the minimality properties of the average systems that would allow the stabilization of the original system by using the average model. The main advantage for using average models is that there is no need to identify the macro-state and the resulting algorithms are rather simple. Of course, the average model can replace the original system only under restricted conditions. The last aspect of hybrid systems considered in this problem is concerned with eigenvalue assignment for hybrid system models, which in this case deals with the Lyapunov exponents. The result is given in Reference 8 and Appendix E. The largest Lyapunov exponent determine the stability of such systems, and its assignment using control gains determines the ability to stabilize such systems.

Other aspects of hybrid systems dealing with realization and its relationship to implementation and filtering is provided in Section V.

SECTION III

FAST AND SLOW DECOMPOSITION

The approximation used to model the nonlinear systems exhibited fast and slow behavior both in the switching process and in each individual realization. Such fast ad slow behavior can lead to simplification of the resulting algorithms due to two-time scales decomposition and to reduced order of the filters-controllers via aggregation. The theory of singular perturbation which has been developed to deal with such behavior has been restricted to smooth systems. In our case the switches and the piecewise linear models lead to difficulties that require an extension to the standard linear theory. This section treats the singular perturbation theory for non-smooth systems with fast and slow modes. In particular it extends the theory developed in Reference 9 for quantized systems to general piecewise linear models. The piecewise linear models is considered in Reference 10 and provided also in Appendix F. Usually sliding modes occur in such models in both the fast and slow dynamics. Reference 11 (also given in Appendix G) discusses the conditions for the existence of the sliding modes and how the algorithm can handle the resulting complications. Two additional extensions of the theory are given in References 12 and 13, which are also provided in Appendix H and I respectively. The first extends the theory to the case of stochastic input as most of our models are subject to random inputs. The second extends the quantized system to the vector quantization case. The quantization problem is of interest in this case due to the fact that with the high order of the filter used in the original filtering problem, it is appropriate to use only a few quantization levels to reduce the computational complexity of the problem. In earlier reports the quantization aspect was covered by an approximate stochastic differential equation model with state dependent noise.

SECTION IV

FILTERS FOR POISSON DRIVEN MODELS

This section considers an alternative approach to the modeling of the switching jumps that affects the systems. In particular it considers a self-excited Poisson model as an input to the system. These self-excited inputs may represent varying maneuvers and or control actions that affects the target. It is well-known that the linear filter for such models is not optimal. It is difficult to derive such a linear filter for the case where the average of the input jumps is not zero. The study first considered several alternatives as suboptimal nonlinear detection-estimation schemes to solve the problem. These are summarized in References 14 and 15 and provided in Appendix J and K. The properties of the model and the derivation of the appropriate linear filters for such models are considered in References 16 and 17 and provided in Appendix L and M. Simulation results and the derivation of the error properties of the resulting approximate filters are still being investigated.

SECTION V

REALIZATION, ROBUSTNESS, AND SENSITIVITY

This section addresses several aspects of hybrid systems modeling with particular emphasis to realization and robustness as they affect the accuracy and sensitivity of the implementation used for the filter.

Work on optimal realizations of such systems progressed in two directions: earlier results showing the optimality of the balanced realizations (see Reference 18), in the discrete time case have been extended to the continuous time case and is given in Reference 19 and Appendix N. The results have also been extended to multi-mode systems (see Reference 20 and Appendix O), and general time-varying systems as given in Reference 21 and Appendix P. In these references applications to filtering have been analyzed, and Reference 20 also addressed the optimal implementations of the suboptimal nonlinear filters for the switched Markov models. Optimality conditions for non-infinitesimal perturbations have been given as well. Conditions for optimality over finite sets have been applied to the parameterization of 3-D rotations in Reference 22 and Appendix Q.

Realization problems for hybrid systems (reachability and observability) for generalized systems have been solved. More specifically, results for N-periodic systems have been reported in References 23 and 24 and are provided in Appendix R and S.

In addition to the realization point of view the sensitivity of analog algorithms were investigated from a parameter sensitivity point of view. In particular, a discussion of optical analog computing devices, for matrix computations was presented in Reference 25, and a wider collection of devices were analyzed in Reference 26.

SECTION VI

CONCLUSION

The research addressed several basic aspects of filtering, and control for nonlinear and hybrid models. These models may be used to approximate the nonlinear environment and other uncertainties in air-to-air engagement. Research is continuing on the integration of these approaches and in the implementation algorithms that could lead to a filter tat is applicable to a realistic system.

REFERENCES

[1] E. I. Verriest and A. H. Haddad, "Linear Markov Approximations of Piecewise Linear Stochastic Systems", Stochastic Analysis and Applications, vol. 5, pp. 213-244, 1987.

[2] E. I. Verriest and A. H. Haddad, "Approximate Nonlinear Filters for Piecewise Linear Models," Proc. Conf. on Information Sciences and Systems, Princeton University, pp. 526-529, March 1986.

[3] A. H. Haddad, E. I. Verriest, and P. D. West, "Approximate Nonlinear Filtering for Piecewise Linear Systems," NATO/AGARD Guidance and Control Panel's 44th Symposium, Athens, Greece, 5-8 May 1987.

[4] J. Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems", Proc. 1988 American Control Conference, Atlanta, pp. 41-46, June 1988. To appear also in International J. of Control.

[5] J. Ezzine and A. H. Haddad, "On the Stabilization of Two-Form Hybrid Systems via Averaging", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 579-584, March 1988.

[6] J. Ezzine and A. H. Haddad, "Error Bounds in the Averaging of Hybrid Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 1787-1791, Dec. 1988. To appear also in IEEE Transactions on Automatic Control.

[7] J. Ezzine and A. H. Haddad, "On the Minimality of the Average of Hybrid Systems", Proc. IEEE Conference on Control and Applications, Jerusalem, Israel, pp. RA-6-2/1-4, April 1989.

[8] J. Ezzine and A. H. Haddad, "On Largest Lyapunov Exponent Assignment and Almost Sure Stabilization of Hybrid Systems", Proc. 1989 American Control Conference, Pittsburgh, PA, pp. 805-810, June 1989.

[9] B. S. Heck and A. H. Haddad, "On Linear Singularly Perturbed Systems with Quantized Control," Proc. Annual Conference on Information Sciences and Systems, Johns Hopkins University, pp. 24-29, March 1987. Also in Automatica, vol. 24, pp. 755-764, Nov. 1988.

[10] B. S. Heck and A. H. Haddad, "Singular Perturbation in Piecewise Linear Systems", Proc. 1988 American Control Conference, Atlanta, pp. 1722-1727, June 1988. Also in IEEE Transactions on Automatic Control, vol. 34, pp. 87-90, January 1989.

[11] B. S. Heck and A. H. Haddad, "Extensions of Singular Perturbation Analysis in Piecewise Linear Systems", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 958-963, March 1988.

[12] B. S. Heck and A. H. Haddad, "Singular Perturbation Theory for Piecewise Linear Systems with Random Inputs," Stochastic Analysis and Applications, to appear.

[13] B. S. Heck and A. H. Haddad, "Singular Perturbation Analysis for Linear Systems with Vector Quantized Control", Proc. 1989 American Control Conference, Pittsburgh, PA, pp. 2178-2183, June 1989.

[14] M. A. Ingram and A. H. Haddad, "Optimal and Suboptimal Filtering for Linear Systems Driven by Self-Excited Poisson Processes", Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, pp. 426-435, October 1987.

[15] M. A. Ingram and A. H. Haddad, "A Sequential Detection Approach to State Estimation of Linear Systems Driven by Self-Excited Point Processes", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 2334-2335, Dec. 1988.

[16] M. A. Ingram and A. H. Haddad, "On Linear Systems Driven by Self-Excited Point Processes", Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, pp. 937-938, October 1988.

[17] M. A. Ingram and A. H. Haddad, "A Linear System Driven by a Jump Process with a State-Dependent Rate: Properties and Linea Estimators", submitted for publications.

[18] W. S. Gray and E. I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, CA, pp. 124-128, Dec. 1987.

[19] E. I. Verriest and S. W. Gray, "Robust Design Problems: A Geometric Approach", in Linear Circuits, Systems and Signal Processing: Theory and Application, Byrnes, Martin and Saeks, eds., Elsevier 1988.

[20] E. I. Verriest and A. H. Haddad, "Filtering and Implementation for Air-to-Air Target Tracking", Proc. 1988 American Control Conference, Atlanta, pp.143-148, June 1988.

[21] E. I. Verriest, "Minimum Sensitivity Implementations for Multi-Mode Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 2165-2170, Dec. 1988.

[22] E. I. Verriest, "On three-dimensional Rotations, Coordinate Frames, and Canonical Forms for It All", Proceedings of the IEEE, vol. 75, pp. 1376-1378, October 1988.

[23] E. I. Verriest, "Alternating Discrete Time Systems: Invariants, Parametrization and Realization", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 952-957, March 1988.

[24] E. I. Verriest, "The Operational Transfer Function and Parametrization of N-periodic Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 124-128,

Dec. 1988.

[25] T. K. Gaylord and E. I. Verriest, "Matrix Triangularization Using Arrays of Integrated Optical Givens Rotation Devices", Computer, pp. 59-66, Dec. 1987.

[26] E. I. Verriest, "Algorithms for Optical Computing and Their Sensitivity", to appear in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, North Holland, 1989.

APPENDIX A

J. Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems", Proc. 1988 American Control Conference, Atlanta, pp. 41-46, June 1988. To appear also in International J. of Control.

ON THE CONTROLLABILITY AND OBSERVABILITY OF HYBRID SYSTEMS[1]

Jelel Ezzine and A. H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

ABSTRACT

This paper considers a special class of hybrid systems, whose state space is a cross-product space of an Euclidean space and a finite-state space. Such models may be used to represent systems subject to known abrupt parameter variations, such as commutated networks. They may also be used to approximate some types of time-varying systems. The paper investigates controllability, observability, and stability of hybrid systems. In particular, it derives a necessary and sufficient algebraic condition, a simple algebraic criterion, and a computationally simple algebraic sufficient test for controllability and observability. Moreover, it provides a simple sufficient stability condition.

1. Introduction and Problem Formulation

This paper examines the controllability, observability and related issues of a special class of hybrid systems [1, 2]. The state space of a hybrid system is a cross-product space of an euclidien space and a finite-state space. Basically, hybrid systems are linear piece-wise constant time-varying systems, which are switching among a finite number of constant realizations. Systems of this type can be used to model synchronously switched linear systems [3], networks with periodically varying switches [4], and systems subject to failures [1]. Even though hybrid systems are time-varying they lend themselves to a precise and complete qualitative and quantitative analysis. Among such results we mention the possibility to explicitly compute their transition matrices, to derive and state necessary and sufficient conditions for their stability, and the possibility to derive an algebraic controllability/observability tests similar to the usual one for linear time-invariant systems. This is possible due to the many features hybrid systems share with time-invariant systems. Moreover, because they are time-varying, they offer many useful features due to their variable structure property. In other words, hybrid systems are a mixture of time-invariant systems with which they share the algebraic and geometric structures, and time-varying systems with which they share their variable structure property that will be useful in their control and stabilization.

The hybrid systems considered in this paper are assumed to have the form

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t) \tag{1.1}$$

$$y(t) = C(r(t))x(t) \tag{1.2}$$

where x is the system state vector of dimension n, u is the control input vector of dimension p, y is the output vector of dimension m, and r(t) is the "form index" which is a deterministic scalar sequence taking values in the finite index set $\mathbf{N}=\{1, 2, \ldots, N\}$.

This type of model can be used to represent systems subject to known abrupt parameter variations such as commutated networks or to approximate some types of time-varying systems. This is done by imposing a "deterministic" switching rule on the time behavior of the form index. However, in order to model unknown abrupt phenomena such as component and interconnection failures the form index can be modeled, for example, as a finite-state Markov chain.

The latter problem has received considerable attention within the control community, but many important generalizations remain to be worked out. Chizeck et al [1] denote such a control problem the Jump Linear Quadratic (JLQ) problem since they view it as an extention of the standard Linear Quadratic (LQ) problem. However, very little attention was given to the deterministic version of the problem, even though it shares many features with the JLQ problem. This paper is concerned with the deterministic version of the problem.

Let S_M denotes any sequence of length M of the values taken by r(t), and let δt_i denotes the time interval during which r(t) = i. Throughout the paper the following assumption is made, that S_N contains all the values that r(t) takes. In this case we define

$$T \equiv \sum_{i=1}^{N} \delta t_i \tag{2}$$

as the period of the system. If in addition the sequence in every S_N is the same the system is called a periodic hybrid system. It will be obvious that the assumption that $M \geq N$ in S_M will not affect the results. Hence, the assumption that M = N will be made to simplify the notations.

[1] This research is supported by the U.S. Air Force under contract F08635-84-C-0273 (with the Armament Laboratory) and grant AFOSR-87-0308.

The system takes the realization $\Sigma_i=(A_i,B_i,C_i)$ when $r(t) = i$, with $i\epsilon N$. This realization is called the ith form.

The following is an outline of the paper. Section 2 discusses the stability of hybrid systems where a simple sufficient stability criterion is derived. The observability and controllability of periodic hybrid systems are treated in Sections 3 and 4, respectively. Algebraic observability and controllability tests are obtained. Section 5 extends the results of Sections 3 and 4 to general hybrid systems. In section 6 the stabilizability of hybrid systems is addressed and a simple application is used for illustration purposes. Section 7 concludes the paper.

2. Stability

Even though hybrid systems are time-varying systems it is possible to obtain necessary and sufficient asymptotic stability conditions. We start by studying the stability of periodic hybrid systems. To this end we recall a theorem by Willems [5] that provides a necessary and sufficient conditions under which piece-wise constant periodic systems are uniformly asymptotically stable. Basically, the theorem states that for the system to be asymptotically stable its transition matrix over one period of time has to be a contraction. This theorem can be obviously modified to derive a similar one for hybrid systems which are not necessarily periodic. However the resulting theorem will be difficult to use, since one has to compute $(N-1)N!$ products of exponential matrices and check their eigenvalues.

In order to derive simpler conditions to test for the stability of such systems, a different norm is defined, namely the logarithmic norm [6, 7]. The result is a simpler condition that is sufficient only.

Definition

The logarithmic norm of a matrix A associated with the matrix norm $\|.\|$ is defined by

$$\mu(A) = \lim_{h\to 0^+} (\|I + hA\| - 1)/h \tag{3}$$

The norm satisfies the following inequality

$$\|Exp(At)\| \le Exp(\mu(A)t). \tag{4}$$

This norm is now used to derive the stability condition.

Theorem 1

For the null solution of the hybrid system (1) to be uniformly asymptotically stable, it is sufficient to have

$$\sum_i \mu(A_i)p_i < 0,\ p_i \equiv \delta t_i/T = (t_i-t_{i-1})/T,\ i\epsilon N. \tag{5}$$

The proof is a simple application of the logarithmic norm to Willems' theorem. It is important to note that the above theorem is stated not only for periodic hybrid systems but it applies to the more genaral hybrid systems as defined above too.

3. Observability

Since hybrid systems are a special class of time-varying systems they display interesting properties relative to controllability and observability. It would be appropriate to define the latter properties while keeping in mind the fact that these systems are variable structure systems. We start with the observability criterion since it is simpler to prove. Consequently, the dual controllability criterion is stated by appealing to the duality principle.

Definition

A periodic hybrid system is said to be observable if there exists some finite $t_f \ge t_0+T$ such that the initial state $x(t_0)$ of the unforced system can be determined from the knowledge of $y(t)$ on $[t_0,t_f]$.

Using the above definition it is possible to state an algebraic necessary and sufficient observability criterion very similar to the usual algebraic test. Moreover this algebraic test is expressed as a function of the observability matrices of the different forms. This condition is a generalization of the well known algebraic observability test.

Theorem 2

A periodic N-form hybrid system is observable if and only if the observability matrix

$$\begin{bmatrix} O_1 \\ O_2Exp(A_1(\delta t_1)) \\ \cdot \\ \cdot \\ \cdot \\ O_NExp(A_{N-1}(\delta t_{N-1}))\ldots Exp(A_1(\delta t_1)) \end{bmatrix} \tag{6}$$

has full rank, where O_i is the observability matrix of the ith form, $i\epsilon N$.

Proof

Let us assume that the system is in its ith form at time $t\epsilon[t_i,t_{i+1}]$ then the output is given by the following expression

$$y(t) = C_iExp(A_i(t-t_i)) \prod_{j=i-1}^{1} Exp(A_j(\delta t_j))x(t_0). \tag{7}$$

We now take n-1 derivatives of $y(t)$ in (7) and arrange them in a column vector $Y_i(t) = [y\ y^{(1)}\ y^{(2)} \ldots y^{(n-1)}]'$ which may be expressed as

$$Y_i(t) = O_iExp(A_i(t-t_i)) \prod_{j=i-1}^{1} Exp(A_j(\delta t_j))x(t_0) \tag{8}$$

where O_i is the observability matrix of the ith form. If the same procedure is repeated for all $i\varepsilon\mathbf{N}$ and combined together the following equation results

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_N \end{bmatrix} = \begin{bmatrix} O_1 \mathrm{Exp}(A_1(t-t_0)) \\ O_2 \mathrm{Exp}(A_2(t-t_1))\mathrm{Exp}(A_1 \delta t_1) \\ \cdot \\ \cdot \\ \cdot \\ O_N \mathrm{Exp}(A_N(t-t_{N-1}))\ldots\mathrm{Exp}(A_1 \delta t_1) \end{bmatrix} x(t_0). \quad (9)$$

From this point on the proof is identical to a standard textbook [8; pp. 354].

4. Controllability

At this point the dual algebraic controllability test is introduced. First a dual definition for controllability is proposed and used along with the algebraic observability test to prove the result via the duality principle.

Definition

A hybrid system is said to be state-controllable if for any t_0 each state $x(t_0)$ can be transferred to any final state x_f after one period. Thus there exists a t_f, $t_0+T \leq t_f < \infty$ such that $x(t_f) = x_f$.

Before presenting the algebraic controllability criterion, the dual to the observability criterion given above, the usual controllability test for time-varying systems is used. This is done in order to display certain interesting properties of hybrid systems. If we compute the controllability grammian and use the fact that the system is piece-wise constant we obtain the following theorem.

Theorem 3

A periodic hybrid system of N forms is controllable if and only if

$$W(t_0,t_0+T) = \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} \Phi_i(t,t_0)B_iB_i'\Phi_i'(t,t_0)dt \quad (10)$$

has full rank.

Corollary

A periodic hybrid system is completely controllable if and only if it is controllable.

Proof

See Remark (2.18) in [9], then use the Theorem 3.

Befor proceeding any further, a necessary and sufficient condition for a periodic hybrid system to be uniformly completely controllable is stated. This result will be of importance when stabilizability of such systems is in question.

Theorem 4

A periodic hybrid system is uniformly completely controllable if and only if it is completely controllable.

Proof

If the periodic system is completely controllable, there must exist a finite time $s \geq T$ such that $W(0,s) \geq \varepsilon I > 0$. Therefore the result is proved by using Lemma 1 from Silverman et al [10] and Remark (2.18) in [9].

Having used the usual test we are ready to present an algebraic controllability test similar to the one used in linear time-invariant systems. The following criterion applies for periodic hybrid systems. A similar criterion for general hybrid systems will be introduced in a later section.

Theorem 5

A periodic hybrid system of N forms is controllable if and only if the controllability matrix

$$[C_N, \mathrm{Exp}(A_N(\delta t_N))C_{N-1}, \ldots,$$

$$\mathrm{Exp}(A_N(\delta t_{N-1}))\ldots\mathrm{Exp}(A_2(\delta t_2))C_1] \quad (11)$$

has full rank, where C_i is the usual controllability matrix of the ith form, $i\varepsilon\mathbf{N}$.

Proof

Using the principle of duality and the algebraic observability theorem presented above proves the theorem.

For computational purposes, it is better to rewrite the above controllability matrix as follows

$$[C_N, \mathrm{Exp}(A_N(\delta t_N))\{C_{N-1}, \ldots\{C_4,$$

$$\mathrm{Exp}(A_3(\delta_t 3))\{C_2, \mathrm{Exp}(A_2(\delta t_2))C_1\}]. \quad (12)$$

This way one does not have to compute all of the matrices needed to express (11) and compute its rank. That is the rank is checked sequentially and (12) is augmented appropriately until full rank is achieved. If full rank can not be achieved throughout this sequential test then the system is not controllable. The same observation applies to the observability criterion.

In addition to the above algebraic criteria for controllability and observability, two more tests are introduced. The first test is a simple and geometrically and computationally attractive necessary algebraic test. The second one is a simple algebraic sufficient condition.

Theorem 6

A necessary algebraic condition for a hybrid system to be controllable is

$$\mathrm{rank}[C_1, C_2, \ldots, C_N] \equiv \mathrm{rank}\, C = n. \quad (13)$$

Where C_i is the controllability matrix for the ith form, iϵ**N**.

Proof

We write the state of the system at time s, for $x(t_0) = 0$:

$$x(s) = \int_{t_0}^{s} \Phi(s,\tau)B(\tau)d\tau. \qquad (14)$$

We now use the fact that the system is piece-wise constant and the linearity property of the integral operator to obtain for $s = t_N$

$$x(t_N) = \mathrm{Exp}(A_N\delta t_N)\ldots\mathrm{Exp}(A_2\delta t_2)$$

$$\int_{t_0}^{t_1} \mathrm{Exp}(A_1(t_1-\tau))B_1u(\tau)d\tau + \ldots$$

$$+ \mathrm{Exp}(A_N\delta t_N) \int_{t_{N-2}}^{t_{N-1}} \mathrm{Exp}(A_{N-1}(t_{N-1}-\tau))B_{N-1}u(\tau)d\tau$$

$$+ \int_{t_{N-1}}^{t_N} \mathrm{Exp}(A_N(t_N-\tau))B_Nu(\tau)d\tau. \qquad (15)$$

After expanding the exponential matrices inside every integral, it is found that $x(t_N)$ is an element of the column range space of the controllability matrix $\hat{C}$ given in Theorem 5. Moreover, it is easy to see that

$$\mathrm{rank}\ \hat{C} \le \mathrm{rank}\ C \le n \qquad (16)$$

an inequality that dictates that full rankness of C is a necessary condition for our system to be controllable.

The above proof gives an alternate way to prove the necessity part in Theorem 5. It is also interesting to note that this latter test is independent of the Σ_i's order. This order independence would have been very beneficial, however it does not hold in the sufficiency part of the proof.

Now we state a theorem that gives a simple sufficient algebraic test. With the above simple necessary test this condition will provide an efficient algebraic method to test for the controllability/observability of hybrid systems. This theorem is adapted from a theorem given in [11].

Theorem 7

A sufficient condition for a periodic hybrid system to be controllable is

$$\mathrm{rank}[B_N,\ \mathrm{Exp}(A_N(\delta t_N))B_{N-1},\ \ldots,$$

$$\mathrm{Exp}(A_N(\delta t_{N-1}))\ldots\mathrm{Exp}(A_2(\delta t_2))B_1]$$

$$\equiv \mathrm{rank}\ \hat{C} = n. \qquad (17)$$

Proof

Since $\hat{C}$ has full rank then $\hat{C}\hat{C}' > 0$, i.e. it is positive definite. Also

$$\hat{C}(s_1,s_2,\ldots,s_N)\hat{C}'(s_1,s_2,\ldots,s_N) =$$

$$\sum_{k=1}^{N} \Phi(s_k,t_0)B_kB_k'\Phi'(s_k,t_0) \qquad (18)$$

where $s_k\epsilon[t_k,t_{k-1}]$. Then we have

$$W(t_0,t_N) = \int_{t_0}^{t_N} \Phi(s,t_0)B(s)B'(s)\Phi(s,t_0)ds$$

$$\ge \sum_{k=1}^{N} (\pm \int_{s_k}^{s_k\pm\sigma} \Phi(s,t_0)B(s)B'(s)\Phi(s,t_0)ds)$$

$$= \sigma\hat{C}(s_1,s_2,\ldots,s_N,t_0)\hat{C}'(s_1,s_2,\ldots,s_N,t_0)$$

$$+ o(\sigma), \qquad (19)$$

for σ sufficiently small. If we assume that $\hat{C}$ has full rank then for σ small enough (19) is positive definite. But then (19) implies that $W(t_n,t_0) > 0$ which proves the theorem.

5. Aperiodic Hybrid Systems

In this section we generalize the above results stated for periodic hybrid systems to more general aperiodic hybrid systems. Nevertheless, many of these results apply to general hybrid systems without modification. Therefore we will state only the most important results.

Theorem 8

A hybrid system is controllable if and only if Theorem 5 holds for all possible N! permutations of the form-index set **N**.

Theorem 9

A hybrid system is controllable if Theorem 7 holds for all possible N! permutations of the index-set **N**.

It is obvious that Theorem 6 applies for general hybrid systems too. Moreover, Theorem 6 may also be sufficient under very general conditions. A heuristic argument can be given as follows: Since any matrix exponential is a perturbation of the identity matrix it follows that multiplying any matrix with matrix exponentials will not change its range space dramatically. That is if, for example, C_1 and C_2 have algebraic complementery range spaces (i.e, range(C_1) is perpendicular to range(C_2)) then range(Exp(AT)C_1) will almost always remain an algebraic complement but not necessarily perpendicular to range(C_2). As a matter of fact, Mariton [2] states that he has proved that Theorem 6 is also a sufficient condition.

6. Stabilizability

This section presents some results concerning the control and stabilization of hybrid systems. These results use standard techniques to control and stabilize hybrid systems.

Ikeda et al [12] looked at the relation between controllability properties of the system and various degrees of stability of the closed loop system resulting from linear state variable feedback. Their results are as follows: For any initial time t_0, and any continuous and monotonically nondecreasing function $\delta(.,t_0)$ such that $\delta(t_0,t_0)=0$, the transition matrix $\hat{\Phi}(.,.)$ of the closed loop system can be made to satisfy

$$\|\hat{\Phi}(t,t_0)\| \leq a(t_0)\mathrm{Exp}\{-\delta(t,t_0)\} \text{ for all } t \geq t_0,$$

if and only if the system is completely controllable. Furthermore, in case of a bounded system, for any $m \leq 0$, a bounded feedback matrix can be found such that the transition matrix of the closed loop system is made to satisfy

$$\|\hat{\Phi}(t_2,t_1)\| \leq a\mathrm{Exp}\{-m(t_2-t_1)\} \text{ for all } t_1,\ t_2 \geq t_1,$$

if and only if the system is uniformly completely controllable. Thus, their results can be regarded, in some sense, as extensions of the well known results of closed loop pole assignment for time-invariant systems.

Therefore there is a high degree of flexibility in the stabilization of hybrid systems if they are either completely controllable or uniformly completely controllable.

As an illustration of the above result, a procedure is proposed to stabilize a periodic hybrid system via state feedback when all of the forms are minimal. This design procedure allows the designer to impose or choose an upper bound on the norm of the transition matrix of the hybrid system to be stabilized. Thus the norm of the transition matrix for hybrid systems plays a role similar to the maximum overshoot and time constants in linear time-invariant systems.

In order to impose an upper bound on the norm of the transition matrix a known stability criterion [5] is used: The null solution of (1) is uniformly asymptotically stable if and only if there exists two positive constant c_1 and c_2 such that

$$\|\Phi(t,t_0)\| \leq c_1\mathrm{Exp}(-c_2(t-t_0)) \qquad (20)$$

or all $t \geq 0$. Therefore the use of Theorem 1 leads to the following design criterion

$$\sum_i \mu(A_i)\delta t_i \leq k_1 - k_2T \qquad (21)$$

where $k_1 = \ln(c_1)$ and T is the period of the hybrid system. The k_i's, i=1, 2, are the design parameters that are chosen according to the specifications on the upper bound of the transition matrix of the closed loop system and consequently reflect the desired time response of the system. This is possible whenever (21) is achievable. Consequently, (21) can be always obtained via state feedback since every form is observable. It is important to note that this design procedure applies to both periodic and aperiodic hybrid systems. It should be noted that the minimality condition for every form is not necessary to achieve such a design.

7. Conclusion

This paper considered a special class of linear piece-wise constant time-varying systems. These systems are called hybrid systems because the set of linear time-invariant systems among which the systems are switching is finite. Their state space thus contains both continuous and discrete components.

Since hybrid systems share several features with linear time-invariant systems it was possible to derive the following results: A necessary and sufficient stability condition and a simple sufficient criterion. Algebraic necessary and sufficient controllability-observability tests similar to the usual time-invariant tests. An interesting necessary controllability-observability condition which may also be sufficient, along with a simple sufficient condition.

The necessary controllability/observability condition is a flat block matrix composed from the controllability/observability matrices of every form which makes it independent of the switching order. This order independence along with the fact that the condition is "almost" sufficient make it a very useful test. Therefore identifying the class of hybrid systems for which this condition is necessary and sufficient would be an interesting problem.

Additional work is needed concerning stability theory of this class of systems. The variable structure property seems to be a promising feature in this direction. In addition if one thinks of every system $\Sigma_i=(A_i,B_i,Ci)$ with $i \in \mathbf{N}$ as an operator acting on the state x during δt_i, and these operators are applied in a successive manner, then this process can be viewed as an iterative process [13]. Viewing a hybrid system as an iterative process sheds some light on some complicated issues such as the stability of such systems.

Finally adapting the results of this paper to hybrid systems where the switching is a stochastic process such as a Markov chain may be useful.

ACKNOWLEDGEMENT

We wish to thank Professor Erik I. Verriest for helpful discussions.

REFERENCES

[1] H. J. Chizeck, A. S. Willsky and D. Castanon, "Discrete-time Markovian jump Linear Quadratic Optimal Control," Int J. Control, Vol. 43, No. 1, pp. 213-231, 1986.

[2] M. Mariton, "Controllability, Stability and Pole Allocation for Jump Linear Systems;" Proc. 25th. IEEE Conf. Decision and Control, Athens, Greece, pp. 2193-2194, Dec. 1986.

[3] T. L. Johnson, "Synchronous Switching Linear Systems," Proc. 24th IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 1699-1700, Dec. 1985.

[4] R. W. Brockett and J. R. Wood, "Electrical networks containing controlled switches," in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, San Francisco, pp. 1-11, April 1974.

[5] J. L. Willems, Stability Theory of Dynamical Systems, Nelson, London, 1970.

[6] C. Van Loan, "The Sensitivity of the Matrix Exponential," SIAM J. Num. Anal., Vol. 14, No. 6, Dec. 1977.

[7] T. Strom, "On Logarithmic Norms," SIAM J. Num. Anal., Vol. 12, No. 5, Oct. 1975

[8] T. Kailath, Linear Systems, Prentice-Hall, 1980.

[9] R. E. Kalman, p. L. Falb and m. A. Arbib, Topics in Mathematical System Theory, McGraw-Hill, New York, 1969.

[10] L. M. Silverman and B. D. O. Anderson, "Controllability, Observability and Stability of Linear Systems," SIAM J. Cont., Vol.6, No.1, pp. 121-130, 1968.

[11] D. L. Russel, Mathematics of Finite-dimensional Control Systems, Theory and Design, Marcel Dekker, 1979.

[12] M. Ikeda, H. Medea, and S. Kodama, "Stabilization of Linear Systems," SIAM J. Cont., Vol.10, No.4, pp. 716-729, 1972.

[13] J. N. Tsitsiklis, "On the Stability of Asynchronous Iterative Processes," Proc. 25th IEEE Conf. Decision and Control, Athens, Greece, pp. 1617-1621, Dec. 1986.

[14] R. E. Kalman, Y. C. Ho and K. S. Narendra, "Controllability of Linear Dynamical Systems," in Contribution to Differential Equations, Vol.1, No.2, pp. 189-213, 1962.

APPENDIX B

J. Ezzine and A. H. Haddad, "On the Stabilization of Two-Form Hybrid Systems via Averaging", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 579-584, March 1988.

ON THE STABILIZATION OF TWO-FORM HYBRID SYSTEMS VIA AVERAGING[1]

Jelel Ezzine and A. H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA, 30332-0250

ABSTRACT

This paper discusses some practical methods of analysis and control of two-form hybrid systems. These systems are called hybrid because their state space contains both continuous and discrete components. These are systems that switch among a finite number of linear time-invariant realizations. Such models may be used to represent systems subject to known abrupt parameter variations such as commutated networks or to approximate some types of time-varying systems. This paper restricts the analysis to systems that switch among only two possible linear models.

1. Introduction and Problem Formulation

This paper discusses some practical methods of analysis and control of two-form hybrid systems. Hybrid systems denote a special class of piece-wise constant time-varying systems. The set of constant realizations among which the model is switching is finite and in this paper is restricted to two. Such systems can be used to model synchronously switched linear systems [1], networks with periodically varying switches [2], and systems subject to failures [3]. In particular we examine the stabilization and related issues of two-form hybrid systems via a special averaging technique. Averaging theory may be used in either deterministic or probabilistic contexts. In the probabilistic case averaging is introduced in a natural way by taking expected values. In the deterministic case, however, averaging is introduced via perturbation techniques. Averaging methods have received considerable attention. Brockett and Wood [2] used a deterministic averaging technique to analyse and stabilize a class of bilinear systems which are difficult to analyse or control otherwise. Geman [4] used probabilistic averaging techniques to study the stability of random differential equations. His main interest was to explore the relation between asymptotic stability in the average equation, and asymptotic stability in the random equation: Specifically, when does the first imply the second? Kosut et al [5] applied the theory of averaging to the analysis of the stability of adaptive systems.

Even though hybrid systems are time-varying they lend themselves to a precise and complete qualitative and quantitative analysis. Among such results we mention the possibility to explicitly compute their transition matrices, to derive and state necessary and sufficient conditions for their stability [6], and most interestingly the possibility to derive algebraic controllability and observability tests similar to the usual ones found in the theory of linear time-invariant systems [6]. This is possible due to the many features hybrid systems share with time-invariant systems. Moreover, because they are time-varying, they offer many useful features due to their variable structure property.

The hybrid systems under consideration are assumed to have the form

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t) \qquad (1.1)$$

$$y(t) = C(r(t))x(t) \qquad (1.2)$$

where x is the system state vector of dimension n, u is the control input vector of dimension p, y is the output vector of dimension m, and r(t) is the "form index" which is a deterministic scalar sequence taking values in the finite index set N={1, 2, ..., N}.

Such model may be used to represent systems subject to known abrupt parameter variations such as commutated networks or to approximate some types of time-varying systems [7]. The latter can be done by imposing a "deterministic" switching rule on the time behavior of the form index. However, to model unknown abrupt phenomena such as component and interconnection failures the form index can be modeled, for example, as a finite-state Markov chain (FSMC) [3].

The latter problem has received considerable attention within the control community but much work still remains to be done. Chizeck et al [3] denotes the optimal control problem of such systems the Jump Linear Quadratic (JLQ) problem since they view it as an extention of the standard Linear Quadratic (LQ) problem. However, very little attention was given to the stabilization and control of the deterministic version of the problem, even though it shares many features with the JLQ problem. This paper is concerned with the latter problem.

Let S_M denotes any sequence of length M of the values taken by r(t), and let δt_i denotes the time interval during which r(t) = i. Throughout the paper the following assumption is made, that S_N contains all the values that r(t) takes. In this case we define

$$T \triangleq \sum_{i=1}^{N} \delta t_i \qquad (2)$$

as the period of the system. If in addition the sequence in every S_N is the same the system is called a periodic hybrid system. It will be obvious that making the assumption that M≥N on S_M will not affect the results. The assumption that M = N simplifies the notations. Let the ith form denote the realization $\Sigma_i=(A_i,B_i,C_i)$ associated with the ith form index (i.e., r(t) = i), with i∈N. In this paper N = 2, so that we are concerned with Flip-Flop (F2) systems as a special class of hybrid

[1] This work is supported by the U.S. Air Force under grant AFOSR-87-0308.

systems. The F2-systems switch "back and forth" between two time-invariant systems (two forms), Σ_1, and Σ_2, of identical dimensions. The system spends T_i time-units at Σ_i, $i = 1, 2$. The only condition imposed on the switching is that the system can not spend more than $2T_i$ at Σ_i. That is, no order is imposed on the switching between the two forms.

The following is an outline of the paper. Section 2 starts by introducing the averaging technique based on a Lie algebraic formulation. It also adresses the perturbations induced by the averaging procedure, and offers, where possible, an alternative to averaging. Finally it discusses the controllability of the average system. The stability and stabilization issues of hybrid systems are treated in section 3. Section 4 concludes the paper and points to additional open problems.

2. The Averaging Technique

In this section we introduce a practical averaging technique which will be helpful in the analysis and control of two-form hybrid systems. It will be obvious from the following treatment that the proposed averaging method applies to multi-form hybrid systems as well. The main tools in simplifying the analysis and synthesis of stabilizing controls for such systems will be some basic ideas from linear systems theory combined with tools from Lie algebras, linear algebra, and stability theory of ordinary differential equations.

The averaging methodology to be used in this paper is based on a result from Lie algebras known as the Baker-Campbell-Hausdorff Formula [2]: Given two real matrices A and B there is no guarantee that there exists a real matrix C such that

$$\mathrm{Exp}(A)\mathrm{Exp}(B) = \mathrm{Exp}(C). \tag{3}$$

This will be the case, however, if $\|A\|+\|B\| \leq \ln(2)$ [8], and then C will be given by a <u>convergent</u> infinite expression

$$C = A + B + (1/12)[[A,B],B] + (1/12)[[B,A],A] + \ldots \tag{4}$$

where the symbol $[A,B] \equiv AB-BA$ is the commutator product. This expression is the Baker-Campbell-Hausdorff formula (BCH).

Similar expressions like the BCH formula are used in a large number of useful approximations in physics [9] and switched electrical networks [2]. In this section we show how the BCH formula and related expressions can be used to analyse and stabilize F2-systems and hybrid systems in general.

The concept is similar to the one used in [2] to stabilize bilinear systems. In our case we are interested in the stabilization of F2-systems whose A-matrix satisfies

$$A(t) = \begin{cases} A_1 \text{ for } 0 \leq t < T_1 \\ A_2 \text{ for } T_1 \leq t < T, \end{cases} \tag{5}$$

and it is desired to approximate the expression $\mathrm{Exp}(A_1T_1)\mathrm{Exp}(A_2(T-T_1))$. The BCH formula is used to yield Exp(C) where C is as in (4). Now it is desired to have an expression for C that is independent of order, to agree with the order independence introduced in the definition of F2-systems, then we must compute

$$\ln\tfrac{1}{2}\{\mathrm{Exp}(A)\mathrm{Exp}(B)+\mathrm{Exp}(B)\mathrm{Exp}(A)\}$$

$$= A+B+(1/12)[[A,B],B]+(1/12)[[B,A],A]$$

$$= A+B+(1/12)[[A,B],B-A]. \tag{6}$$

Therefore, we obtain a series of approximations for F2-systems. If the F2-system realization is $\Sigma_1 = (A_1,b_1)$ for T_1, and $\Sigma_2 = (A_2,b_2)$ for T_2, then the first approximation is

$$\Sigma = \{\alpha A_1+(1-\alpha)A_2,\ \alpha b_1+(1-\alpha)b_2\}, \tag{7}$$

with

$$\alpha \equiv T_1/(T_1 + T_2),$$

and the second approximation is

$$\Sigma = \{[\alpha A_1+(1-\alpha)A_2+(1/12)\alpha(1-\alpha)[[A_1,A_2],\ (1-\alpha)A_2-\alpha A_1],\ \alpha b_1+(1-\alpha)b_2\}. \tag{8}$$

Some comments about these two approximate expressions are in order. The first order approximation can be interpreted at least in two different ways. The first interpretation is a probabilistic one; it says that the average system can be viewed as the probabilistic average of the hybrid system with $P(\Sigma=\Sigma_1) = \alpha$ and $P(\Sigma=\Sigma_2) = 1-\alpha$. This is consistent with the frequency interpretation idea especially when we are interested in long time-range behavior of the system. The second interpretation comes from the theory of variable structure systems (VSS) and the Filippov's continuation technique. The latter technique was introduced to study the behavior of the system in chattering mode. The above first order approximation is nothing but a Filippov's average system. Therefore, a hybrid system can be viewed as a VSS system in chattering mode where the switching manifolds are solution orbits of the average system.

It was shown in [2], via an example, that in some special cases the second correction term is more important then the first, which, in fact, might vanish. Thus, the usefulness of the second approximation. However, in [2] there was no attempt to analyse the errors introduced by the BCH formula and the averaging method. Obviously, there are two very important issues in using such a formula and averages derived from it. The first one is the error introduced by only using few terms in the BCH expression while computing the average matrix. The second one is the difference between the actual system, in our case the F2-system, or in [2] the bilinear system, and the average system used to reflect the average behavior of the system under consideration. Both of these issues have to be addressed because of their paramount importance, especially the difference between the actual system and its average which is a function of the error introduced by truncating the BCH formula expression. Since the latter problems require a lengthy discussion only a summary of the results is given in this paper.

In what follows we present some results related to the accuracy of the usage of a truncated BCH formula. Using the BCH formula, one can obtain an approximation $\hat{C}$ of C, to any desired order. Consequently, C can be written as $C = \hat{C}+\tilde{C}$, where $\tilde{C}$ is the unknown error due to the approximation. Therefore, the induced error in computing Exp(C) by using the approximate matrix $\hat{C}$ is

$$E \equiv \mathrm{Exp}(CT)-\mathrm{Exp}(\hat{C}T). \qquad (9)$$

The first order approximation of E can be expressed in terms of the solution of a linear time-invariant matrix differential equation:

Proposition 1 [10]

Let E_1 denote the first order approximation in ε to E, then E_1 satisfies the following matrix differential equation

$$\dot{Y}(t) = \hat{C}Y + \varepsilon\tilde{C}_p\mathrm{Exp}(\hat{C}t), \quad Y(0) = 0. \qquad (10)$$

where $\tilde{C} \equiv \varepsilon\tilde{C}_p$, and ε is a positive scalar.

In order to compute upper bounds for E_1 and E the following results may be used.

Proposition 2

Assume that $\|\mathrm{Exp}(\hat{C}t)\| \leq M(t)\mathrm{Exp}(\beta(t)t)$, with $\beta(t)$ a scalar function, then

$$\|E_1\| \leq (\|\tilde{C}\|/2\|\hat{C}\|)M(t)\times$$

$$\mathrm{Exp}\{\beta(t)t\}(\mathrm{Exp}\{2\|\hat{C}\|t\}-1). \qquad (11)$$

The use of results in [10], and the assumption that M(t) is monotonic yields

$$\|E\| \leq t\|\tilde{C}\|M^2(t)\mathrm{Exp}\{(\beta(t) + M(t)\|\tilde{C}\|)t\}. \qquad (12)$$

Sometimes it is possible to avoid the computation of A-"average". That is, under certain conditions, it is possible, via state feedback, to make the F2-system time-invariant in A. That is given Σ_1 and Σ_2, and the appropriate conditions satisfied by the given forms, one can compute a feedback gain matrix

$$K = [K_1,K_2] \qquad (13)$$

which will make the A-matrices of both forms equal and cosequently render the A-matrix of the hybrid system, upon closing the loop via K_i for Σ_i, constant. Moreover, this constant A matrix is given by the following expression

$$A = A_1 - B_1K1 = A_2 - B2K_2. \qquad (14)$$

In this section necessary and sufficient conditions are derived for the existence of K and a compact computation recipe based on the Kronecker-product and the generelized-inverse techniques is given.

Theorem 1

Given the F2-system

$$\dot{x}(t) = A_ix + B_iu, \quad i = 1,2. \qquad (15)$$

such that

$$\mathrm{Range}[A_1-A_2,B_1,-B_2] = \mathrm{Range}[B_1,-B_2], \qquad (16)$$

then, there exists a minimum-norm $G \equiv [K_1,K_2]$ such that

$$A_1 - B_1K_1 = A_2 - B_2K_2. \qquad (17)$$

Moreover, the G matrix is given by

$$s(G) = \{[B_1,-B_2]^\dagger \otimes I_n\}s(A_1-A_2), \qquad (18)$$

where s(.) is the stacking operator.

Proof

Starting with the forms Σ_i, i=1, 2, the ith model is given by

$$\dot{x}(t) = A_ix + B_iu. \qquad (19)$$

If we define $A \equiv A_i-\delta A_i$ the above system can be written as follows

$$\dot{x}(t) = Ax + \delta A_ix + B_iu, \qquad (20)$$

and the next step is to compute a gain matrix K_i such that

$$\delta A_ix + B_iu = \delta A_ix - B_iK_ix =$$

$$(\delta A_i - B_iK_i)x = 0 \text{ for all } x. \qquad (21)$$

This is equivalent to

$$B_iK_i = \delta A_i, \qquad (22)$$

which is nothing but an algebraic equation for the unknown K_i. Therefore, for K_i to exist one needs the well known condition

$$\mathrm{rank}[B_i,\delta A_i] = \mathrm{rank}[B_i], \; i=1, 2 \qquad (23)$$

which is equivalent to the existence of some matrix K_i such that

$$\delta A_i = B_iK_i, \; i=1, 2. \qquad (24)$$

Substructing the first equation from the second yields

$$A_1 - A_2 = B_1K_1 - B_2K_2 \qquad (25)$$

which can be written as follows

$$[A_1 - A_2] = [B_1,-B_2][K_1',K_2']' \qquad (26)$$

which is itself an algebraic equation with the K_i, i=1,2, as unknown and the condition given in the theorem is the one needed for the existence of both gains. Equation (18) is nothing but a compact way to write such equations. As a matter of fact it is very useful when numerical techniques are used to solve the problem.

Corollary

If $\mathrm{Range}[A_1-A_2,B_1-B_2] = \mathrm{Range}[B_1-B_2]$, then

$$K \equiv K_1 = K_2, \qquad (27)$$

and the gain matrix is given by:

$$s(K) = \{[B_1-B_2]^\dagger \otimes I_n\}s(A_1-A_2). \qquad (28)$$

Proof

When $K_1 = K_2 = K$ is needed the proof of the above theorem is changed accordingly to yield the results stated in the theorem.

We now return to the average system. One of the key assumptions made to design the regulator via averaging is the controllability of the average system. This assumption is not unreasonable since the controllability property of linear time-invariant systems is generic. However, one can construct hybrid systems such that their averages are not controllable [11]. In [11] a sufficient condition that identifies a class of hybrid systems for which the hybrid system's controllability guarantees the controllability of the average system was given. The result is stated in the following theorem.

Theorem 2
The average system of a hybrid system is controllable if

a. Rank $[C_1, C_2, \ldots, C_N] = n$.
b. All forms are simultaneously diagonalizable.

3. Stability

Even though Hybrid systems are time-varying systems it is possible to obtain necessary and sufficient asymptotic stability conditions [6]. However, the latter condition is computationally time consuming and a simple sufficient asymptotic stability condition was presented to alleviate the computational burden. In this section the same sufficient condition is rederived using other means, which may be generalized.

In order to rederive this sufficient condition a brief introduction to the notion of logarithmic norm is given. The logarithmic norm (also known as the logarithmic derivative, the measure of a matrix) was introduced in 1958 separately by Dahlquist [12] and Lozinskij [13] as a tool to study the growth of solutions to ordinary differential equations and the error growth in discretization methods for their approximate solution. It is formally defined as follows:

Definition
The logarithmic norm associated with the matrix norm $\|\cdot\|$ is defined by

$$\mu(A) = \lim_{h\to 0^+} (\|I + hA\| - 1)/h. \qquad (29)$$

The explicit expression for the logarithmic norm associated with the Euclidian norm is

$$\mu(A) = \max\{\mu : \mu \in \lambda((A+A^*)/2)\}. \qquad (30)$$

Then the following inequality is true:

$$\|\mathrm{Exp}(At)\| \leq \mathrm{Exp}(\mu(A)t). \qquad (31)$$

Now we are ready to apply the logarithmic norm to derive a simple sufficient condition to test for the stability of hybrid systems.

Theorem [14]
Let $t \to A(t)$ be a regulated function from $[0, \infty)$ to C^{nxn}. Then the solution of

$$\dot{x}(t) = A(t)x(t) \qquad (32)$$

satisfies the inequalities

$$\|x(t_0)\| \mathrm{Exp}\{-\int_{t_0}^{t} \mu[-A(t')]dt'\}$$

$$\leq \|x(t)\| \leq \|x(t_0)\| \mathrm{Exp} \int_{t_0}^{t} \mu[A(t')]dt'. \qquad (33)$$

Basically the theorem states that the rate of change of the norm of the state vector $x(t)$ is bounded by $\mu(A(t))$ and to insure stability this bound must be negative.

Theorem 3
For the null solution of the hybrid system (1) to be uniformly asymptotically stable, it is sufficient to have

$$\sum_{i} \mu(A_i)p_i < 0, \quad p_i \equiv (t_i - t_{i-1})/T, \; i \in N. \qquad (34)$$

Before giving the proof of the theorem we introduce a different way to represent hybrid systems (1)-(2). This new formulation has the advantage of simplifying certain proofs. This new representation is as follows

$$\dot{x}(t) = \{\sum_{i=1}^{N} v_i(t)A_i\}x(t) + \{\sum_{i=1}^{N} v_i(t)B_i\}u(t) \qquad (35)$$

$$y(t) = \{\sum_{i=1}^{N} v_i(t)C_i\}x(t) \qquad (36)$$

where $v_i(t) = 1$ when the system is governed by the ith realization Σ_i, and $v_i(t) = 0$ otherwise. The $v_i(t)$ function is called the ith indicator function. It is evident from the definition of hybrid systems that at any point in time only one of the N indicator functions is one. Now we prove the theorem.

Proof
Using the above representation the homogeneous part of a hybrid system can be written as

$$\dot{x}(t) = \{\sum_{i=1}^{N} v_i(t)A_i\}x(t) \qquad (37)$$

Using Theorem 27 in [14] one can write

$$\|x(t)\| \leq \|x(t_0)\|\mathrm{Exp}\{\int_{t_0}^{t} \mu(A(s))ds\}$$

$$= \|x(t_0)\| \mathrm{Exp}\{\int_{t_0}^{t} \mu[\sum_{i=1}^{N} v_i(s)A_i]ds\} \qquad (38)$$

Using Theorem 5(e, d) in [14] yields

$$\|x(t)\| \leq \|x(t_0)\| \mathrm{Exp}\{\sum_{i=1}^{N} \int_{t_0}^{t} v_i(s)\mu[A_i]ds\}$$

$$= \|x(t_0)\| \mathrm{Exp}\{\sum_{i=1}^{N} \mu[A_i] \int_{t_0}^{t} v_i(s)ds\}$$

$$= \|x(t_0)\| \mathrm{Exp}\{(\sum_{i=1}^{N} \mu[A_i](1/t-t_0) \int_{t_0}^{t} v_i(t)ds)(t-t_0)\}$$

$$= \|x(t_0)\| \mathrm{Exp}\{(\sum_{i=1}^{N} p_i\mu[A_i])(t-t_0)\} \qquad (39)$$

with

$$p_i \equiv (1/(t-t_0)) \int_{t_0}^{t} v_i(s)ds. \qquad (40)$$

This completes the proof after taking the limits.

This simple sufficient condition states that for a hybrid system to be uniformly asymototically stable the weighted average of the logarithmic norms of each realization has to be negative. Therefore, this sufficient condition allows for unstable forms. That is, as long as the stable forms dominate, the overall system is asymptotically stable. This domination can occur in two ways: either the stable forms are strongly stable (i.e., highly negative logarithmic norms), or the time span of the stable forms is large relative to the time span of the unstable ones or a combination of the latter two reasons.

The above interpretation provides a mathematical rationale to the observations made by Chizeck et al [3] while analyzing such systems.

4. Stabilization

This section presents some results concerning the control and stabilization of hybrid systems. These results use standard techniques to control/stabilize hybrid systems.

Definition
A hybrid system is stabilizable if there exists a constant feedback gain matrix K such that the closed loop hybrid system is asymptotically stable.

Theorem 4
A hybrid system is stabilizable if

a. The average system is stabilizable,
b. The following inequality is satisfied

$$\sum_{i=1}^{N} \mu[\delta A_i - \delta B_i K]p_i < 0, \qquad (41)$$

where K is a stabilizing gain matrix of the average system and $\delta\Sigma_i \equiv (\delta A_i, \delta B_i)$ is the difference between the ith realization and the average system.

Proof
Given a hybrid system with a stabilizable average then there exists at least one constant gain matrix K such that the average closed loop matrix (A-BK) is Hurwitz. Therefore, $x_{average}(t)$ is asymptotically stable. But the actual system response is composed of two components, the average system component and the error component. That is

$$x(t) = x_{avrage}(t) + e(t) \qquad (42)$$

where the error dynamics are

$$\dot{e}(t) = \sum_i v_i(t)[\delta A_i - \delta B_i K]e. \qquad (43)$$

Condition b is a sufficient requirement for e(t) to be asymptotically stable which proves the theorem.

The following example is given to illustrate the results.

Example
Given the following F2-system $\Sigma(\alpha)$

$$\dot{x}_1(t) = -\alpha x_1 + x_2 + u, \qquad (44)$$

$$\dot{x}_2(t) = (1 - \alpha)x_2 + \alpha u, \qquad (45)$$

where $\alpha \equiv T_1/T$, $\Sigma_1 = \Sigma(1)$ and $\Sigma_2 = \Sigma(0)$. The above system is the exact average system. The transfer-function of the average system is given by

$$H(s) = (s + (2\alpha - 1))/(s + \alpha)(s - 1 + \alpha). \qquad (46)$$

For $\alpha > 0.5$, (46) is a minimum-phase transfer function, otherwise it is not. Using usual techniques the minimum-phase case can be stabilized with an output feedback gain K. For α=.8 and K=5 the closed loop average system's poles are s_1=-1.13 and s_2=-10.1. However, the logarithmic norm test applied to the error dynamics gives an upper bound equal to zero, implying that the error dynamics are not unstable. A graph of the (α,K)-stabilizability domain and two phase-space simulations are given in Fig. 1 and Fig. 2, respectively, to illustrate the stability results and the effect of the feedback gain K on the dynamics of the closed loop system.

5. Conclusion

An averaging method based on the Baker-Campbell-Hausdorff formula was introduced for computing the average of a hybrid system. In order to be able to find how well the average system is approximating the actual system upper bounds of the error induced by averaging are given. Furthermore, it was shown that under certain conditions one can avoid the averaging of the A-matrix and therefore minimize the errors introduced by the averaging procedure. This is done via state feedback by making the A-matrix constant.

The controllability property of the average system is a key assumption in the stabilization procedure given in the paper. For that reason a sufficient condition that identifies a class of hybrid systems for which the average is controllable was given. Therefore, the class of hybrid systems with a controllable average is a research topic in need of further investigation.

The stability of hybrid systems is still far from being solved. This is mainly due to the fact that hybrid systems are time-varying systems. In the paper a sufficient stability condition was derived. This condition is based on the logarithmic norm concept. One important point to be investigated about this stability condition is how conservative it is? The variable structure property seems to be a promising feature in this direction. Furthermore if one thinks of every system $\Sigma_i = (A_i, B_i, Ci)$ with $i \in N$ as an operator acting on the state x during δt_i, and these operators are applied in a successive manner, then this process can be viewed as an iterative process [15]. Viewing a hybrid system as an iterative process sheds some light on some complicated issues such as the stability of such systems.

Finally adapting the results of this paper to hybrid systems where the switching is a stochastic process such as a Markov chain can be easely done.

REFERENCES

[1] T. L. Johnson, "Synchronous Switching Linear Systems," Proc. 24th. IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 1699-1700, Dec. 1985.

[2] R. W. Brockett and J. R. Wood, "Electrical Networks Containing Controlled Switches," in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, San Francisco, pp. 1-11, April 1974.

[3] H. J. Chizeck, A. S. Willsky and D. Castanon, "Discrete-Time Markovian-Jump Linear Quadratic Optimal Control," Int J. Control, Vol. 43, No. 1, pp. 213-231, 1986.

[4] S. Geman, "Some Averaging and Stability Results for Random Differential Equations," SIAM J. App. Math., Vol. 36, No. 1, Feb. 1979.

[5] R. L. Kosut, B. D. O. Anderson and I. M. Y. Mareels, "Stability Theory for Adaptive Systems: Method of Averaging and Persistency of Excitation," IEEE Trans. Automatic Control, Vol. AC-32, NO. 1, Jan. 1987.

[6] Jelel Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems," Proc. 1988 American Control Conference, Atlanta, GA, June 1988.

[7] J. A. Richards, "Aanalysis of Periodically Time-Varying Systems," Springer-Verlag, 1983.

[8] J. G. F. Belinfante, "Explicit Version of the Campbell-Baker-Hausdorff Formula: Integral Representation for ln $e^{x}e^{y}$," unpublished.

[9] R. M. Wilcox, "Exponential Operators and Parameter Differentiation in Quantum Physics," J. of Math. Physics, Vol.8, pp. 962-982, 1967.

[10] J. Ezzine and C. D. Johnson, "Analysis of Continuous/Discrete Model Parameter Sensitivity via a Perturbation Technique," Proceedings of The Eighteenth Southeastern Symposium on System Theory, pp. 545-550, 1986.

[11] Jelel Ezzine and A. H. Haddad, "On the Controllability of the Average of Hybrid Systems," submitted for publication.

[12] C. Van Loan, "The Sensitivity of the Matrix Exponential," SIAM J. Num. Anal., Vol. 14, No. 6, Dec. 1977.

[13] T. Strom, "On Logarithmic Norms," SIAM J. Num. Anal., Vol. 12, Oct. 1975.

[14] C. D. Desoer and M. Vidyasagar, "Feedback Systems: Input-Output Properties," Academic Press, 1975.

[15] J. N. Tsitsiklis, "On the Stability of Asynchronous Iterative Processes," Pro. 25th. IEEE Conf. Decision and Control, Athens, Greece, pp. 1617-1621, Dec. 1986.

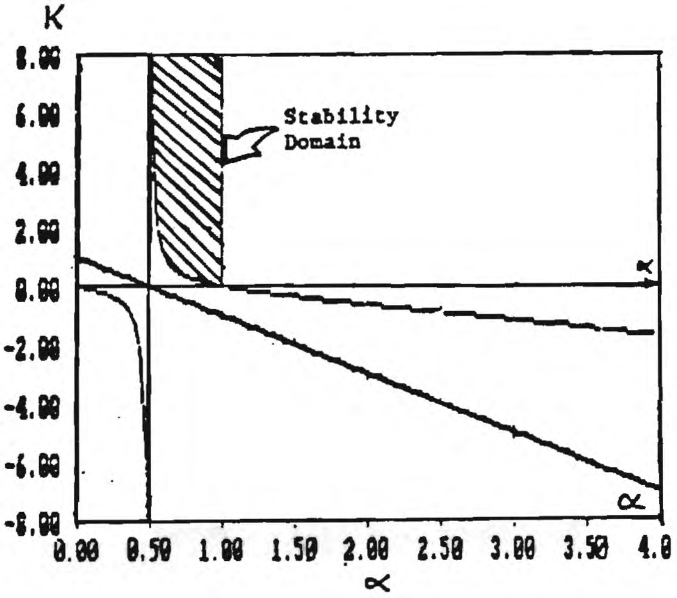

Fig. 1 (α,K) - Stabilizability Domain

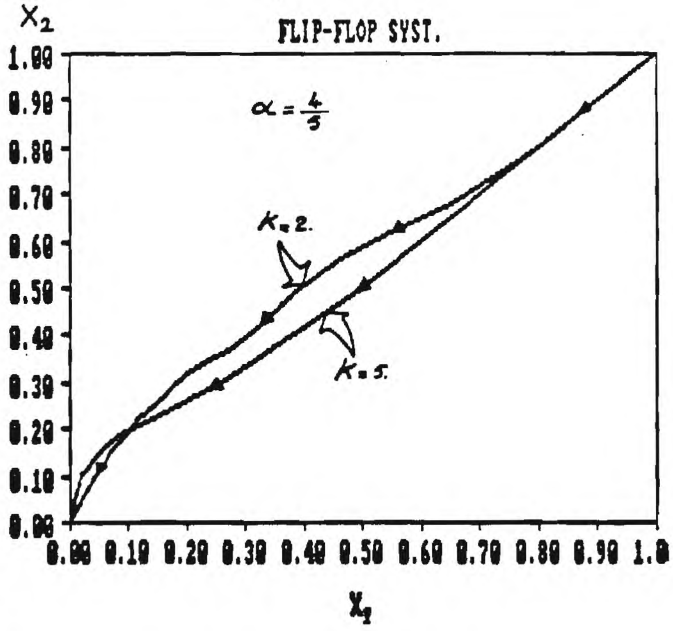

Fig. 2 Phase - Space Simulations

APPENDIX C

J. Ezzine and A. H. Haddad, "Error Bounds in the Averaging of Hybrid Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 1787-1791, Dec. 1988. To appear also in IEEE Transactions on Automatic Control.

Proceedings of the 27th Conference
on Decision and Control
Austin, Texas • December 1988

ERROR BOUNDS IN THE AVERAGING OF HYBRID SYSTEMS[1]

J. Ezzine

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

A. H. Haddad

Department of EE/CS
Northwestern University
Evanston, IL 60208

ABSTRACT

This paper analyzes the errors introduced by the averaging of hybrid systems. These systems involve linear systems which can take a number of different realizations based on the state of an underlying finite state process. The averaging technique (based on a formula from Lie algebras known as the Baker-Campbell-Hausdorff (BCH) formula) provides a single system matrix as an approximation to the hybrid system. The two errors discussed are: a) The error induced by the truncation of the BCH series expansion, and b) The error between the actual hybrid system and its average. A simple sufficient stability test is proposed to check the asymptotic behavior of this error. In addition, conditions are derived that allow the use of state feedback to arrive at a time-invariant system matrix instead of averaging.

1. INTRODUCTION AND PROBLEM FORMULATION

Hybrid systems are a special class of piece-wise constant time-varying systems. Such models switch at different time instants among a finite set of linear time-invariant realizations. Systems of this type can be used to model systems subject to known abrupt parameter variations such as synchronously switched linear systems [1], networks with periodically varying switches [2] or to approximate some types of time-varying systems [3]. This is achieved by imposing a deterministic switching rule [4]. To model unknown abrupt phenomena such as systems subject to failures [5], the switching can be modeled, for example, as a finite-state Markov chain (FSMC). An earlier review of hybrid systems may be found in [6].

Averaging theory, which is used in a deterministic or probabilistic context, is an approach to the approximation of such systems by a single constant linear model. In the probabilistic case averaging is introduced in a natural way by taking expected values. In the deterministic case, however, averaging is introduced via perturbation techniques. Brockett and Wood [2] used a deterministic averaging technique to analyze and stabilize a class of bilinear systems which are very hard to analyze or control otherwise. Geman [7] used probabilistic averaging techniques to study the stability of random differential equations. His main interest was to explore the relation between asymptotic stability in the average equation, and asymptotic stability in the random equation. Specifically, when does the first imply the second? Kosut et al. [8] applied the theory of averaging to the analysis of the stability of adaptive systems. Ezzine and Haddad [9] used an averaging technique very similar to the one used in [2] to analyze and stabilize hybrid systems via a nonswitching gain. As a matter of fact, Mariton et al. [10] showed that nonswitching control gains may be preferable, in addition to the fact that they are much easier to implement.

In this paper the averaging procedure used in [2,9] is considered further. In [2] there was no attempt to analyze the errors introduced by the BCH formula and the averaging method. However, there are two very important issues in using such a formula and averages derived from it as mentioned in [9]. The first is the error introduced by truncating the BCH expression while computing the average matrix. The second is the difference between the actual system, in [9] the F2-system, or in [2] the bilinear system, and the average system used to approximate the average behavior of the system under consideration. This paper addresses both issues by providing bounds on the resulting errors.

Furthermore, the paper provides conditions under which the BCH formula can be avoided. Instead of using the BCH formula to compute an average system matrix, state feedback is used to obtain a constant closed loop matrix for the system.

The class of hybrid systems considered in this paper are assumed to have the form

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t) \tag{1a}$$

$$y(t) = C(r(t))x(t); \tag{1b}$$

where x is the system state vector of dimension n, u is the control input vector of dimension p, y is the system output vector of dimension m, and r(t) is the "form index" which is a deterministic scalar sequence taking values in the finite index set $\mathbf{N}=\{1, 2, \ldots, N\}$. Let the ith form denote the realization $\Sigma_i=(A_i,B_i,C_i)$ associated with the ith form index (i.e., r(t)=i), for $i \in \mathbf{N}$.

It is assumed that any r(t) sequence is composed of a succession of N-termed blocks. Every block is a permutation of the index set **N**. It is important to note that the succession of the blocks is <u>completely arbitrary</u> (e.g., for N=3, a possible r(t)-sequence is: 123, 321, 213, 213, 312, ...). The time interval during which r(t)=i is denoted by δt_i. In this case we define

$$T \equiv \sum_{i=1}^{N} \delta t_i \tag{2}$$

as the <u>period</u> of the system. Piece-wise constant periodic systems are a special class of hybrid systems. Therefore, from an application point of view the subsequent results can, at least, be applied to the periodic case. However, the primary motivation is to derive results that can be applied to the case where the switching is governed by a FSMC.

The following is an outline of the paper. Section 2 begins with an overview of the averaging technique for hybrid systems. It also addresses the perturbations induced by the averaging procedure. Two important perturbation errors are identified, and the

[1] This work is supported by the U.S. Air Force under grant AFOSR-87-0308.

88CH2531-2/88/0000-1787$1.00 © 1988 IEEE

first one is analyzed. Section 3 discusses the second error, and also offers, where possible, an alternative to averaging. Section 4 concludes the paper and points to additional open questions.

2. DERIVATION OF THE PERTURBATION BOUNDS

This section addresses the accuracy of the averaging technique. First, upper bounds for the errors introduced by using a truncated BCH formula are derived.

The averaging methodology to be analyzed in this paper is based on a formula from Lie algebras known as the Baker-Campbell-Hausdorff (BCH) formula [2,11,12]: Given two real matrices A and B there is no guarantee that there exists a real matrix C such that

$$\mathrm{Exp}(A)\mathrm{Exp}(B) = \mathrm{Exp}(C). \qquad (3)$$

However, if $\|A\|+\|B\|\leq \ln(2)$, then C exists [11], and will be given by a convergent infinite expression (the BCH formula)

$$C = A+B+(1/12)[[A,B],B]+(1/12)[[B,A],A]+\ldots \qquad (4)$$

where the symbol $[A,B]\equiv AB-BA$ (i.e., the commutator product).

Similar expressions like the BCH formula are used in a large number of useful approximations in physics [12] and switched electrical networks [2]. In the sequel we show how the BCH formula and related expressions can be used to compute the average of N-form hybrid systems. However, for notational simplicity F2-systems (i.e., two-form hybrid systems) only are treated.

The averaging idea introduced in [2] and used in [9] to stabilize hybrid systems is outlined in the following section. Given an F2-system such that

$$A(t) = \begin{cases} A_1 & \text{for } 0 \leq t < T_1, \\ A_2 & \text{for } T_1 \leq t < T. \end{cases} \qquad (5)$$

then an approximation for $\mathrm{Exp}(A_1T_1)\mathrm{Exp}(A_2(T-T_1))$ is desired. The BCH formula provides the approximation as Exp(CT), where C is as in (4). Now if we require an expression for C independent of the order of the product, to agree with the order independence introduced in the definition of F2-systems, then we must compute

$$\ln\tfrac{1}{2}\{\mathrm{Exp}(A)\mathrm{Exp}(B)+\mathrm{Exp}(B)\mathrm{Exp}(A)\}$$
$$\approx A+B+(1/12)[[A,B],B]+(1/12)[[B,A],A]$$
$$=A+B+(1/12)[[A,B],B-A]. \qquad (6)$$

Therefore, we obtain a series of approximate averages for F2-systems. If the F2-system realization is $\Sigma_1=(A_1,b_1)$ for period T_1, and $\Sigma_2=(A_2,b_2)$ for period T_2, then the first order approximation is

$$\Sigma = (\alpha A_1+(1-\alpha)A_2,\ \alpha b_1+(1-\alpha)b_2), \qquad (7)$$

with $\alpha \equiv T_1/(T_1 + T_2)$. Second-order terms may be included to obtain the approximation

$$\Sigma = ([\alpha A_1+(1-\alpha)A_2+(1/12)\alpha(1-\alpha)[[A_1,A_2],$$
$$(1-\alpha)A_2-\alpha A_1],\ \alpha b_1+(1-\alpha)b_2). \qquad (8)$$

Higher-order approximations may also be derived. Consequently, if we let $\hat{C}$ denote the approximating matrix, then C can be written as $C=\hat{C}+\tilde{C}$, where $\tilde{C}$ is the error due to the approximation. Therefore, the induced error in computing Exp(CT) by using the matrix $\hat{C}$ is

$$E \equiv \mathrm{Exp}(CT)-\mathrm{Exp}(\hat{C}T). \qquad (9)$$

The solution formula for inhomogeneous differential equations can be used to derive an exact expression of E [13]:

$$E = \int_0^T \{\mathrm{Exp}(\hat{C}(T-s))\tilde{C}\mathrm{Exp}((\hat{C}+\tilde{C})s)\}ds. \qquad (10)$$

In this section a useful approximate expression for E is derived using perturbation techniques. To do so we define $\tilde{C}\equiv\varepsilon C_p$, where ε is a scalar. It is recalled [13] that if $(\hat{C},\tilde{C})$ commute, that is,

$$\hat{C}\tilde{C} = \tilde{C}\hat{C}, \qquad (11)$$

then

$$\mathrm{Exp}((\hat{C}+\varepsilon C_p)T) = \mathrm{Exp}(\hat{C}T)\mathrm{Exp}(\varepsilon C_pT)$$
$$= \mathrm{Exp}(\hat{C}T)(I + \varepsilon C_pT + \varepsilon^2C_p^2T^2/2! + \ldots). \qquad (12)$$

To find $\mathrm{Exp}((\hat{C}+\tilde{C})T)$, when $\hat{C}$ and C_p do not commute, one can use an iterative technique similar to the one used to derive the exact expression for E [13]. Hence one can write

$$\mathrm{Exp}((\hat{C}+\tilde{C})T) = \mathrm{Exp}(\hat{C}T) +$$
$$\varepsilon\mathrm{Exp}(\hat{C}T) \int_0^T \mathrm{Exp}(-\hat{C}s)C_p\mathrm{Exp}(\hat{C}s)ds + O(\varepsilon^2). \qquad (13)$$

This is the Liouville-Neumann series solution for the integral equation. It is a convergent perturbation series for all ε [13].

However, these exact expressions, given as a series in ε, do not lend much insight to qualitative analysis questions. In that regard, transforming those integral expressions for E into differential equations might be more useful. We first introduce the following definition:

$$E_1 \equiv E - O(\varepsilon^2)$$
$$= \varepsilon\mathrm{Exp}(\hat{C}T) \int_0^T \mathrm{Exp}(-\hat{C}s)C_p\mathrm{Exp}(\hat{C}s)ds, \qquad (14)$$

where E_1 is the first order approximation to E in ε. Now E_1 can be expressed as follows.

Proposition 1 [14]

Let E_1 denote the first order approximation in ε to E, then E_1 satisfies the following matrix differential equation:

$$\dot{Y} = \hat{C}Y + \varepsilon C_p\mathrm{Exp}(\hat{C}t), \qquad Y(0) = 0. \qquad (15)$$

As a consequence of the above representation one can use the theory of linear matrix differential equations to study the qualitative behavior of E_1. For example, one can show that if all eigenvalues of $\hat{C}$ have negative real parts then $E_1(t) \to 0$ as $t \to \infty$.

Moreover, it is possible to derive a general explicit expression for E_1. To do so we first recall the well known result [12]

$$\mathrm{Exp}(sA)B\mathrm{Exp}(-sA) = \sum_{i=0}^{\infty} (s^i/i!)\{A^i,B\}, \qquad (16)$$

with $\{A^0,B\}=B$ and $\{A^{n+1},B\}=[A,\{A^n,B\}]$. Using this identity the general explicit expression for E_1 will follow

$$-E_1 = \varepsilon\mathrm{Exp}(\hat{C}t) \sum_{i=1}^{\infty} (t^i/i!)\{(-\hat{C})^{i-1}, C_p\}. \qquad (17)$$

At this point, we are ready to compute upper bounds for E_1 and E.

Proposition 2

Assume that $\|\mathrm{Exp}(\hat{C}t)\|\leq M(t)\mathrm{Exp}(\beta(t)t)$, with $\beta(t)$ a scalar function, then

$$\|E_1\|\leq(\|\tilde{C}\|/2\|\hat{C}\|)M(t)\mathrm{Exp}\{\beta(t)t\}(\mathrm{Exp}\{2\|\hat{C}\|t\}-1), \qquad (18)$$

and with the added assumption of M(t) being monotone then

$$\|E\| \leq t\|\tilde{C}\|M^2(t)\mathrm{Exp}\{(\beta(t) + M(t)\|\tilde{C}\|)t\}. \qquad (19)$$

Proof

Using (16) in the evaluation of the integral in (14) yields

$$E_1 = \mathrm{Exp}(-At) \sum_{i=1}^{\infty} (t^i/i!)\{A^{i-1},B\}. \qquad (20)$$

Taking the norm of both sides in (20) and using the following inequality

$$\|\{A^n,B\}\| \leq 2^n \|A^n\| \|B\| = \|2A\|^n \|B\| \qquad (21)$$

leads to

$$\|E_1\| \leq \|\mathrm{Exp}(-At)\| \sum_{i=1}^{\infty} (t^i/i!)\|2A\|^{i-1}\|B\|, \qquad (22)$$

which after simple algebra results in (18). Equation (19) follows in the same manner by using the monotonicity property of M(t).

3. STABILITY OF THE ERROR DYNAMICS

In this section the dynamics of the error between the average system and the actual hybrid system are derived. The stability of the error dynamics is discussed and two stability criteria are introduced.

Given a homogeneous N-Form hybrid system with state vector x(t), and one of its time-invariant averages with state vector $x_a(t)$, the error is given as

$$e(t) \equiv x(t) - x_a(t) \qquad (23)$$

From this definition it is easy to see that the error dynamics are governed by the following hybrid system

$$\dot{e}(t) = (\sum_{i=1}^{N} v_i(t)\ \delta A_i)e(t) \qquad (24)$$

where $\delta A_i \equiv A_i - A_a$ and the indicator functions $v_i(t)$ are defined by: $v_i(t)=1$ when the original hybrid system is described by the ith realization Σ_i, and $v_i(t)=0$ otherwise.

At this point two stability criteria are introduced to check the stability of (24). The first criterion is a necessary and sufficient condition [4] and the second one is a sufficient test only [9, 4]. Because the first condition is computationally involved and is a generalization of a well known result (see [4]) we choose to present the second one. Interestingly enough, the second test is more general in the sense that it can be easily generalized to a larger class of hybrid systems.

In order to state this sufficient condition a brief introduction to the notion of logarithmic norm is given.

The logarithmic norm (also known as the logarithmic derivative, the measure of a matrix) was introduced in 1958 separately by Dahlquist [15] and Lozinskij [16] as a tool to study the growth of solutions to ordinary differential equations and the error growth in discretization methods for their approximate solution. It is formally defined as follows:

Definition

The logarithmic norm associated with the matrix norm $\|.\|$ is defined by

$$\mu(A) = \lim_{h\to 0^+} (\|I + hA\| - 1)/h \qquad (25)$$

Explicit expression for the logarithmic norm associated with the Euclidean norm is

$$\mu(A) = \max\{\mu : \mu \in \lambda((A+A^*)/2)\}, \qquad (26)$$

where $\lambda(A)$ is the set of eigenvalues corresponding to the matrix A. Then the following inequality is true:

$$\|\mathrm{Exp}(At)\| \leq \mathrm{Exp}(\mu(A)t). \qquad (27)$$

Theorem 3

For the null solution of the hybrid system (24) to be uniformly asymptotically stable, it is necessary that

$$\sum_i \mu(-A_i)p_i > 0, \qquad (28a)$$

and sufficient to have

$$\sum_i \mu(A_i)p_i < 0, \qquad (28b)$$

where

$$p_i \equiv (t_i-t_{i-1})/T,\ i\varepsilon N.$$

Proof

We start by showing that the sufficient condition holds. Using Theorem 27 in [17] one can write

$$\|e(t)\| \leq \|e(t_0)\|\mathrm{Exp}\{ \int_{t_0}^{t} \mu(A(s))ds\}$$

$$= \|e(t_0)\|\ \mathrm{Exp}\{ \int_{t_0}^{t} \mu[\sum_{i=1}^{N} v_i(t)A_i]ds \}. \qquad (29)$$

Using Theorem 5(e, d) in [17] and after some algebra we get

$$\lim_{t\to\infty}\|e(t)\| \leq \lim_{t\to\infty}\|e(t_0)\|\ \mathrm{Exp}\{ \sum_{i=1}^{N} \int_{t_0}^{t} v_i(t)\mu[A_i]ds\}$$

$$= \lim_{t\to\infty} \|e(t_0)\| \, \mathrm{Exp}\{(\sum_{i=1}^{N} p_i \mu[A_i])(t-t_0)\} \qquad (30)$$

with

$$p_i = \lim_{t\to\infty} (1/(t-t_0)) \int_{t_0}^{t} v_i(s)ds. \qquad (31)$$

which completes the proof of (28b).

The necessary condition (28a) is shown similarly by using the fact that $\mathrm{Exp}\text{-}\{\mu(-A)\} \le \|\mathrm{Exp}At\|$.

This simple sufficient condition states that for a hybrid system to be uniformly asymptotically stable the weighted average of the logarithmic norms of each realization has to be negative. Therefore, this sufficient condition allows for unstable forms. That is, as long as the stable forms dominate the overal system is asymptotically stable. This domination can occur in three ways: either the stable forms are strongly stable (i.e., highly negative logarithmic norms) or the time span of the stable forms is large relative to the time span of the unstable ones or a combination of both reasons. This stability property of hybrid systems was reported in [5] via examples.

The difference between (28b) and (28a) could be used as a measure of the conservativeness of (28b).

Sometimes it is possible for F2-systems to avoid the computation of an average matrix and, consequently, minimize the errors induced by averaging [9]. That is, under certain conditions, it is possible, via state feedback, to make the closed loop F2-system A-matrix time-invariant. That is, given Σ_1 and Σ_2 satisfying appropriate conditions, one can compute a feedback gain matrix

$$K = [K_1 \mid K_2] \qquad (32)$$

which will make the A-matrices of both forms equal, so that the A-matrix of the closed loop hybrid system (using gain K_i for Σ_i) becomes a constant. Moreover, this constant A matrix is given by the following equation

$$A = A_1 - B_1K_1 = A_2 - B_2K_2. \qquad (33)$$

In [18], Mariton proposed a technique quite similar to this idea. He showed that it is possible to solve the Jump Linear Quadratic (JLQ) problem by making the performance index independent of the different realizations of the form-index r(t). His approach renders the cost incurred by any realization of r(t) the same. In other words, it makes all realizations equal in that sense.

Even though the goals seem similar, the approaches are not. In contrast to [18], the present equalization is direct; the homogeneous parts of the two forms are made equal via feedback. In this section a sufficient condition is given for the existence of K and a simple computational algorithm based on the Kronecker-product and the generelized-inverse techniques is proposed. As stated above, the following results hold for N=2 only.

Theorem 4

Given the F2-system

$$\dot{x} = A_ix + B_iu, \ i = 1,2. \qquad (34)$$

such that

$$\mathrm{Range}[A_1-A_2 \mid B_1 \mid -B_2] = \mathrm{Range}[B_1 \mid -B_2], \qquad (35)$$

then, there exists a minimum-norm $G \equiv [K_1 \mid K_2]$ such that

$$A_1 - B_1K_1 = A_2 - B_2K_2. \qquad (36)$$

Moreover, the G matrix is given by

$$s(G) = \{[B_1 \mid -B_2]^{\dagger} \otimes I_n\}s(A_1-A_2), \qquad (37)$$

where s(.) is the stacking operator, $(.)^{\dagger}$ is the generelized-inverse, and $\otimes$ is the Kronecker-product.

Proof

Starting with the forms Σ_i, i=1, 2, the ith model is given by

$$\dot{x}(t) = A_ix + B_iu. \qquad (38)$$

If we define $A \equiv A_i - \delta A_i$ the above system can be written as follows

$$\dot{x}(t) = Ax + \delta A_ix + B_iu, \qquad (39)$$

The next step is to compute a gain matrix K_i such that

$$\delta A_ix + B_iu = \delta A_ix - B_iK_ix$$
$$= (\delta A_i - B_iK_i)x = 0 \ \text{ for all } x. \qquad (40)$$

This is equivalent to

$$B_iK_i = \delta A_i, \qquad (41)$$

which is an algebraic equation for the unknown K_i. Substructing the first equation from the second yields

$$A_1 - A_2 = B_1K_1 - B_2K_2 \qquad (42)$$

which can be written as follows

$$[A_1 - A_2] = [B_1,-B_2][K_1',K_2']' \qquad (43)$$

which is itself an algebraic equation with K_i, i=1, 2, as unknown and the condition given in the theorem is the one needed for the existence of both gains. Equation (37) is a compact way to write these equations, and is also useful when numerical techniques are used to solve the problem.

Corollary

If
$$\mathrm{Range}[A_1-A_2 \mid B_1-B_2] = \mathrm{Range}[B_1-B_2], \qquad (44)$$

then
$$K \equiv K_1 = K_2, \qquad (45)$$

and the gain matrix is given by:

$$s(K) = \{[B_1-B_2]^{\dagger} \otimes I_n\}s(A_1-A_2). \qquad (46)$$

To illustrate the above results consider the following example.

Example

Given the following F2-system

$$A_1 = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}, \; b_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \; b_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

it is easy to check that after closing the loop via the gain $K=(1 \;\; 2)$, the system A-matrix becomes the identity for $i=1, 2$.

4. CONCLUSION

In this paper the errors introduced by averaging hybrid systems have been analyzed. Two errors have been identified and discussed. The first is the error induced by trucating the BCH expression while computing the average A-matrix. The second is the error between the actual hybrid system and its average. The paper provides two upper bounds for the first error along with a simple sufficient stability criterion for the second one.

It is important to note that the results of this paper are independent of the averaging technique used to compute the average system. Thus, the results can be used, for instance, when the switching is governed by a stochastic process such as a Markov chain and the average system is the probabilistic average of the hybrid system. However, the stability tests used in the paper have to be adjusted accordingly. As a matter of fact, imposing some ergodicity conditions on the dynamics of the hybrid system yields exactly the same sufficient stability condition when properly interpreted.

A different direction that might be helpful in deriving better stability conditions for such systems is to think of every system $\Sigma_i=(A_i,B_i,C_i)$ with $i\epsilon N$ as an operator acting on the state x during δt_i, and these operators are applied in a successive manner, then this process can be viewed as an iterative process [19]. Viewing a hybrid system as an iterative process sheds some light on the complexity of the stability of such systems.

It would be interesting to compare the present work and [9] to [20] where a different type of averaging hybrid systems is discussed.

REFERENCES

[1] T. L. Johnson, "Synchronous Switching Linear Systems," Proc. 24th IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 1699-1700, Dec. 1985.

[2] R. W. Brockett and J. R. Wood, "Electrical Networks Containing Controlled Switches," in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, San Francisco, pp. 1-11, April 1974.

[3] J. A. Richards, Analysis of Periodically Time-Varying Systems, Springer-Verlag, 1983.

[4] J. Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems," Proc. 1988 American Control Conf., Atlanta, GA, pp. 41-46, June 1988.

[5] H. J. Chizeck, A. S. Willsky and D. Castanon, "Discrete-time Markovian Jump Linear Quadratic Optimal Control," Int J. Control, Vol. 43, No. 1, pp. 213-231, 1986.

[6] D. D. Sworder, "Control of systems subject to sudden changes in character," Proc. IEEE, Vol. 64, No. 8, pp. 1219-1225, 1976.

[7] S. Geman, "Some Averaging and Stability Results for Random Differential Equations," SIAM J. App. Math., Vol. 36, No. 1, Feb. 1979.

[8] R. L. Kosut, B. D. O. Anderson and I. M. Y. Mareels, "Stability Theory for Adaptive Systems: Method of Averaging and Persistency of Excitation," IEEE Trans. Automat. Cont., Vol. AC-32, NO. 1, pp. 26-34, Jan. 1987.

[9] J. Ezzine and A. H. Haddad, "On the Stabilization of Two-Form Hybrid Systems Via Averaging," Proc. 22nd Annual Conf. on Information Sciences and Systems, Princeton University, pp. 579-584, March 1988.

[10] M. Mariton and P. Bertrand, "Non-switching Control Strategies for Continuous-time Jump Linear Quadratic Systems," Proc. 24th IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 916-921, Dec. 1985.

[11] J. G. F. Belinfante, "Explicit Version of the Campbell-Baker-Hausdorff Formula: Integral Representation for ln e^xe^y," unpublished.

[12] R. M. Wilcox, "Exponential operators and parameter differentiation in Quantum Physics," J. of Math. Physics, Vol.8, pp. 962-982, 1967.

[13] R. Bellman, Introduction to Matrix Analysis, McGraw-Hill, 1960.

[14] J. Ezzine and C. D. Johnson, "Analysis of Continuous-Discrete Model Parameter Sensitivity via a Perturbation Technique," Proc. Eighteenth Southeastern Symposium on System Theory, pp. 545-550, 1986.

[15] C. Van Loan, "The Sensitivity of the Matrix Exponential," SIAM J. Num. Anal., Vol. 14, No. 6, Dec. 1977.

[16] T. Strom, "On Logarithmic Norms," SIAM J. Num. Anal., Vol. 12, No. 5, Oct. 1975

[17] C. A. Desoer and M. Vidyasagar, Feedback Systems: Input-Output Properties, Academic Press, 1975.

[18] M. Mariton, "The Equalizing Solution of the JLQ Problem," Proc. 26th IEEE CDC, Los Angeles, CA, Dec. 1987.

[19] J. N. Tsitsiklis, "On the Stability of Asynchronous Iterative Processes," Pro. 25th. IEEE Conf. Decision and Control, Athens, Greece, pp. 1617-1621, Dec. 1986.

[20] D. A. Castanon et al. "Asymptotic Analysis, Approximation and Aggregation Methods for Stochastic Hybrid Systems," Proc. 1980 JACC, San Francisco, CA, paper TA3-D, 1980.

APPENDIX D

J. Ezzine and A. H. Haddad, "On the Minimality of the Average of Hybrid Systems", Proc. IEEE Conference on Control and Applications, Jerusalem, Israel, pp. RA-6-2/1-4, April 1989.

ON THE MINIMALITY OF THE AVERAGE OF HYBRID SYSTEMS**

Jelel Ezzine

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

A. H. HADDAD

Department of EE/CS
Northwestern University
Evanston, IL 60208-3118

ABSTRACT

The stabilization of hybrid systems with a non-switching gain is cheaper and simpler to implement than the switching one. One approach to the design of a non-switching gain is based on the averaging of the hybrid system. For obvious reasons, the non-switching gain exists if the average system is controllable. In this paper, the minimality of the average system is investigated and a sufficient criterion is derived. Furthermore, these results also shed some light on the topology of minimal LTI systems in parameter space.

1. INTRODUCTION AND PROBLEM FORMULATION

Hybrid systems are a special class of piece-wise constant time-varying systems. Such models switch at different time instants among a finite set of linear time-invariant realizations. Systems of this type can be used to model systems subject to known abrupt parameter variations such as synchronously switched linear systems[1], networks with periodically varying switches[2] or to approximate some types of time-varying systems[3]. This is achieved by imposing a deterministic switching rule[4]. However, to model unknown abrupt phenomena such as systems subject to failures[5] the switching can be modeled, for example, as a finite-state Markov chain (FSMC). An earlier review of hybrid systems may be found in Sworder's paper[6].

Averaging theory, which is used in a deterministic or probabilistic context, is an approach to the approximation of such systems by a single constant linear model. In the probabilistic case averaging is introduced in a natural way by taking expected values. In the deterministic case, however, averaging is introduced via perturbation techniques. Brockett and Wood[2] used a deterministic averaging technique to analyze and stabilize a class of bilinear systems which are very hard to analyze or control otherwise. Geman[7] used probabilistic averaging techniques to study the stability of random differential equations. His main interest was to explore the relation between asymptotic stability in the average equation, and asymptotic stability in the random equation. Specifically, when does the first imply the second? Kosut et al.[7] applied the theory of averaging to the analysis of the stability of adaptive systems. Ezzine and Haddad[9] used an averaging technique very similar to the one used in Brockett et al.'s paper[2] to analyze and stabilize hybrid systems via a nonswitching gain. As a matter of fact, Mariton et al.[10] showed that nonswitching control gains may be preferable, in addition to the fact that they are much easier to implement.

The paper also considers deterministic hybrid systems, when such systems capture the essence of stochastic hybrid systems. In this case the addition to the fact that they are much easier to implement.

The hybrid systems considered in this paper are assumed to have the form

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t) \qquad (1.1)$$

$$y(t) = C(r(t))x(t) \qquad (1.2)$$

where x is the system state vector of dimension n, u is the control input vector of dimension p, y is the output vector of dimension m, and r(t) is the "form index" which is either a deterministic or a stochastic scalar sequence taking values in the finite index set $\mathbf{N}=\{1, 2, \dots, N\}$.

The system takes the realization $\Sigma_i=(A_i,B_i,C_i)$ when $r(t) = i$, with $i\in\mathbf{N}$. This realization is called the ith form. δt_i denotes the time interval during which $r(t) = i$. In addition we define

$$T \equiv \sum_{i=1}^{N} \delta t_i \qquad (2)$$

as the <u>period</u> of the system.

Sometimes it is more convenient to represent the hybrid system in an equivalent different form which leads to the following representation

$$\dot{x}(t) = \{\sum_{i=1}^{N} v_i(t)A_i\}x(t) + \{\sum_{i=1}^{N} v_i(t)B_i\}u(t) \qquad (3.1)$$

$$y(t) = \{\sum_{i=1}^{N} v_i(t)C_i\}x(t) \qquad (3.2)$$

where $v_i(t) = 1$ when the system is governed by the ith realization Σ_i, and $v_i(t) = 0$ otherwise. The $v_i(t)$ function is called the <u>ith indicator function</u>. It is evident from the definition of hybrid systems that at any point in time only one of the N indicator functions takes the value one.

** This research is supported by the U.S. Air Force under contract F08635-84-C-0273 (with the Armament Laboratory) and grant AFOSR-87-0308.

This paper mostly addresses the case where where $r(t)$ is a stochastic process, which is assumed to be governed by a Finite State Markov Chain (FSMC) with probabilities

$$\Pr\{r(t+\tau) = j \mid r(t) = i\} = p_{ij}(\tau), \qquad (4.1)$$

for continuous-time systems. In case the dynamics of the hybrid system are discrete the transition probabilities are givin by

$$\Pr\{r(t+1) = j \mid r(t) = i\} = p_{ij}. \qquad (4.2)$$

It is also assumed, Throughout this research work, unless stated otherwise, that the FSMC is stationary and irreducible[11].

sequence $r(t)$ is assumed to be deterministic and to be composed of a succession of N-termed blocks. Every block is a permutation of the index set **N**. It is important to note that the succession of the blocks is completely arbitrary (e.g., for N=3, a possible r(t)-sequence is: 123,321,213,213,312,...).

The following is an outline of the paper. Section 2 begins with an overview of the averaging techniques for hybrid systems. In section 3 the controllability of hybrid systems is recalled along with a necessary contollability condition and a stabilization result. Section 4 deals with the controllability of the average of a hybrid system. This result is used to derive the main theorem of the paper, which identifies a class of hybrid systems for which the average is minimal. Section 5 concludes the paper and points to additional open questions.

2. THE AVERAGE SYSTEM

Two averaging techniques are concidered in this paper. The first one is the probabilistic average of the hybrid system when the switching is governed by an irreducible FSMC. The second one is the first order average of the hybrid system when the switching is deterministic as discussed above. The latter average is based on the Backer-Campbell-Hausdorff formula2.

The two averages are identical in form. In fact, both averages are weighted averages. In the probabilistic case the weights are the components of the stationary probability vector of the FSMC. In the deterministic case the weights are the relative time-spans spent by each form. Therefore, the following representation, of the average, is adopted for both cases in the rest of the paper

$$\Sigma_a \equiv \alpha_1\Sigma_1 + \alpha_2\Sigma_2 + \ldots + \alpha_N\Sigma_N, \qquad (5)$$

where the $\alpha_i \neq 0$ are the steady state probabilities of $r(t)=i$ for the stochastic case, and are defined as

$$\alpha_i = \delta t_i / T \qquad (6)$$

for the deterministic case with

$$T = \sum_{i=1}^{N} \delta t_i \qquad (7)$$

3. CONTROLLABILITY OF HYBRID SYSTEMS

Before dealing with the controllability of the average system a definition of the controllability of deterministic hybrid systems[4] is proposed along with few related results.

Definition:
A deterministic hybrid system is said to be state-controllable if for any t_0 each state $x(t_0)$ can be transferred to any final state x_f after one period. Thus there exists a t_f, $t_0+T \leq t_f < \infty$ such that $x(t_f)=x_f$.

The next result is a necessary algebraic controllability condition. Basically the theorem says that in order for the hybrid system to be controllable it is necessary that the sum of the controllable subspaces of the forms to be equal to the whole space.

Theorem[4]
A necessary algebraic condition for a deterministic hybrid system to be controllable is

$$\text{rank}[\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N] \equiv \text{rank}\ \bar{C} = n. \qquad (8)$$

Where $\mathcal{C}_i$ is the controllability matrix for the ith form, $i \in \mathbf{N}$.

Due to the importance of this theorem a heuristic proof presented in Ezzine et al.s paper[4] that shows that the above condition is almost sufficient will be repeated here. The heuristic argument can be given as follows: Since any matrix exponential is a perturbation of the identity matrix it follows that multiplying any matrix with matrix exponentials will not change its range space drastically. That is if, for example, $\mathcal{C}_1$ and $\mathcal{C}_2$ have algebraic complementery range spaces (i.e, range($\mathcal{C}_1$) is perpendicular to range($\mathcal{C}_2$)) then range(Exp(AT)$\mathcal{C}_1$) will almost always remain an algebraic complement but not necessarely perpendicular to range($\mathcal{C}_2$). As a matter of fact, Mariton[12] states that he has proved that theorem 6 is also a sufficient condition when the switching is governed by a continuous FSMC.

Using the above definition it is possible to prove in a classical way[4] that deterministic hybrid systems are uniformly completely controllable iff they are controllable. Therefore, the above algebraic condition plays almost exactly the same role as the usual algebraic condition for LTI systems.

At this point and in the light of the preceding paragraphs we would like to mention the work of Ikeda et al.[13]. In their work they looked at the relation between controllability properties of the system and various degrees of stability of the closed loop system resulting from linear feedback of the state variables. Their results are as follows: For any initial time t_0, and any continuous and monotonically nondecreasing function $\delta(.,t_0)$ such that $\delta(t_0,t_0)=0$, the transition matrix $\bar{\Phi}(.,.)$ of the closed loop system can be made such that $\|\bar{\Phi}(t,t_0)\| \leq a(t_0)\text{Exp}\{-\delta(t,t_0)\}$

for all $t \geq t_0$, iff the system is completely controllable. Furthermore, in case of a bounded system, for any $m \leq 0$, a bounded feedback matrix can be found such that $\|\Phi(t_2,t_1)\| \leq a\mathrm{Exp}\{-m(t_2-t_1)\}$ for all t_1, $t_2 \geq t_1$, iff the system is uniformly completely controllable. Thus, their results can be regarded, in some sense, as extensions of the well known results of closed loop pole assignment for time-invariant systems.

Hence, there is a high degree of flexibility in the stabilization of hybrid systems if they are controllable or, equivalently, uniformly completely controllable.

As an illustration of the above results we recall, with a slight generalization, a stabilization theorem[4] for hybrid systems where averaging is used.

Definition

A hybrid system is almost surely stabilizable if there exists a constant feedback gain matrix K such that the closed loop hybrid system is asymptotically stable.

Theorem[4]

A hybrid system is almost surely stabilizable if

a. The average system is stabilizable,

b. The following inequality is almost surely satisfied

$$\sum_{i=1}^{N} \mu[\delta A_i - \delta B_i K] p_i < 0, \qquad (9)$$

where K is a stabilizing gain matrix of the average system, $\delta\Sigma_i \equiv (\delta A_i, \delta B_i)$ is the difference between the ith realization and the average system, and $\mu(.)$ is the logarithmic norm of $(.)$[14].

4. CONTROLLABILITY OF THE AVERAGE SYSTEM

We now turn to the controllability of the average system in light of the above theorem. One of the key assumptions made to design the regulator via averaging is the controllability of the average system. This assumption is not unreasonable since the controllability property of linear time invariant systems is generic. However, one can construct hybrid systems such that their averages are not controllable; such a system is a 2-form hybrid system (F2-system):

$$\Sigma_1 = (A_1, b_1) \quad \text{and} \quad \Sigma_2 = (A_2, b_2)$$

with

$$A_1 = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

It is easy to check that neither Σ_1 nor Σ_2 is controllable but that the F2-system is[4]. After derivation of the average system the determinant of its controllability matrix as a function of α was computed. The latter determinant is given by the following third order polynomial:

$$P(\alpha) = \alpha(2\alpha^2 - 3\alpha + 1).$$

The zeros of the above polynomial are $\alpha_1 = 0.$, $\alpha_2 = 1.$ and $\alpha_3 = .5$. The two first zeros α_1 and α_2 are a consequence of the fact that Σ_1 and Σ_2 are both uncontrollable. However, the zero $\alpha_3 = .5$ is a result of the averaging, therefore it refutes the claim that the average system of a controllable hybrid system is necessarily controllable.

In th sequel we give a sufficient condition that identifies a class of hybrid systems for which the hybrid system's controllability guaranties the controllability of the average system.

Theorem 1

The average system of a N-form hybrid system is controllable if

a- rank $[\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N] = n$, where $\mathcal{C}_i$ is the controllability matrix of the ith form,

b- All forms are simultaneously diagonalizable,

Proof:

The average system $\Sigma = (A,B)$ is given by

$$A = \sum_{i=1}^{N} \alpha_i A_i, \quad \text{and} \quad B = \sum_{i=1}^{N} \alpha_i B_i \qquad (10)$$

with

$$\sum_{i=1}^{N} \alpha_i = 1, \text{ and } \alpha_i \neq 0 \text{ for } i = 1, 2, .., N. \qquad (11)$$

Since all forms, $i=1, 2, \ldots, N$, are diagonalizable with the same similarity transformation then

$$T^{-1} A_i T = \Lambda_i, \qquad (12)$$

$$T^{-1} B_i = \Gamma_i \quad \text{for } i=1, 2, \ldots, N. \qquad (13)$$

The T matrix is the common modal matrix for all the forms, and Λ_i is the diagonal matrix corresponding to the i-th form.

Using the above result the average system can be transformed to the following form

$$\Lambda \equiv T^{-1}AT = T^{-1}(\Sigma_i\, \alpha_i A_i)T = \Sigma_i\, \alpha_i \Lambda_i, \qquad (14)$$

and

$$\Gamma \equiv T^{-1}B = T^{-1}(\Sigma_i\, \alpha_i B_i) = \Sigma_i\, \alpha_i \Gamma_i. \qquad (15)$$

Because of assumption (b) Λ will be diagonal too. Now, invoking condition (a) every 1ow in the Γ matrix must have at least one nonzero entry[15] which concludes the proof.

Using theorem 1 and the duality principle it is easy to derive the following theorem.

Theorem 2
The average system of a N-form hybrid system is minimal provided that

a- rank$[\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N] = n$, where $\mathcal{C}_i$ is the controllability matrix of the ith form,

b-

$$\text{rank}\begin{bmatrix} O_1 \\ O_2 \\ . \\ . \\ . \\ O_N \end{bmatrix} = n,$$

where Oi is the observability matrix of the ith form,

c- All forms are simultaniously diagonalizable.

Using this theorem it is easy to design a constant regulator for hybrid systems, for which the above three conditions hold, using standard LTI-design techniques. For an example the interested reader is referred to Ezzine et al.'s work[9].

5. CONCLUSION

The main result presented in this paper is a theorem that identifies a class of hybrid systems for which the average is minimal. The minimality of the average system is crucial if the hybrid system is to be stabilized via a nonswitching feedback gain. Furthermore, this result sheds some light on the topology of minimal LTI systems in parameter space.

The above theorem can, probably, be generalized to the case where all forms are simultaniously transformable to jordan canonical forms with the additional condition that the geometric multiplicities of all eigenvalues of the avarage system are equal to one.

It is, again, obvious that the necessary part of the controllability criterion of the hybrid system plays an important role in the controllability of the average system. This need for this condition makes it still more important to be studied closely.

The issue of the minimality of average systems can be studied in a more systematic way by defining an appropriate topology on the parameters space; works such as Cobb's paper[16], Eising's paper[17], and references therein can provide a good start in this direction. Nevertheless, the paper's results point to directions that can be helpful in defining such topology. Moreover, the last theorem tells us that given a set of simultaniously diagonalizable systems then every element in the set of all averages, as defined in this paper, is controllable. This is an interesting result in its own right.

Another research direction concerning the minimality of the average system is the application of the results in Anderson's paper[18] where structural controllability and matrix nets are studied.

REFERENCES

[1] T. L. Johnson, "Synchronous Switching Linear Systems," Proc. 24th. IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 1699-1700, Dec. 1985.

[2] R. W. Brockett and J. R. Wood, "Electrical networks containing controlled switches," in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, San Francisco, pp. 1-11, April 1974.

[3] J. A. Richards, "Analysis of Periodically Time-Varying Systems," Spring er-Verlag, 1983.

[4] Jalel Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems," Proc. 1988 American Control Conf., Atlanta, GA, June 1988.

[5] H. J. Chizeck, A. S. Willsky and D. Castanon, "Discrete-time Markovian jump Linear Quadratic Optimal Control," Int J. Control, Vol. 43, No. 1, pp. 213-231, 1986.

[6] D. D. Sworder, "Control of systems subject to sudden changes in character," Proc. IEEE, Vol. 64, No. 8, pp. 1219-1225, 1976.

[7] S. Geman, "Some Averaging and Stability Results for Random Differential Equations," SIAM J. App. Math., Vol. 36, No. 1, Feb.1979.

[8] R. L. Kosut, B. D. O. Anderson and I. M. Y. Mareels, "Stability Theory for Adaptive Systems: Method of Averaging and Persistency of Exitation," IEEE Trans. Automat. Cont., Vol. AC-32, NO. 1, JAN. 1987.

[9] Jalel Ezzine and A. H. Haddad, "On the Stabilization of Two-Form Hybrid Systems Via Averaging," Proc. 22nd Annual Conf. on Information Sciences and Systems, Princeton, March 1988.

[10] M. Mariton and P. Bertrand, "Non-switching Control Strategies for Continuous-time Jump Linear Quadratic Systems," Proc. 24th Conf.Decision and Control, Ft. Lauderdale, pp. 916-921, Dec. 1985.

[11] J, L, Doob, Stochastic Processes, John Wiley and Son, Inc., 1953.

[12] M, Mariton, "Controllability, Stability and Pole Allocation for Jump Linear Systems," Proc. 25th. IEEE CDC, Athen, Greece, pp. 2193-2194, 1986.

[13] M. Ikeda, H. Medea, and S. Kodama, "Stabilization of linear systems," SIAM J. Cont., Vol.10, No.4, pp. 716-729, 1972.

[14] C. D. Desoer and M. Vidyasagar, Feedback Systems: Input-Output Properties, Academic Press, 1975.

[15] W. L. Brogan, Modern Control Theory, Prentice-Hall/Quantum, 1982.

[16] J. D. Cobb, "On the Topology of Spaces of Controllable and Observable Systems," IEEE Trans. Automat. Cont., Vol. AC-31, No. 6, June 1986.

[17] R. Eising, "The Distance Between a System and the Set of Uncontrollable Systems," Proc. MTNS, Beer-Sheva, Israel, June 1983.

[18] B. D. O. Anderson, and Hui-Min Hong, "Structural Controllability and Matrix Nets," Int. J. Control, Vol. 35, No. 3, pp. 397-416, 1982.

APPENDIX E

J. Ezzine and A. H. Haddad, "On Largest Lyapunov Exponent Assignment and Almost Sure Stabilization of Hybrid Systems", Proc. 1989 American Control Conference, Pittsburgh, PA, pp. 805-810, June 1989.

ON LARGEST LYAPUNOV EXPONENT ASSIGNMENT AND ALMOST SURE STABILIZATION OF HYBRID SYSTEMS[1]

Jelel Ezzine

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

A. H. Haddad

Department of EE/CS
Northwestern University
Evanston, IL 60208

ABSTRACT

This paper develops techniques for the analysis, control and stabilization of hybrid systems. These systems switch among a finite set of linear time-invariant models with switching behavior governed by a Finite State Markov Chain. The relationship of these techniques to standard methodologies for linear time-invariant systems is also considered.

1. INTRODUCTION AND PROBLEM FORMULATION

1.1 INTRODUCTION

Many real systems such as power systems [1-3] exhibit variations in their structures or abrupt changes in their inputs or internal variables and other system parameters. Standard linear time-invariant systems models can not adequatly represent these systems. Consequently, a new class of systems has been proposed to model such systems [4-14]. This class is called hybrid systems due to the existence of both discrete and continuous variables in their state space. Such systems have been explicitly or implicitly used in past research [1].

Systems of this type can be used to model networks with periodically varying switches [15,16], synchronously switched linear systems [17], multi-rate Sampled-data systems [18-22], systems subject to failures [1,23-25], manufacturing systems [24,25], large scale flexible structures [13], and last but not least macroeconomic models [5].

The objective of this paper is to develop methodologies or, at least, to provide the necessary foundations for the analysis and design of such systems, with emphasis on their stabilization. These design tools are expected to aid in the design of controllers to stabilize such systems and to achieve reliable performance despite the changes.

It is customary to assume that these systems switch among a finite set of linear models according to an irreducible FSMC. Hence, the approach is based on the mathematical theory of ergodic stochastic processes. Interestingly enough, the concepts of eigenvalues and eigenspaces are generalizable within the ergodic theory framework. Therefore, the key design idea of eigenvalues assignment for hybrid systems remains meaningful despite the time-variation and random nature of these systems.

Earlier major works on the subject exhibit two important points. The first one is the almost exclusive usage of Stochastic Dynamic Programming (SDP) as a tool to adress the optimal control and stabilization of hybrid systems. The second one is the difficulties related to the solution (i.e., existence and uniqueness issues) of the coupled Riccati-like equations derived via SDP. These equations play the same role played by the familiar Riccati equations in LQR theory.

SDP is a very convenient tool to solve optimization problems. However, besides the well known "curse of dimentionality", the systematic application of SDP does not allow the user to gain insight about this complex problem. In general the use of SDP obscures most of the useful properties of these systems. These properties can reveal the geometric and algebraic structures of hybrid systems. As a matter of fact, many of the results presented in [26], [14], and [27] show that these systems, despite their time-variation and random nature, share many useful properties with LTI systems.

Our approach avoids the difficulties faced by previous researchers and exploits the similarities between LTI systems and hybrid systems. In particular, due to the simplicity and success of the eigenvalues assignment design technique in the stabization of LTI systems, we will follow a similar approach.

In order to address the stabilization problem of hybrid systems one needs a simple stability criterion for this class of systems. Hence, some conditions for determining the stability of hybrid systems are first developed based on a generalization of the eigenvalues concept. The stabilization approach uses the largest Lyapunov exponent along with some controllability properties of hybrid systems.

The following is an outline of the paper. After the problem formulation, Section 2 introduces the material needed to discuss the almost sure stabilization of hybrid systems. This section also addresses the stability of both continuous and discrete time hybrid systems and simple sufficient stability criteria are derived. In section 3 the almost sure stabilizability result is discussed. This result is a generalization of Wonham stabilization theorem to this class of systems. Section 4 concludes the paper.

1.2 PROBLEM FORMULATION

The hybrid systems considered in this paper are assumed to have the form

[1] This work is supported by the U.S. Air Force under grant AFOSR-87-0308.

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t) \quad (1.a)$$

$$y(t) = C(r(t))x(t) \quad (1.b)$$

where x is the system state vector of dimension n, u is the control input vector of dimension p, y is the output vector of dimension m, and r(t) is the "form index" which is either a deterministic or a stochastic scalar sequence taking values in the finite index set N={1, 2, ..., N}.

The system takes the realization $\Sigma_i=(A_i,B_i,C_i)$ when r(t) = i, with iεN. This realization is called the ith form. Let δt_i denote the time interval during which r(t) = i. In addition we define

$$T \equiv \sum_{i=1}^{N} \delta t_i \quad (2)$$

as the period of the system.

Sometimes it is more convenient to represent the hybrid system in an equivalent different form which leads to the following representation

$$\dot{x}(t) = \{ \sum_{i=1}^{N} v_i(t)A_i\}x(t) + \{ \sum_{i=1}^{N} v_i(t)B_i\}u(3)a)$$

$$y(t) = \{ \sum_{i=1}^{N} v_i(t)C_i\}x(t) \quad (3.b)$$

where $v_i(t)$ = 1 when the system is governed by the ith realization Σ_i, and $v_i(t)$ = 0 otherwise. The $v_i(t)$ function is called the ith indicator function. It is evident from the definition of hybrid systems that at any point in time only one of the N indicator functions takes the value one.

Most of our work will address the situation where r(t) is a stochastic process, more presisely, r(t) will be assumed to be a Finite State Markov Chaine (FSMC) with transition probabilities

$$P\{r(t+\delta t_i) = j | r(t) = i\} = p_{ij}, \quad (4.a)$$

for continuous-time systems, and

$$P\{r(t+1) = j | r(t) = i\} = p_{ij}, \quad (4.b)$$

for discrete-time systems. It is also assumed, unless stated otherwise, that the FSMC is stationary and irreducible [28].

Sometimes r(t) is defined as a special deterministic process. It will be obvious from the definition that r(t) is very similar to the FSMC defined above. This is done in order to show that a deterministic formulation is sufficient to answer certain questions. In this case it is assumed that any r(t) sequence is composed of a succession of N-termed blocks, where every block is a permutation of the index set **N**. It is important to note that the succession of the blocks is completely arbitrary (e.g., for N=3, a possible r(t)-sequence is: 123,321,213,213,312,...).

2. STABILITY OF HYBRID SYSTEMS

2.1 Lyapunov Exponents

In 1892, A. M. Lyapunov founded the theory of characteristic exponents that bear his name [29].

His intention was to determine criteria for the stability (of the origine x≡0) of

$$\dot{x} = A(t)x, \quad x(0; x_0) = x_0 \in R^n, \ t \in R^+, \quad (5)$$

A(t) is continuous and bounded.

For constant A the eigenvalues of A determine the stability behavior of (5). For periodic A(t) Floquet theory shows that the results for constant A remain true if the real parts of eigenvalues are replaced by the characteristic exponents of A(t) [30, chapters 3 and 13].

The Lyapunov exponent of a solution $x(t; x_0)$ is defined by

$$\lambda(x_0) \equiv \limsup_{t\to\infty} (1/t)\mathrm{Log}\|x(t; x_0)\|. \quad (6)$$

Lyapunov proved that for every solution with $x_0 \neq 0$, $\lambda(x_0)$ is finite. Moreover, the set of all possible numbers which are Lyapunov exponents of some nonzero solution of (5) is finite, with cardinality p, such that 1≤p≤n and $\lambda_p < \ldots < \lambda_1$.

Furthermore, Lyapunov proved that the subspaces

$$L_i \equiv \{x_0 \in R^n : \lambda(x_0) \leq \lambda_i, \ i=1,\ldots, p+1\}$$

form a filtration of R^n, i.e.,

$$0 = L_{p+1} \quad L_p \quad \ldots \quad L_1 = R^n,$$

with $\dim L_i = k_i$, such that

$$k_{p+1} = 0 < k_p < \ldots < k_1 = n,$$

and

$$\lambda(x_0) = \lambda_i \text{ iff } x_0 \in L_i \backslash L_{i+1}, \ i = 1, \ldots, p.$$

The numbers λ_i together with their multiplicities d_i are called the Lyapunov spectrum of (5). The asymptotic behavior of (5) is dictated by λ_1. That is, (5) is exponentially stable iff $\lambda_1<0$.

Unfortunatly, it is in general not true that $\lambda_1<0$ implies the stability of the following nonlinear system:

$$\dot{x} = A(t)x + f(t, x). \quad (7)$$

However, for a special class of f(t, x) the above is true if (5) is what Lyapunov calls regular. For regular systems the following holds

$$\lambda_1 = \lim_{t\to\infty} (1/t)\mathrm{Log}\|x(t; x_0)\|. \quad (8)$$

For example, (5) is regular if A is constant or periodic. In the latter case λ_i are the characteristic exponents.

Regularity is hard to verify for a particular system, but it happens with probability one in many cases involving a flow with an invariant probability measure [31]. This is how Birkoff's ergodic theorem [32] and ergodic theory in general comes in to exploit Lyapunov powerful spectral theory.

As shown above there are many similarities between eigenvalues and eigenspaces of constant matrices and the Lyapunov spectral theory of time-varying systems. The main similarity that will be of major importance in this paper is the stability role played by the largest Lyapunov exponent λ_1; λ_1 plays the same role as the largest eigenvalue plays in the stability of time-invariant systems.

2.2 Continuous-time Hybrid Systems

Eventhough the sign of λ_1 is a necessary and sufficient test for the stability of a hybrid system, it is almost impossible to compute. In this section we will use the Lyapunov exponent along with the logarithmic norm concept to derive a simple sufficient stability test for continuous-time hybrid systems.

In order to derive uncomplicated conditions for the stability of such systems, a different tool is used, namely the logarithmic norm [33,34], resulting in a simpler sufficient condition.

The logarithmic norm (also known as the logarithmic derivative, the measure of a matrix) was introduced in 1958 separately by Dahlquist [33] and Lozinskij [34] as a tool to study the growth of solutions to ordinary differential equations and the error growth in discretization methods for their approximate solution. It is formally defined as follows:

Definition 1

The logarithmic norm of a matrix A associated with the matrix norm $\|.\|$ is defined by

$$\mu(A) = \lim_{h\to 0^+} (\|I + hA\| - 1)/h \qquad (9)$$

Explicit expression for the logarithmic norm associated with the Euclidean norm is

$$\mu(A) = \max\{\mu : \mu \, \varepsilon \, \lambda((A+A^*)/2)\}. \qquad (10)$$

Then the following inequality is true:

$$Exp(-\mu(-A)t) \le \|Exp(At)\| \le Exp(\mu(A)t). \qquad (11)$$

One very important property of the logarithmic norm follows from the fact that it may be shown to be the smallest element of

$$S = \{s : \|Exp(At)\| \le Exp(st), t\ge 0\}. \qquad (12)$$

Therefore it gives an optimal bound on the exponential behavior of $\|Exp(At)\|$ for $t\ge 0$. It may be concluded therefor that

$$\sup_{t\ge 0}\|Exp(At)\| = 1 \text{ iff } \mu(A)\le 0. \qquad (13)$$

In the case where A is normal square matrix (i.e., $A^*A=AA^*$), then

$$\|Exp(At)\| = Exp(\alpha(A)t) = Exp(\mu(A)t)(14)$$

where $\alpha(A)$ is the maximal real part of the eigenvalues of A. This norm is now used to derive the stability condition.

Theorem 1

For the null solution of the hybrid system (1) to be a.s. exponentially stable, it is necessary that

$$\sum_i \mu(-A_i)p_i > 0, \qquad (15.a)$$

and sufficient to have

$$\sum_i \mu(A_i)p_i < 0, \qquad (15.b)$$

where the p_i's are the steady-state probabilities of the irreducible FSMC and $i\varepsilon N$.

The simple sufficient condition states that for a hybrid system to be a.s. uniformly asymptotically stable the average of the logarithmic norms of each realization has to be negative. Therefore, this sufficient condition allows for unstable forms. That is, as long as the stable forms dominate, the overal system is a.s. exponentially stable. This domination can occure in two ways: either the stable forms are strongly stable (i.e., highly negative logarithmic norms) or the time span of the stable forms is large relative to the time span of the unstable ones or a combination of both reasons. This stability feature of hybrid systems was reported in [23] via examples. The critera given above holds for hybrid systems where the switching is deterministic as well, in which case the p_i's represent relative time span of the forms.

It is clear from the proof of the above theorem that the sufficient almost sure exponential stability criterion is an upper bound for the largest Lyapunov exponent λ_1, and is simple to compute. Therefore, this sufficient criterion can play a beneficial role in testing for the stability of hybrid systems as well as in the design of controllers to stabilize such systems, as stated in the following:

Theorem 2

The largest Lyapumov exponent λ_1 of the null solution of the hybrid system (1) satisfies the following inequality

$$\lambda_1 = \lim_{t\to\infty} (1/t)Log\|x(t; x_o)\| \le \sum_{i\varepsilon N} p_i\mu(A_i), \qquad (16)$$

where the p_i's are the steady-state probabilities of the irreducible FSMC.

The differance between the upper and lower bounds given in theorem 1 give a measure of the conservativeness of (15.a). This problem is considered further in the sequel.

2.3 Discrete-time Hybrid Systems

The study of the stability, sample-wise, of discrete-time hybrid systems is similar to the study of the solutions of stochastic linear difference equations with randomly varying parameters. This lead us to the study of the following problem:

Let $\{M_n, n \, \varepsilon \, N\}$ be a sequence of random, nxn, matrices. To each $x_o \, \varepsilon \, R^n$ one associates the process $\{X_n, n \, \varepsilon \, N\}$ with values in R^n, which is the solution to

$$X_{n+1} = M_nX_n, \quad n \, \varepsilon \, N, \text{ and } X_o = x_o. \qquad (17)$$

We have $X_{n+1} = M_n...M_1x_o$. One important question is what is the asymptotic behavior of this process.

Furstenberg-Kesten Theorem [35]

Let $\{M_i,\ i \in N\}$ be a stationary, metrically transitive (i.e., ergodic) stochastic process with values in the set of nxn matrices such that $E\{\mathrm{Log}^+\|M_0\|\}<\infty$. Then, with probability one

$$\lim_n (1/n)E\{\mathrm{Log}\|S_n\|\} = \inf_n (1/n)E\{\mathrm{Log}\|S_n\|\}$$

$$= \overline{\lim_n}(1/n)\mathrm{Log}\|S_n\| \equiv \lambda_1 \in R\cup\{-\infty\}, \qquad (18)$$

where $S_n \equiv M_n ... M_1$. λ_1 is the largest Lyapunov exponent of the process.

Remark

The Furstenberg-Kesten Theorem (FKT) is a generalization of Birkoff's ergodic theorem to the case of matrix valued functions.

The largest Lyapunov exponent λ_1 of a discrete-time hybrid system plays the same role as the one played by the largest Lyapunov exponent of continous-time hybrid systems. That is, for a discrete-time hybrid system to be a.s. uniformly asymptotically stable it is necessary and sufficient to have $\lambda_1<0$. However, it would be very impractical to use (18) to compute λ_1 (i.e., compute (18)). Therefore, one way to avoid this difficulty is to use the Furstenberg-Kesten Theorem to derive simpler criteria to test for the a.s. stability of the system.

As it is stated in the FKT, λ_1 is an infimum of a particular set. Consequently, if any of the elements of this set is negative one concludes, using the property of the infimum, that the system is a.s. exponentially stable. However, by not knowing exactly λ_1, it is not possible to tell how stable the system is. That is, by exploiting this property to alleviate the computational burden, we are loosing some qualitative insight about the dynamical behavior of the system. This qualitative insight can be crucial for application purposes. First we need additional definitions.

Definition 2

A set $S=\{H_i,\ i=1, ...,N\}$ of nxn matrices is nilpotent provided there is a a k-termed sequence $\{H_i\}$, such that k is finite and the matrix

$$\Pi_k \equiv H_k \dots H_2H_1 \qquad (19)$$

is nilpotent. The least number θ for which the power of the matrix Π_k is null is called the index of nilpotency.

In case S is a singleton the above definition is identical to the usual definition of nilpotency.

Proposition 1

If S contains a nilpotent matrix then the set is nilpotent and its index of nilpotency θ is less or equal to the index of nilpotency of the nilpotent matrix.

Definition 3

The set $S=\{H_i,\ i=1, ..., N\}$ of nxn matrices is said to be contractive provided that there is a k-termed finite sequence $\{H_i\}$ and a norm, such that

$$\|H_k \dots H_2H_1\| \le \alpha < 1. \qquad (20)$$

At this point we would like to recall a theorem that will play a key role in the sequel. This theorem is the converse of a well known result relating the norm of a matrix to its spectral radius (i.e., $\alpha(A) \le \|A\|$). The following theorem asserts that there exists an induced norm for which the inequality in the previous result can be reversed after adding an arbitrarily small positive number to the matrix spectral radius.

Theorem [36]

For any $\varepsilon>0$ and any nxn real A matrix, there is a (vector) norm on R^n such that the corresponding induced norm satisfies

$$\|A\| \le \alpha(A) + \varepsilon. \qquad (21)$$

Now we are ready to apply the above definitions and theorems to derive simple sufficient a.s. exponential stability tests for discrete-time hybrid systems. As mentioned earlier these results will, in general, answer the stability issue but will not provide enough information about the qualitative behavior of the system. That is, how fast or how slow the system is converging or diverging. However, the first result is an exact result (i.e., exact λ_1). It allows a complete quantitative and qualitative analysis in an important special case.

Theorem 3

A homogeneous N-form hybrid system with a stationary irreducible FSMC, is a.s. exponentially stable with $\lambda_1=-\infty$, provided that the set **N** contains a nilpotent set.

The first result says that a homogeneous N-form hybrid system is a.s. stable, and its $\lambda_1=-\infty$, if the set of N-matrices of the hybrid system contains a nilpotent set.

This result is the analog of the stability result of a homogeneous difference equation with a nilpotent matrix. These difference equations converge to zero in no more than n steps. That is, by analogy, hybrid systems with a nilpotent set of matrices are very fast systems. This is confirmed by the fact that $\lambda_1=-\infty$.

The next result is more general but less powerful in the sense that it does not provide us with λ_1.

Theorem 4

A homogeneous N-form hybrid system with a stationary irreducible FSMC, is a.s. exponentially stable, provided that the set **N** contains a contractive set.

As an attempt to alleviate the shortcomings of the latter theorem we provide an upper bound for λ_1 similar to the one given for continuous-time hybrid systems. However, this bound does not involve the logarithmic norm concept, but it is derived via a simple computation.

Theorem 5

The LSR of a homogeneous N-form hybrid system with a stationary irreducible FSMC satisfies the following inequality

$$\lambda_1 \le \sum_{i=1}^{N} p_i \mathrm{Log}\|A_i\|. \qquad (22)$$

The next result is based on the work of Katz and Thomasian [37]. Along with the FKT it provides a mean by which one can estimate the least number of matrix multiplications in the λ_1-formula for which the probability of having a large error is minimized.

Theorem 6

Given a homogeneous N-form hybrid system with a stationary and irreducible FSMC, a positive integer m and $\varepsilon>0$. Then

$$P\{|(1/n)\sum_{i=1}^{N}\mathrm{Log}\|A_i\| - LSR^+|\geq\varepsilon \text{ for some } n\geq m\}$$
$$\leq 2\alpha\mathrm{Exp}\{-\beta\varepsilon^2 m\}, \qquad (23)$$

where

$$\beta = p^{3N}/2^8\delta^2N^2, \qquad (24.a)$$

$$\alpha = 8N/p^N(1 - \mathrm{Exp}\{-\beta\varepsilon^2\}), \qquad (24.b)$$

with

$$\delta = \max_{i\varepsilon N}\mathrm{Log}\|A_i\| - \min_{j\varepsilon N}\mathrm{Log}\|A_j\|, \qquad (25.a)$$

$$LSR^+ = \sum_{i\varepsilon N} p_i\mathrm{Log}\|A_i\|. \qquad (25.b)$$

and p is the largest entry in the steady-state probability vector.

This result tells us when to stop the time average of the λ_1 upper bound and still get a good approximation. This might seem useless since we have a simple expression for LSR^+. However, the LSR^+ is an upper bound for λ_1, therefore, this stopping rule can be used as an approximate stopping rule in computing the LSR.

3. ALMOST SURE STABILIZATION OF HYBRID SYSTEMS

3.1 Lyapunov SPECTRAL RADIUS ASSIGNMENT

In this section we provide the main result of the paper; a simple sufficient a.s. stabilization theorem.

Theorem 7

An m inputs n states N-form hybrid system with a stationary and irreducible FSMC is a.s. stabilizable by arbitrarily assigning its λ_1 provided that there is a completely reachable K-periodic system, with m inputs and n states, embedded in the N forms, with $K\leq N$.

This result is based on the pole-assignment for discrete-time linear periodic system. Hernandez and Urbano [38] extended the pole assignment technique to linear periodic systems. Their result is used here to extend the pole assignment to stochastic hybrid systems by assigning the largest Lyapunov exponent λ_1.

It is possible under special assumptions to arbitrarily assign λ_1 without requiring the complete reachability of a periodic system. That is, by imposing a geometric relation among the N forms. This way the complete reachability condition is weakened considerably as it is stated in the next theorem

Theorem 8

Given a N-form hybrid system with a stationary and irreducible FSMC such that

a. Rank $[C_1, C_2, \ldots, C_N] = n$,

b. $\{A_i-B_iK_i,\ i\varepsilon N\}$ belongs to a solvable Lie algebra,

then the hybrid system can be made exponentially stable with $\lambda_1=-\infty$ a.s.

Actually what the theorem says is that the Lyapunov spectral radius can always be made negative, more precisely $-\infty$, if the two conditions of the theorem are met. The first condition, as discussed several times in [26], is a necessary condition for the hybrid system to be controllable, therefore, it is the weakest controllability condition. This condition is the strongest part of this theorem. However, the second condition is quite restrictive and maybe impractical.

4. CONCLUSION

The a.s. exponential stability criteria presented in this paper are simple to compute, consequently they alleviate the computational shortcomings of Lyapunov exponents. However, these tests are only sufficient and they can be quite conservative, hence they require further study.

The a.s. stabilizability theorem can be viewed as a generalization of Wonham stabilization theorem. Actually what the theorem says is that the Lyapunov spectral radius can always be made negative with probability one, more precisely $-\infty$, if certain conditions are met.

One additional problem of interest is to find wether the eigenspaces idea carries over to hybrid systems and its usefulness.

REFERENCES

[1] A. S. Debs, and A. R. Benson, "Security assessment of power systems," Proc of the Engi. Foundat. Conf. on Systems Engineering for Power: Status and Prospects, Publication No. CONF-750867, Henniker, N. H., 1975.

[2] A. S. Willsky, and B. C. Levy, "Stochastic stability research for complex power systems," DOE Contract Report, LIDS, MIT, 1979.

[3] J. Zaborsky, and co-authors, "Towards a comprehensive analysis and operating practice of the laege compound HV-AC-DC system," Final report to U.S. DOE, Washingto Univ., St-Louis, Missouri, 1982.

[4] N. N. Krasovskii, and E. A. Lidskii, "Analytical designs of controllers in systems with random attributes I, II, III," Automation and Remate Contr., Vol. 2, Academic Press, New York, 1962.

[5] T. Kazangey, and D. D. Sworder, D. D., "Effective federal policies for regulating residential housing," in Proc. Summer Computer Simulation Conf., pp. 1120-1128, 1971.

[6] D. D. Sworder, "Feedback control of a class of linear systems with jump parameters," IEEE Trans. Auto. Contr., Vol. AC-14, No. 1, pp. 9-14, 1969.

[7] J. D. Birdwell, D. A. Castanon, and M. Athans, "On reliable control system designs with and without feedback reconfigurations," Proc. IEEE Conf. on Dec. and Contr. Theory, 1979

[8] H. J. Chizeck, and A. S. Willsky, "Towards fault-tolerant optimal control," IEEE Conf. on Deci. and Contr. Theory, 1978.

[9] G. S. Ladde, and D. D. Siljak, "Multiplex control systems: Stochastic stability and dynamic reliability," Int. J. Contr., Vol. 38, No. 3, pp. 515-524, 1983.

[10] M. Mariton, and P. Bertrand, "Outputfeedback for a class of linear systems with stochastic jump parameters," IEEE Trans. Auto. Contr., Vol AC-30, No. 9, pp. 898-903, 1985.

[11] M. Mariton, and P. Bertrand, "Robust jump linear quadratic control: a mode stabilizing solution," IEEE Trans. Auto. Contr., Vol. AC-30, No. 11, pp. 1145-1149, 1985.

[12] M. Mariton, and P. Bertrand, "Non-switching control strategies for continuous-time jump linear quadratic systems," Proc. 24th Conf. Decision and Control, Ft. Lauderdale, pp. 916-921, 1985.

[13] M. Mariton, "Controllability, stability and pole allocation for jump linear systems," Proc. 25th. IEEE Conf. Decision and Control, Athens, Greece, pp. 2193-2194, 1986.

[14] M. Mariton, "Stochastic controllability of linear systems with mrkovian jumps," Automatica, Vol. 23, No. 6, pp. 783-785, 1987.

[15] R. W. Brockett, and J. R. Wood, "Electrical networks containing controlled switches," in Applications of Lie groups theory to nonlinear networks problems, Supplement to IEEE International Symposium on Circuit Theory, San Francisco, pp. 1-11, 1974.

[16] Y. Sun, "Networks containing periodic switches: A unified approach and applications," Ph. D. Thesis, Dept. of Elect. Eng. and Comp. Scien., Univ. of Calif. Berkeley, 1967.

[17] T. L. Johnson, "Synchronous Switching Linear Systems," Proc. 24th IEEE Conf. Decision and Control, Ft. Lauderdale, FL, pp. 1699-1700, 1985.

[18] J. Tokarzewski, "Sufficient stabilizability conditions for multirate sampled-data systems," INT. J. Contr., Vol. 39, No. 2, pp. 257-277, 1984.

[19] D. P. Stanford, and L. T. Conner JR., "Controllability and stabilizability in multi-pair systems," SIAM J. Cont., Vol. 18, No. 5, pp. 488-497, 1980.

[20] D. P. Stanford, "Stability for multi-rate sampled-data system," SIAM J. Cont., Vol. 17, No. 3, pp. 390-399, 1979.

[21] R. E. Kalman, "Control of randomly varying linear dynamical systems," Proc. Sympos. Appl. Math., Amer. Math. Soc., Vol. 13, Providence, R.I., pp. 287-298, 1960.

[22] W. L. De Koning, "Stationary optimal control of stochastically sampled continuous-time systems," Automatica, Vol. 24, No. 1, pp. 77-79, 1988.

[23] H. J. Chizeck, A. S. Willsky, and D. Castanon, "Discrete-time Markovian jump Linear Quadratic Optimal Control," Int J. Control, Vol. 43, No. 1, pp. 213-231, 1986.

[24] R. Akella, and P. R. Kumar, "Optimal control of production rate in a failure prone manufacturing system," IEEE Trans. Automat. Contr., Vol. AC-31, No. 81, pp. 116-126, 1986.

[25] R. S. Ratner, and D. G. Luenberger, " Performance-adaptive renewal policies for linear systems," IEEE Trans. Auto. Contr., Vol. AC-14, pp. 344-351, 1969.

[26] J. Ezzine, and A. H. Haddad, "On the controllability and observability of hybrid systems," Proc. 1988 American Control Conference, Atlanta, GA, pp. 41-45, June 1988.

[27] J. Yuandong, and H. J. Chizeck, "Controllability, observability and continuous-time marcovien jump linear quadratic control," Preprint.

[28] J. L. Doob, Stochastic Processes. JohnWiely, New York, 1953.

[29] A. M. Lyapunov, "Problème général de la stabilité du mouvement," Reprint Ann. of Math. Studies, Vol. 17, Princeton Univ. Press 1949, Princeton.

[30] E. A. Coddington, and N. Levinson, Theory of Ordinary Differential Equations. Mc Graw-Hill, New York, 1955.

[31] V. I. Oseledec, "A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems," Trans. Moscow Math. Soc., Vol. 19, pp. 197-231, 1968.

[32] A. I. Khinchin, Mathematical Foundations of Statistical Mechanics. Dover, NY, 1949.

[33] C. Van Loan, "The sensitivity of the matrix exponential," SIAM J. Num. Anal., Vol. 14, No. 6, pp. 971-981, December 1977.

[34] T. Strom, "On logarithmic norms," SIAM J. Num. Anal., Vol. 12, No. 5, pp. 741-753, October 1975.

[35] H. Furstenberg, and R. Kesten, "Products of random matrices," Ann. Math. Statist., Vol. 31, pp. 457-469, 1960.

[36] C. D. Desoer, and M. Vidyasagar, Feedback Systems: Input-Output Properties. Academicress, 1975.

APPENDIX F

B. S. Heck and A. H. Haddad, "Singular Perturbation in Piecewise Linear Systems", Proc. 1988 American Control Conference, Atlanta, pp. 1722-1727, June 1988. Also in IEEE Transactions on Automatic Control, vol. 34, pp. 87-90, January 1989.

SINGULAR PERTURBATION IN PIECEWISE-LINEAR SYSTEMS[1]

B.S. Heck and A.H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

This paper analyzes piecewise-linear systems which are singularly perturbed. A technique is developed that allows decoupling of such systems into fast and slow subsystems for analysis and design. The results of a numerical example are included to demonstrate this technique.

1. INTRODUCTION

Piecewise-linear systems which are singularly perturbed are found in many applications including electrical circuits and flight controls. The piecewise-linearity may be due to nonlinear elements such as saturation or may result from a linearization about various operating points of a nonlinear plant. These types of systems are numerically very stiff and, hence, are difficult to analyze. This problem may be alleviated by using singular perturbation theory to separate the system into reduced-order models, one containing the slow dynamics and one containing the fast dynamics. Reduced-order models are easier to use in analysis and design by lessening the computation complexity. In addition, time-integration of the lower order systems instead of the full order model reduces computation time since a larger time step can be used for the slow dynamic model. The use of standard singular perturbation techniques, however, requires that the system dynamical equations be smooth [1,2] ruling out their use on piecewise-linear systems. This paper extends the general method of singular perturbation for application to continuous piecewise-linear systems.

1.1 Problem formulation

The system considered in this paper may be represented in the following form:

$$\dot{x} = f_1(x,z), \qquad x(t_0) = x_0 \tag{1}$$

$$\mu\dot{z} = f_2(x,z), \qquad z(t_0) = z_0 \tag{2}$$

where: f_1 and f_2 are continuous piecewise-linear functions, $\mu>0$ is a small parameter, and $x \varepsilon R^p$ and $z \varepsilon R^r$. The functions are affine in specific regions of the state space (R^{p+r}) where a region is typically defined as an intersection of half-spaces. For example, equations (1) and (2) are represented in the i^{th} region by the following "linear" system:

$$\dot{x} = A_{11}{}^{i}x + A_{12}{}^{i}z + w_1{}^{i} \tag{3}$$

$$\mu\dot{z} = A_{21}{}^{i}x + A_{22}{}^{i}z + w_2{}^{i} \tag{4}$$

For the purposes of this paper, the i^{th} region is defined by the set $S_i=\{(x,z)\colon d_{i-1}< K_x x+K_z z \le d_i\}$ where K_x and K_z are row vectors and $d_{i-1} < d_i$ are scalars. By this definition, the type of regions allowed are parallel in that the boundaries do not intersect. An example of a physical system which has this description is one in which the piecewise-linear element is in a scalar feedback loop. The reason for the restriction will be discussed in Section 2.

The system given in equations (1) and (2) contains both fast and slow dynamics. The variable x is primarily slow while z has both fast and slow components. Starting from the initial conditions of equations (1) and (2), the fast part of z quickly dies out and z converges to a quasi-steady-state value (i.e., the slow component) in a short time interval $[t_0,t_0+\delta)$ known as the boundary layer. The fast component of z is then known as the boundary layer solution. The solution of the system outside of the boundary layer is termed the outer solution. It is desired to decouple system (1)-(2) into fast and slow models which yield the boundary layer solution and the outer solution, respectively. The boundary layer solution is then used as a correction term to the outer solution so that the combination is an approximation for the original system with errors of order $O(\mu)$. A technique to decouple the system is developed in this paper.

The following is an outline of the paper. Section 2 discusses the boundary layer solution and developes a reduced-order model to approxi-

[1] This research is supported by the U.S. Air Force under contract F08635-84-C-0273 (with the Armament Laboratory) and AFOSR-87-0308.

mate this solution. The outer solution along with a corresponding reduced-order model is discussed in Section 3. A numerical example is presented in Section 4 to demonstrate the techniques developed in this paper. Section 5 concludes the paper.

2. BOUNDARY LAYER SOLUTION

The fast dynamics of the system are most prominant during the boundary layer and can be decoupled from the slow dynamics by introducing an expanded time scale $\tau = (t-t_0)/\mu$. Examination of equation (1) shows that x stays relatively constant with respect to τ assuming that $A_{11}{}^i$, $A_{12}{}^i$ and $w_1{}^i$ are bounded in all regions S_i [1]. Equation (2) may be rewritten as follows:

$$\frac{d\hat{z}}{d\tau} = \hat{f}(\hat{z}) \qquad (5)$$

where $\hat{z}(\tau)=z(\mu\tau+t_0)$ and $\hat{f}(\hat{z})=f_2(x_0,\hat{z})$. The function $\hat{f}$ is a continuous piecewise-linear mapping from R^r into R^r. The state space in R^r is partitioned into regions where the function is affine; e.g., the i^{th} region is defined as the set $R_i=\{z\colon\ d_{i-1} < K_x x_0+K_z z \le d_i\}$. A degenerate case where $K_z=0$ results in the existence of only one region in R^r so that $\hat{f}$ is affine everywhere. The initial quasi-steady-state value, $z_s(t_0)$, of z(t) is a stable equilibrium point of (5). Note that the equilibrium point of the degenerate case is easily found.

The equilibrium point(s) of (5) for the nondegenerate case can be found using solution techniques developed for piecewise-linear resistive networks. Many papers have been written on finding the solution x of the equation f(x)=y where f is a continuous piecewise-linear function, e.g. [3-9]. Fujisawa and Kuh show in [4] that a continuous piecewise-linear function satisfies a Lipshitz condition. The following theorem from [4] gives sufficient conditions for the existence and uniqueness of the solution.

Theorem 1: Let f be a continuous piecewise-linear mapping of R^r into itself and let $J^i{}_k$ denote the matrix composed of the first k rows and columns of the Jacobian matrix J^i in region R_i. The mapping is a homeomorphism of R^r onto itself if, for each k=1,2,...,r, the determinants of the kxk matrices

$$J^1{}_k, J^2{}_k, \ldots, J^r{}_k$$

do not vanish and have the same sign.

This previous work is used in finding the equilibrium point(s) of system (5) by solving $\hat{f}(\hat{z})=0$. In this application, $J^i = A_{22}{}^i$ and each $A_{22}{}^i$ is assumed to be Hurwitz for stability purposes. The conditions of Theorem 1 may be stringent and various other sufficient conditions for the existence and uniqueness of the solution are given in [9-11]. Also, reference [12] discusses nonunique solutions.

2.1 Algorithm to Solve for Equilibrium Point

The Katzenelson algorithm is widely used insolving for x in the equation

$$f(x) = y \qquad (6)$$

where $f\colon R^r \rightarrow R^r$ is continuous and piecewise-linear. The basic outline of this algorithm used in solving $\hat{f}(\hat{z})=0$ is given below. More details of the general method are given in [4]. Let $W^j = A_{21}{}^j x_0 + w^j$ $\forall j$, and denote the iteration number on z_s and λ by superscripts.

```
0) initialize by letting $i=1$ and $z_s{}^i=z_0$
1) solve $z = -(A_{22}{}^j)^{-1}W^j$, where region $R_j$ contains $z_s{}^i$
2) if z lies in region $R_j$ then $z_s = z$ and stop
3) otherwise, let $R_k$ be the region containing z;
   if $k>j$ then $d = d_j$ and then let $j=j+1$
   if $k<j$ then $d = d_{j-1}$ and then let $j=j-1$
4) solve $\lambda^i = (K_z z_s+K_x x_0-d)/K_z(z_s{}^i-z)$
5) solve $z_s{}^{i+1} = z_s{}^i - \lambda^i(z_s{}^i-z)$
6) let $i=i+1$ and go to 1)
```

It is shown in [4] that if the piecewise-linear function is a homeomorphism (e.g., it satisfies the conditions of Theorem 1) then the algorithm will converge in a finite number of steps.

2.2 Boundary Layer Approximation

A fast model approximating the dynamics occurring in the boundary layer can be found once the equilibrium point of system (5) is known. The boundary layer solution is then given as $\hat{z}_f(\tau) = \hat{z}(\tau)-z_s(t_0)$. In this application, z_s must be found implicitly because f_2 is not smooth. Therefore, the fast model approximating the boundary layer solution is given in terms of $\hat{z}$. In the i^{th} region the fast model is given by

$$\frac{d\hat{z}}{d\tau} = A_{21}{}^i x_0 + A_{22}{}^i \hat{z} + w_2{}^i, \quad \hat{z}(0)=z_0 \qquad (7)$$

$$\hat{z}_f(\tau) = \hat{z}(\tau) - z_s(t_0)$$

where the i^{th} region is defined by the set $R_i=\{\hat{z}\colon d_{i-1} < K_x x_0 + K_z \hat{z} \le d_i\}$.

For the purposes of this paper, it is assumed that there exists exactly one equilibrium point which is asymptotically stable. Multiple stable equilibrium points may be handled by partitioning the state space into domains of attraction for the various equilibrium points and the analysis in this paper holds for each domain of attraction.

Asymptotic stability is assumed in this system though there is no known general method for determining asymptotic stability of piecewise-linear systems. Depending on the specific system under consideration, a Lyapunov function may be found. Another possibility is to use standard SISO frequency domain techniques or hyperstability. For using hyperstability notions, system (5) may be rewritten as

$$\frac{d\hat{z}}{d\tau} = A\hat{z} + Bu \qquad (8)$$

where A is chosen to be stable, B is the identity I, and u is defined in the i^{th} region to be $u = \Delta A^i\hat{z} + A_{12}{}^i x_0 + w_2{}^i$ where $\Delta A^i = A_{22}{}^i - A$. If the nonlinearity in the feedback loop satisfies the Popov integral inequality, then the necessary and sufficient condition for asymptotic stability is that the transfer matrix $(sI-A)^{-1}$ must be strictly positive real [13].

The errors in this approximation, which are of order $O(\mu)$, are due to the substitution of x_0 for x in (7) and in the definition of the regions. Substituting $x = x_0 + O(\mu)$ in (7) and in R_i yields the system

$$\frac{d\tilde{z}}{d\tau} = A_{21}{}^i(x_0+O(\mu)) + A_{22}{}^i\tilde{z} + w_2{}^i, \quad \tilde{z}(0)=z_0 \qquad (9)$$

$$R_i = \{\tilde{z}: d_{i-1}+O(\mu) < K_x x_0+K_z\tilde{z} \leq d_i+O(\mu)\}$$

where $\tilde{z}$ represents the actual response. In the interior of any particular region, both the approximation and the actual model are linear. Previous results on singular perturbation theory in linear systems show that if $\tilde{z}(\tau')=\hat{z}(\tau')+O(\mu)$ then $\tilde{z}(\tau'')=\hat{z}(\tau'')+O(\mu)$ for $\tau''>\tau'$ as long as both $\tilde{z}$ and $\hat{z}$ stay within the region. The problems that may arise due to a boundary crossing are eliminated if the class of systems allowed is restricted to those in which the vector field intersects a boundary hyperplane at a large enough angle (i.e. $O(\mu^0)$). In these systems if either $\tilde{z}$ or $\hat{z}$ crosses into another region, the other must also cross into that region. The resulting error in the approximation remains of order $O(\mu)$. These conditions are summarized in the following theorems. Note that the restriction placed on the class of systems is sufficient and not necessary for proving that the approximation error is of order $O(\mu)$.

Theorem 2: Let the vector field near a boundary at $d_i=K_z\tilde{z}+K_x x_0+O(\mu)$ in the space R^r be given by

$$f(\tilde{z}) = A_{21}{}^i\,(x_0+O(\mu)) + A_{22}{}^i\,\tilde{z} + w_2{}^i. \qquad (10)$$

Assume that $f(\tilde{z})$ does not vanish near the boundary. If $f(\tilde{z})$ intersects the boundary with an angle of order $O(\mu^0)$, then the difference between the solutions of (7) and (9) is $O(\mu)$.

Proof: Assume $\tilde{z}$ crosses the d_i boundary at τ' and $\hat{z}$ has not crossed yet. Prior to crossing $\tilde{z} = \hat{z} + O(\mu)$. The normal vector of the boundary hyperplane is given by $n=K_z{}^T/\|K_z\|$. Since $f(z)\cdot n = O(\mu^0)$, then

$$K_z(A_{21}{}^i\,(x_0+O(\mu)) + A_{22}{}^i\,\tilde{z} + w_2{}^i) = O(\mu^0). \qquad (11)$$

It follows that

$$K_z(A_{21}{}^i\,x_0 + A_{22}{}^i\,\hat{z} + w_2{}^i) = O(\mu^0) \qquad (12)$$

Define $\hat{s}$ and $\tilde{s}$ by

$$\hat{s} = K_z\hat{z} - d_i' \qquad (13)$$

$$\tilde{s} = K_z\tilde{z} - d_i' + O(\mu) \qquad (14)$$

where $d_i'=d_i-K_x x_0$. Assume $\tilde{s},\hat{s}>0$. For $\tilde{z}$ to cross the boundary, $\frac{d\tilde{s}}{d\tau}<0$ where $\frac{d\tilde{s}}{d\tau}$ is given by expression (11). Correspondingly, $\frac{d\hat{s}}{d\tau} < 0$ where $\frac{d\hat{s}}{d\tau}$ is given by expression (12). At the boundary crossing, $\tilde{s}(\tau')=0$ so that $K_z\tilde{z}-d_i'=O(\mu)$. It follows that $\hat{s}(\tau')=O(\mu)$.

Since $\frac{d\hat{s}}{d\tau}=O(\mu^0)$ then $\frac{\Delta\hat{s}}{\Delta\tau} = O(\mu^0)$. Hence, $\Delta\tau=O(\mu)$ since $\Delta\hat{s}=\hat{s}(\tau')=O(\mu)$. Therefore, if $\tilde{z}$ crosses a boundary into a new region at τ', then $\hat{z}$ must also cross into the same region at a time τ'' such that $\tau''=\tau'+O(\mu)$.

It remains to be shown that the time difference of $O(\mu)$ in the boundary crossing has $O(\mu)$ effect on the solution. Let $A = A_{22}{}^i$ and $\Delta A = A_{22}{}^j - A$ where R_j is the new region and $\tau_0<\tau'$ be such that both $\hat{z}(\tau_0)$ and $\tilde{z}(\tau_0)$ lie in region R_i. Then the solution of (7) for $\tau>\tau'$ is

$$\hat{z}(\tau) = \Phi(\tau,\tau_0)\hat{z}(\tau_0) + \int_{\tau'}^{\tau}\Phi(\tau,\sigma)(\Delta A\hat{z}+A_{21}{}^j x_0+w_2{}^j)d\sigma$$
$$+ \int_{\tau_0}^{\tau'}\Phi(\tau,\sigma)(A_{21}{}^i x_0+w_2{}^i)\,d\sigma \qquad (15)$$

where $\Phi(\tau,\tau') = \exp[A(\tau-\tau')]$. Since the integrands are bounded in both integrals and $\tau''-\tau' = O(\mu)$, equation (15) is rewritten as

$$\hat{z}(\tau) = \Phi(\tau,\tau_0)\hat{z}(\tau_0) + \int_{\tau''}^{\tau}\Phi(\tau,\sigma)(\Delta A\hat{z}+A_{21}{}^j x_0+w_2{}^j)d\sigma$$
$$+\int_{\tau_0}^{\tau''}\Phi(\tau,\sigma)(A_{21}{}^i x_0+w_2{}^i)\,d\sigma + O(\mu) \qquad (16)$$

Similarly, the solution to equation (9) is found to match the form of equation (16) exactly. Hence, $\tilde{z}(\tau)=\hat{z}(\tau)+O(\mu)$. ■

Theorem 3: Let the vector field near a boundary at $d_i=K_z\hat{z}+K_x x_0$ in the space R^r be given by

$$\hat{f}(\hat{z}) = A_{21}{}^i\,x_0 + A_{22}{}^i\,\hat{z} + w_2{}^i. \qquad (17)$$

Assume that $\hat{f}(\hat{z})$ does not vanish near the boundary. If $\hat{f}(\hat{z})$ intersects the boundary with an angle of order $O(\mu^0)$, then the difference between the solutions of (7) and (9) is $O(\mu)$.

Proof: The proof is very similar to that of Theorem 2. The gist of the proof is to show that if $\hat{z}$ crosses the boundary prior to a crossing of $\tilde{z}$, then $\tilde{z}$ must cross within a time of order $O(\mu)$. The time delay in crossing affects the error in the approximation only by order $O(\mu)$.

Using the results of Theorems 2 and 3 it is seen that the errors in the approximation are of order $O(\mu)$. The restriction given in Section 1 that the regions of linearity be parallel is used in the proof of the theorems but is not a necessary condition. The difficulty is showing that if a solution crosses a boundary near an intersection of boundaries then the approximation will remain within an error of order $O(\mu)$.

3. OUTER SOLUTION

A reduced-order model for system (1)-(2) is developed below with approximation errors of order $O(\mu)$ for the time outside of the boundary layer. Assuming that the fast subsystem given in equation (7) is asymptotically stable to its equilibrium point, the fast component of z is negligible outside of the boundary layer. Therefore, the variables of the reduced-order slow model are x and the quasi-steady-state value z_s of z. Here, z_s is the equilibrium point of (7) when x_0 is replaced with x. Hence, the quasi-steady-state value of z is a continuous implicit function of x. (Continuity is shown below.) The value of z_s can be determined by using the Katzenelson algorithm (see Section 2.1) with x substituted for x_0. The algorithm is initialized with z_s^1 equal to the previous value of z_s. Due to continuity, a small change in x results in a small change in z_s. Hence, in time-integrating the system, generally only steps 0)-3) are used to find a new z_s at each time-step. Continuity of z_s as a function of x is shown in the proof of the following theorem.

Theorem 4: Let $f:R^r \rightarrow R^r$ be a continuous piecewise-linear mapping defined in the i^{th} region by

$$f(z) = A_{21}{}^i x + A_{22}{}^i z + w_2{}^i \qquad (18)$$

If f is a homeomorphism then the equilibrium point z_s of (18) is given by a continuous function of x.

Proof: Since f is a homeomorphism, a unique solution for z_s exists for any x. Let x_1 be given with resulting z_s given by $z_{s,1}$.
Let S_i denote the region of $(x_1, z_{s,1})$ in R^{p+r}. Suppose $(x_1, z_{s,1})$ lies in the interior of region S_i. Then $z_{s,1}$ can be written as

$$z_{s,1} = -(A_{22}{}^i)^{-1}(A_{21}{}^i x_1 + w_2{}^i) \qquad (19)$$

It is clear that z_s is a continuous function of x at x_1 supposing that there exists a $\delta > 0$ such that (x, z_s) lies in region S_i for all x such that $\|x_1 - x\| < \delta$. Defining $M = K_x - K_z(A_{22}{}^i)^{-1}A_{21}{}^i$ and $d_j{}' = d_j + K_z(A_{22}{}^i)^{-1}w_2{}^i$ (for $j = i-1, i$), a δ is given by

$$\delta = \min \left[\|M^{\dagger}(d_i{}' - Mx_1)\| \ , \ \|M^{\dagger}(Mx_1 - d_{i-1}{}')\| \right]$$

where $M^{\dagger} = M^T(MM^T)^{-1}$. Therefore, z_s is a continuous function of x for all x such that (x, z_s) lies in the interior of a region.
Suppose x_1 is given so that $(x_1, z_{s,1})$ lies on a boundary, say $d_i = K_x x_1 + K_z z_{s,1}$. Choose x_2 close to x_1 resulting in $z_s = z_{s,2}$. If $(x_2, z_{s,2})$ lies in region S_i then the above analysis is applied and z_s is considered to be continuous from the closed halfspace in region S_i. If $(x_2, z_{s,2})$ lies in region S_{i+1}, then

$$z_{s,2} = -(A_{22}{}^{i+1})^{-1}(A_{21}{}^{i+1} x_2 + w_2{}^{i+1}) \qquad (20)$$

A consequence of the continuity of f is that

$$-(A_{22}{}^i)^{-1}(A_{21}{}^i x_1 + w_2{}^i) \quad + $$
$$(A_{22}{}^{i+1})^{-1}(A_{21}{}^{i+1} x_1 + w_2{}^{i+1}) = 0 \qquad (21)$$

Adding equation (21) to equation (20), subtracting the result from (19) and taking the norm of both sides yields:

$$\|z_{s,1} - z_{s,2}\| = \|(A_{22}{}^{i+1})^{-1}A_{21}{}^{i+1}(x_2 - x_1)\| \leq$$
$$\|(A_{22}{}^{i+1})^{-1}A_{21}{}^{i+1}\|\|x_2 - x_1\|$$

Hence, z_s satisfies a Lipshitz condition in the open halfspace in region S_{i+1}. Therefore, z_s is continuous for x such that (x, z_s) lies on a boundary hyperplane. Thus, z_s is a continuous function of x. ■

The reduced-order slow model of system (1)-(2) for t outside of the boundary layer, i.e. $t > t_0 + \delta$, is given as follows:

$$\dot{x}_s = A_{11}{}^i x_s + A_{12}{}^i z_s + w_1{}^i, \quad x_s(t_0) = x_0 \qquad (22)$$

where z_s is an implicit function of x and is found using the Katzenelson algorithm.

The error in the approximation is due entirely to the fact that $z = z_s + O(\mu)$. This error is analogous to the error of approximating x by x_0 in the boundary layer solution. Therefore, the effect of the error can be analyzed similarly as in Theorems 2 and 3 showing that the errors in the solution are of order $O(\mu)$.

4. EXAMPLE

The techniques previously described for separating a piecewise-linear singularly perturbed system are demonstrated on the example below. The model represents a linear system with a saturation nonlinearity in the feedback loop. Such types of models exist in both flight controls and in electrical circuits. The system is given by

$$\dot{x} = A_{11}x + A_{12}z - B_1 u \qquad (23)$$
$$\mu\dot{z} = A_{21}x + A_{22}z - B_2 u \qquad (24)$$
$$u = \begin{cases} -1 \quad , & \text{if } K_x x + K_z z < -1 \\ K_x x + K_z z, & \text{if } |K_x x + K_z z| \leq 1 \\ 1 \quad , & \text{if } K_x x + K_z z > 1 \end{cases}$$

where $\mu = 0.1$. The parameter matrices are given as follows:

$$A_{11} = \begin{bmatrix} -3 & 0.4 \\ 0 & 0 \end{bmatrix} \quad A_{12} = \begin{bmatrix} 0 & 0 \\ 0.345 & 0 \end{bmatrix} \quad B_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} 0 & -0.524 \\ 0 & 0 \end{bmatrix} \quad A_{22} = \begin{bmatrix} -0.465 & 0.262 \\ 0 & -1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$K_x = [1 \ 0.861] \quad K_z = [1.220 \ 0.310]$$

The initial conditions are given as $x(0) = z(0) = [2. \ 3.]'$.

The substitution of u into (23)-(24) yields a piecewise-linear model, with three regions:

$S_1=\{(x,z):K_x x+K_z z<-1\}$, $S_2=\{(x,z):|K_x x+K_z z|\leq 1\}$ and $S_3=\{(x,z):K_x x+K_z z>1\}$. The initial condition is in S_3.

The reduced-order models given in the form of equations (7) and (22) are used in finding the time response. Comparisons between these results and those obtained by time-integrating the full order model are shown in Figure 1 through Figure 4. Note that the approximation matches the actual response very closely, i.e. within an error of order $O(\mu)$. The computation time for the approximation was roughly one-third of that for the actual system. Furthermore, as the value of μ decreases, the approximation becomes more accurate and the relative computation time decreases due to the numerical stiffness in the actual system.

5. SUMMARY

A singular perturbation technique is developed in this paper which allows for a decoupling of a continuous piecewise-linear system into slow and fast subsystems. Under the assumption of asymptotic stability, the fast variable is found to decay in the boundary layer to its quasi-steady-state solution. This quasi-steady-state solution is given by a continuous implicit function of the slow variable. The solution is found using the finite step algorithm given in the paper. Sufficient conditions for the approximation to be accurate to an order of $O(\mu)$ are given. The technique developed is successfully illustrated via a numerical example.

REFERENCES

[1] P.V. Kokotovic, R.E. O'Malley and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--An Overview," Automatica, vol. 12, pp. 123-132, March 1976.

[2] J.J. Levin, "The Asymptotic Behavior of the Stable Initial Manifolds of a System of Nonlinear Differential Equations," Tran. Am. Math. Soc., vol. 85, pp. 357-368, 1957.

[3] J. Katzenelson, "An Algorithm for Solving Nonlinear Resistive Networks," Bell Syst. Tech. J., vol. 44, pp. 1605-1620, 1965.

[4] T. Fujisawa and E.S. Kuh, "Piecewise-linear Theory of Nonlinear Networks," SIAM J. Appl. Math., vol. 22, pp. 307-328, March 1972.

[5] L.O. Chua, "Efficient Computer Algorithms for Piecewise-Linear Analysis of Resistive Networks," IEEE Trans. Circuits and Systems, vol. 18, pp. 73-85, Jan. 1971.

[6] S.N. Stevens and P-M Lin, "Analysis of Piecewise-Linear Resistive Networks Using Complimentary Pivot Theory," IEEE Trans. Circuits and Systems, vol. 28, pp. 429-441, May 1981.

[7] T. Ohtsuki, T. Fujisawa and S. Kumagai, "Existence Theorem and a Solution Algorithm for Piecewise-Linear Resistor Networks," SIAM J. Mathematical Analysis, vol. 8, pp. 69-99, Feb. 1977.

[8] S.M. Kang and L.O. Chua, "A Global Representation of Multidimensional Piecewise-Linear Functions with Linear Partitions," IEEE Trans. Circuits and Systems, vol. CAS-25, pp. 938-940, Nov. 1978.

[9] W.C. Rheinboldt and J.S. Vandergraft, "On Piecewise Affine Mappings in R^n," SIAM J. Appl. Math., vol. 29, pp. 680-689, Dec. 1975.

[10] V.C. Prasad and P.B.L. Gaur, "Homeomorphism of Piecewise-Linear Resistive Networks," Proc. IEEE, vol. 71, pp. 175-177, Jan. 1983.

[11] M. Kojima and R. Saigal, "On the Relationship Between Conditions that Insure a PL Mapping is a Homeomorphism," Mathematics of Operations Research, vol. 5, pp. 101-109, Feb. 1980.

[12] S-M Lee and K-S Chao, "Multiple Solutions of Piecewise-Linear Resistive Networks," IEEE Trans. Circuits and Systems, vol. CAS-30, pp. 84-89, Feb. 1984.

[13] Y. D. Landau, Adaptive Control - The Model Reference Approach, Marcel Dekker, 1979.

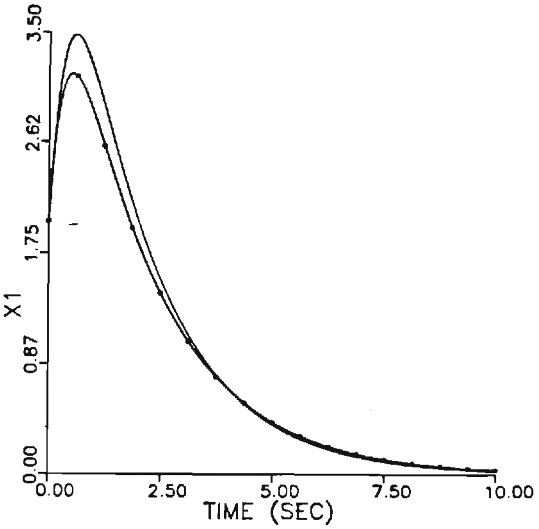

Figure 1: Response of x_1 to initial condition for actual system (solid line) and approximated system.

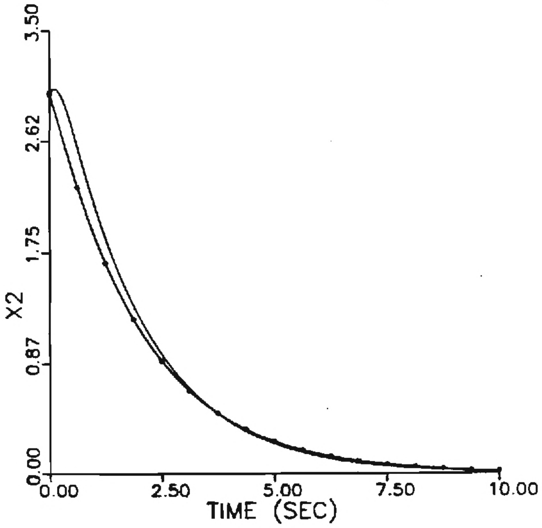

Figure 2: Response of x_2 to initial condition for actual system (solid line) and approximated system.

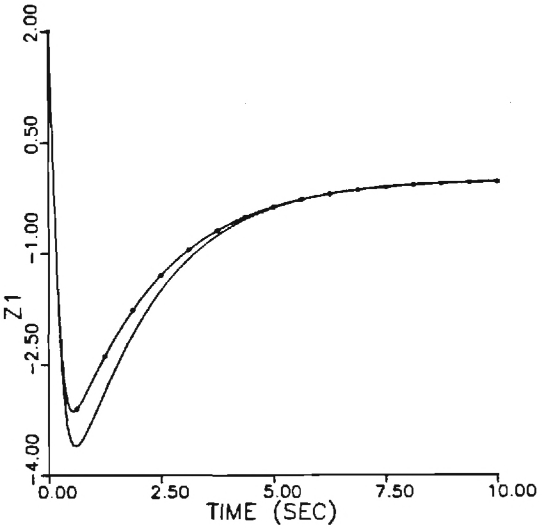

Figure 3: Response of z_1 to initial condition for actual system (solid line) and approximated system.

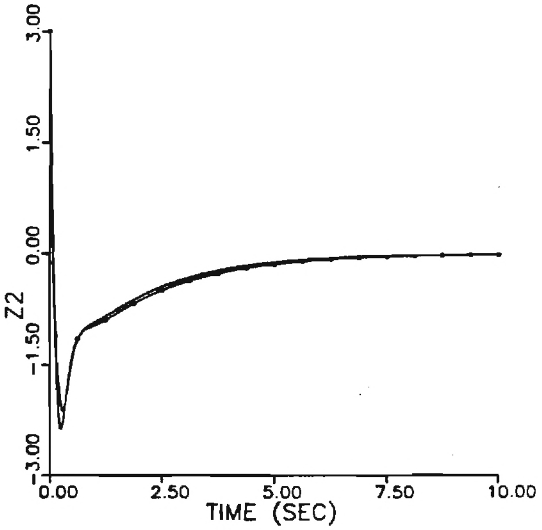

Figure 4: Response of z_2 to initial condition for actual system (solid line) and approximated system.

T-AC/34/1//23302

Singular Perturbation in Piecewise-Linear Systems

B. S. Heck
A. H. Haddad

Reprinted
IEEE TRANSACTIONS ON AUTOMATIC CONTROL
Vol. 34, No. 1, January 1989

Technical Notes and Correspondence

Singular Perturbation in Piecewise-Linear Systems

B. S. HECK AND A. H. HADDAD

Abstract—**This note analyzes piecewise-linear systems which are singularly perturbed. A technique is developed that allows decoupling of such systems into fast and slow subsystems for analysis and design. The results of a numerical example are included to demonstrate this technique.**

I. Introduction

Piecewise-linear systems which are singularly perturbed are found in many applications including electrical circuits and flight controls. The piecewise linearity may be due to nonlinear elements such as saturation or may result from a linearization about various operating points of a nonlinear plant. These types of systems are numerically very stiff and, hence, are difficult to analyze. This problem may be alleviated by using singular perturbation theory to separate the system into reduced-order models, one containing the slow dynamics and one containing the fast dynamics. The use of standard singular perturbation techniques, however, requires that the system dynamical equations be smooth [1], [2] ruling out their use on piecewise-linear systems. This note extends the general method of singular perturbation for application to continuous piecewise-linear systems.

A. Problem Formulation

The system considered in this note may be represented in the following form:

$$\dot{x}=f_1(x, z) \qquad x(t_0)=x_0 \tag{1}$$

$$\mu\dot{z}=f_2(x, z) \qquad z(t_0)=z_0 \tag{2}$$

where f_1 and f_2 are continuous piecewise-linear functions, $\mu > 0$ is a small parameter, and $x \in R^p$ and $z \in R^r$. The functions are affine in specific regions of the state space (R^{p+r}) where a region is typically defined as an intersection of halfspaces. For example, (1) and (2) are represented in the ith region by the following linear system:

$$\dot{x}=A^i_{11}x+A^i_{12}z+w^i_1 \tag{3}$$

$$\mu\dot{z}=A^i_{21}x+A^i_{22}z+w^i_2. \tag{4}$$

For the purposes of this note, the ith region is defined by the set $S_i = \{(x, z):d_{i-1} < K_x x + K_z z \leqq d_i\}$ where K_x and K_z are row vectors and $d_{i-1} < d_i$ are scalars. By this definition, the type of regions allowed are parallel in that the boundaries do not intersect. An example of a physical system which has this description is one in which the piecewise-linear element is in a scalar feedback loop. The reason for the restriction will be discussed in Section II.

The following is an outline of the note. Section II discusses the boundary layer solution and develops a reduced-order model to approximate this solution. The outer solution along with a corresponding reduced-order model is discussed in Section III. A numerical example is presented in Section IV to demonstrate the techniques developed in this note. Section V concludes the note.

II. Boundary Layer Solution

The fast dynamics of the system are most prominent during the boundary layer and can be decoupled from the slow dynamics by introducing an expanded time scale $\tau = (t - t_0)/\mu$. Examination of (1) shows that x stays relatively constant with respect to τ assuming that A^i_{11}, A^i_{12}, and w^i_1 are bounded in all regions S_i [1]. Equation (2) may be rewritten as follows:

$$\frac{d\hat{z}}{d\tau}=\hat{f}(\hat{z}) \tag{5}$$

where $\hat{z}(\tau) = z(\mu\tau + t_0)$ and $\hat{f}(\hat{z}) = f_2(x_0, \hat{z})$. The function $\hat{f}$ is a continuous piecewise-linear mapping from R^r into R^r. The state space in R^r is partitioned into regions where the function is affine, e.g., the ith region is defined as the set $R_i = \{z:d_{i-1} < K_x x_0 + K_z z \leqq d_i\}$. A degenerate case where $K_z = 0$ results in the existence of only one region in R^r so that $\hat{f}$ is affine. The initial quasi-steady-state value, $z_s(t_0)$, of $z(t)$ is a stable equilibrium point of (5). Note that the equilibrium point of the degenerate case is easily found.

The equilibrium point(s) of (5) for the nondegenerate case can be found using solution techniques developed for piecewise-linear resistive networks. Many papers have been written on finding the solution x of the equation $f(x) = y$ where f is a continuous piecewise-linear function, e.g., [3]–[9]. Fujisawa and Kuh show in [4] that a continuous piecewise-linear function satisfies a Lipshitz condition. The following theorem from [4] gives sufficient conditions for the existence and uniqueness of the solution.

Theorem 1: Let f be a continuous piecewise-linear mapping of R^r into itself and let J^i_k denote the matrix composed of the first k rows and columns of the Jacobian matrix J^i in region R_i. The mapping is a homeomorphism of R^r onto itself if, for each $k = 1, 2, \cdots, r$, the determinants of the $k \times k$ matrices

$$J^1_k, J^2_k, \cdots, J^n_k$$

do not vanish and have the same sign.

This previous work is used in finding the equilibrium point(s) of system (5) by solving $\hat{f}(\hat{z}) = 0$. In this application, $J^i = A^i_{22}$ and each A^i_{22} is assumed to be Hurwitz for stability purposes. The conditions of Theorem 1 may be stringent and various other sufficient conditions for the existence and uniqueness of the solution are given in [9]–[11]. Also, [12] discusses nonunique solutions.

A. Algorithm to Solve for Equilibrium Point

The Katzenelson algorithm is widely used in solving for x in the equation

$$f(x)=y \tag{6}$$

where $f:R^r \rightarrow R^r$ is continuous and piecewise linear. The basic outline of this algorithm used in solving $\hat{f}(\hat{z}) = 0$ is given below. More details of the general method are given in [4]. Let $W^j = A_{21}{}^j x_0 + w^j \; \forall j$, and denote the iteration number on z_s and λ by superscripts.

0) Initialize by letting $i = 1$ and $z^i_s = z_0$.
1) Solve $z = -(A^j_{22})^{-1}W^j$, where region R_j contains z^i_s.
2) If z lies in region R_j, then $z_s = z$ and stop.
3) Otherwise, let R_k be the region containing z;
   If $k > j$, then $d = d_j$ and then let $j = j + 1$
   If $k < j$, then $d = d_{j-1}$ and then let $j = j - 1$.
4) Solve $\lambda^i = (K_z z_s + K_x x_0 - d)/K_z(z^i_s - z)$.
5) Solve $z^{i+1}_s = z^i_s - \lambda^i(z^i_s - z)$.
6) Let $i = i + 1$ and go to 1).

Manuscript received October 4, 1987; revised February 16, 1988 and April 13, 1988. This work was supported by the U.S. Air Force under Contract F08635-84-C-0273 (with the Armament Laboratory) and by AFOSR-87-0308.

B. S. Heck is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250.

A. H. Haddad was with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250. He is now with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208.

IEEE Log Number 8823302.

0018-9286/89/0100-0087$01.00

It is shown in [4] that if the piecewise-linear function is a homeomorphism (e.g., it satisfies the conditions of Theorem 1), then the algorithm will converge in a finite number of steps.

B. Boundary Layer Approximation

A fast model approximating the dynamics occurring in the boundary layer can be found once the equilibrium point of system (5) is known. The boundary layer solution is then given as $\hat{z}_f(\tau) = \hat{z}(\tau) - z_s(t_0)$. In this application, z_s must be found implicitly because f_2 is not smooth. Therefore, the fast model approximating the boundary layer solution is given in terms of $\hat{z}$. In the *i*th region the fast model is given by

$$\frac{d\hat{z}}{d\tau}=A_{21}^{i}x_0+A_{22}^{i}\hat{z}+w_2^i \qquad \hat{z}(0)=z_0$$

$$\hat{z}_f(\tau)=\hat{z}(\tau)-z_s(t_0) \tag{7}$$

where the *i*th region is defined by the set $R_i = \{\hat{z}: d_{i-1} < K_x x_0 + K_z\hat{z} \leqq d_i\}$.

For the purposes of this note, it is assumed that there exists exactly one equilibrium point which is asymptotically stable. Multiple stable equilibrium points may be handled by partitioning the state space into domains of attraction for the various equilibrium points and the analysis in this note holds for each domain of attraction.

Asymptotic stability is assumed in this system although there is no known general method for determining asymptotic stability of piecewise-linear systems. Depending on the specific system under consideration, a Lyapunov function may be found. Another possibility is to use standard SISO frequency domain techniques or hyperstability. For using hyperstability notions, system (5) may be rewritten as

$$\frac{d\hat{z}}{d\tau}=A\hat{z}+Bu \tag{8}$$

where A is chosen to be stable, B is the identity I, and u is defined in the *i*th region to be $u = \Delta A^i\hat{z} + A_{12}^i x_0 + w_2^i$ where $\Delta A^i = A_{22}^i - A$. If the nonlinearity in the feedback loop satisfies the Popov integral inequality, then the necessary and sufficient condition for asymptotic stability is that the transfer matrix $(sI - A)^{-1}$ must be strictly positive real [13].

The errors in this approximation, which are of order $O(\mu)$, are due to the substitution of x_0 for x in (7) and in the definition of the regions. Substituting $x = x_0 + O(\mu)$ in (7) and in R_i yields the system

$$\frac{d\tilde{z}}{d\tau}=A_{21}^{i}(x_0+O(\mu))+A_{22}^{i}\tilde{z}+w_2^i \qquad \tilde{z}(0)=z_0$$

$$R_i=\{\tilde{z}:\, d_{i-1}+O(\mu)<K_x x_0+K_z\tilde{z} \leqq d_i+O(\mu)\} \tag{9}$$

where $\tilde{z}$ represents the actual response. In the interior of any particular region, both the approximation and the actual model are linear. Previous results on singular perturbation theory in linear systems show that if $\tilde{z}(\tau') = \hat{z}(\tau') + O(\mu)$, then $\tilde{z}(\tau'') = \hat{z}(\tau'') + O(\mu)$ for $\tau'' > \tau'$ as long as both $\tilde{z}$ and $\hat{z}$ stay within the region. The problems that may arise due to a boundary crossing are eliminated if the class of systems allowed is restricted to those in which the vector field intersects a boundary hyperplane at a large enough angle [i.e., $\neq O(\mu)$].[1] In these systems if either $\tilde{z}$ or $\hat{z}$ crosses into another region, the other must also cross into that region. The resulting error in the approximation remains of order $O(\mu)$. These conditions are summarized in the following theorems. Note that the restriction placed on the class of systems is sufficient and not necessary for proving that the approximation error is of order $O(\mu)$.

Theorem 2: Let the vector field near a boundary at $d_i = K_z\tilde{z} + K_x x_0 + O(\mu)$ in the space R' be given by

$$f(\tilde{z})=A_{21}^{i}(x_0+O(\mu))+A_{22}^{i}\tilde{z}+w_2^i. \tag{10}$$

Assume that $f(\tilde{z})$ does not vanish near the boundary. If $f(\tilde{z})$ does not intersect the boundary with an angle of order $O(\mu)$, then the difference between the solutions of (7) and (9) is of order $O(\mu)$.

Proof: Assume $\tilde{z}$ crosses the d_i boundary at τ' and $\hat{z}$ has not crossed yet. Prior to crossing $\tilde{z} = \hat{z} + O(\mu)$. The normal vector of the boundary hyperplane is given by $n = K_z^T/\|K_z\|$. Since $f(z)\cdot n \neq O(\mu)$, then

$$K_z(A_{21}^{i}(x_0+O(\mu))+A_{22}^{i}\tilde{z}+w_2^i)\neq O(\mu). \tag{11}$$

It follows that

$$K_z(A_{21}^{i}x_0+A_{22}^{i}\hat{z}+w_2^i)\neq O(\mu). \tag{12}$$

Define $\hat{s}$ and $\tilde{s}$ by

$$\hat{s}=K_z\hat{z}-d_i' \tag{13}$$

$$\tilde{s}=K_z\tilde{z}-d_i'+O(\mu) \tag{14}$$

where $d_i' = d_i - K_x x_0$. Assume $\tilde{s}, \hat{s} > 0$. For $\tilde{z}$ to cross the boundary, $d\tilde{s}/d\tau < 0$ where $d\tilde{s}/d\tau$ is given by expression (11). Correspondingly, $d\hat{s}/d\tau < 0$ where $d\hat{s}/d\tau$ is given by expression (12). At the boundary crossing, $\tilde{s}(\tau') = 0$ so that $K_z\tilde{z} - d_i' = O(\mu)$. It follows that $\hat{s}(\tau') = O(\mu)$. Since $d\hat{s}/d\tau \pm O(\mu)$, then $\Delta\hat{s}/\Delta\tau \neq O(\mu)$. Hence, $\Delta\tau = O(\mu)$ since $\Delta\hat{s} = \hat{s}(\tau') = O(\mu)$. Therefore, if $\tilde{z}$ crosses a boundary into a new region at τ', then $\hat{z}$ must also cross into the same region at a time τ'' such that $\tau'' = \tau' + O(\mu)$.

It remains to be shown that the time difference of $O(\mu)$ in the boundary crossing has $O(\mu)$ effect on the solution. Let $A = A_{22}^i$ and $\Delta A = A_{22}^j - A$ where R_j is the new region and let $\tau_0 < \tau'$ be such that both $\hat{z}(\tau_0)$ and $\tilde{z}(\tau_0)$ lie in region R_i. Then the solution of (7) for $\tau > \tau'$ is

$$\hat{z}(\tau)=\Phi(\tau,\tau_0)\hat{z}(\tau_0)+\int_{\tau'}^{\tau}\Phi(\tau,\sigma)(\Delta A\hat{z}+A_{21}^{j}x_0+w_2^j)\,d\sigma$$
$$+\int_{\tau_0}^{\tau'}\Phi(\tau,\sigma)(A_{21}^{i}x_0+w_2^i)\,d\sigma \tag{15}$$

where $\Phi(\tau, \tau') = \exp[A(\tau - \tau')]$. Since the integrands are bounded in both integrals and $\tau'' - \tau' = O(\mu)$, (15) is rewritten as

$$\hat{z}(\tau)=\Phi(\tau,\tau_0)\hat{z}(\tau_0)+\int_{\tau''}^{\tau}\Phi(\tau,\sigma)(\Delta A\hat{z}+A_{21}^{j}x_0+w_2^j)\,d\sigma$$
$$+\int_{\tau_0}^{\tau''}\Phi(\tau,\sigma)(A_{21}^{i}x_0+w_2^i)\,d\sigma+O(\mu). \tag{16}$$

Similarly, the solution to (9) is found to match the form of equation (16) exactly. Hence, $\tilde{z}(\tau) = \hat{z}(\tau) + O(\mu)$. ■

Theorem 3: Let the vector field near a boundary at $d_i = K_z\hat{z} + K_x x_0$ in the space R' be given by

$$f(\hat{z})=A_{21}^{i}x_0+A_{22}^{i}\hat{z}+w_2^i. \tag{17}$$

Assume that $f(\hat{z})$ does not vanish near the boundary. If $f(\hat{z})$ does not intersect the boundary with an angle of order $O(\mu)$, then the difference between the solutions of (7) and (9) is of order $O(\mu)$.

Proof: The proof is very similar to that of Theorem 2. The gist of the proof is to show that if $\hat{z}$ crosses the boundary prior to a crossing of $\tilde{z}$, then $\tilde{z}$ must cross within a time of order $O(\mu)$. The time delay in crossing affects the error in the approximation only by order $O(\mu)$.

Using the results of Theorems 2 and 3 it is seen that the errors in the approximation are of order $O(\mu)$. The restriction given in Section I that the regions of linearity be parallel is used in the proof of the theorems but is not a necessary condition. The difficulty is showing that if a solution crosses a boundary near an intersection of boundaries, then the approximation will remain within an error of $O(\mu)$.

III. Outer Solution

A reduced-order model for system (1) and (2) is developed below with approximation errors of order $O(\mu)$ for the time outside of the boundary layer. Assuming that the fast subsystem given in (7) is asymptotically stable to its equilibrium point, the fast component of z is negligible outside of the boundary layer. Therefore, the variables of the reduced-order slow model are x and the quasi-steady-state value z_s of z. Here, z_s is the

[1] $A \neq O(\mu)$ is used to mean $\|A\|/\mu \to +\infty$ as $\mu \to 0$.

equilibrium point of (7) when x_0 is replaced with x. Hence, the quasi-steady-state value of z is a continuous implicit function of x. (Continuity is shown below.) The value of z_s can be determined by using the Katzenelson algorithm (see Section II-A) with x substituted for x_0. The algorithm is initialized with z_s^1 equal to the previous value of z_s. Due to continuity, a small change in x results in a small change in z_s. Hence, in time-integrating the system, generally only steps 0)–2) are used to find a new z_s at each time-step. Continuity of z_s as a function of x is shown in the proof of the following theorem.

Theorem 4: Let $f: R^r \to R^r$ be a continuous piecewise-linear mapping defined in the ith region by

$$f(z)=A_{21}^i x+A_{22}^i z+w_2^i. \tag{18}$$

If f is a homeomorphism, then the equilibrium point z_s of (18) is given by a continuous function of x.

Proof: Since f is a homeomorphism, a unique solution for z_s exists for any x. Let x_1 be given with resulting z_s given by $z_{s,1}$. Let S_i denote the region of $(x_1, z_{s,1})$ in R^{p+r}.

Suppose $(x_1, z_{s,1})$ lies in the interior of region S_i. Then $z_{s,1}$ can be written as

$$z_{s,1}=-(A_{22}^i)^{-1}(A_{21}^i x_1+w_2^i). \tag{19}$$

It is clear that z_s is a continuous function of x at x_1 supposing that there exists a $\delta > 0$ such that (x, z_s) lies in region S_i for all x such that $\|x_1 - x\| < \delta$. Defining $M = K_x - K_z(A_{22}^i)^{-1}A_{21}^i$ and $d_j' = d_j + K_z(A_{22}^i)^{-1}w_2^i$ (for $j = i - 1, i$), a δ is given by

$$\delta=\min\,[\|M^\dagger(d_i'-Mx_1)\|,\ \|M^\dagger(Mx_1-d_{i-1}')\|]$$

where $M^\dagger = M^T(MM^T)^{-1}$. Therefore, z_s is a continuous function of x for all x such that (x, z_s) lies in the interior of a region.

Suppose x_1 is given so that $(x_1, z_{s,1})$ lies on a boundary, say $d_i = K_x x_1 + K_z z_{s,1}$. Choose x_2 close to x_1 resulting in $z_s = z_{s,2}$. If $(x_2, z_{s,2})$ lies in region S_i, then the above analysis is applied and z_s is considered to be continuous from the closed halfspace in region S_i. If $(x_2, z_{s,2})$ lies in region S_{i+1}, then

$$z_{s,2}=-(A_{22}^{i+1})^{-1}(A_{21}^{i+1}x_2+w_2^{i+1}). \tag{20}$$

A consequence of the continuity of f is that

$$-(A_{22}^i)^{-1}(A_{21}^i x_1+w_2^i)+(A_{22}^{i+1})^{-1}(A_{21}^{i+1}x_1+w_2^{i+1})=0. \tag{21}$$

Adding (21) to (20), subtracting the result from (19) and taking the norm of both sides yields

$$\|z_{s,1}-z_{s,2}\|=\|(A_{22}^{i+1})^{-1}A_{21}^{i+1}(x_2-x_1)\| \leqq \|(A_{22}^{i+1})^{-1}A_{21}^{i+1}\|\ \|x_2-x_1\|.$$

Hence, z_s satisfies a Lipshitz condition in the open halfspace in region S_{i+1}. Therefore, z_s is continuous for x such that (x, z_s) lies on a boundary hyperplane. Thus, z_s is a continuous function of x. ■

The reduced-order slow model of system (1) and (2) is given in the ith region of R^p, $\{x_s: d_{i-1} < K_x x_s + K_z z_s \leqq d_i\}$, as follows:

$$\dot{x}_s=A_{11}^i x_s+A_{12}^i z_s+w_1^i \qquad x_s(t_0)=x_0 \tag{22}$$

where z_s is an implicit function of x_s and is found using the Katzenelson algorithm. The actual variables, x and z, are approximated by x_s and z_s for t outside of the boundary layer, i.e., $t > t_0 + \delta$.

The error in the approximation is due entirely to the fact that $z = z_s + O(\mu)$. This error is analogous to the error of approximating x by x_0 in the boundary layer solution. Therefore, the effect of the error can be analyzed similarly as in Theorems 2 and 3 showing that the errors in the solution are of order $O(\mu)$.

IV. Example

The techniques previously described for separating a piecewise-linear singularly perturbed system are demonstrated in the following example.

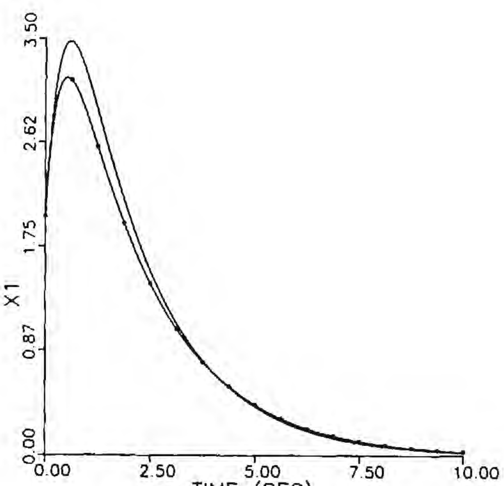

Fig. 1. Response of x_1 to initial condition for actual system (solid line) and approximated system.

The model represents a linear system with a saturation nonlinearity in the feedback loop. Such types of models exist in both flight controls and in electrical circuits. The system is given by

$$\dot{x}=A_{11}x+A_{12}z-B_1u \tag{23}$$

$$\mu\dot{z}=A_{21}x+A_{22}z-B_2u \tag{24}$$

$$u=\begin{cases}-1, & \text{if } K_x x+K_z z<-1\\ K_x x+K_z z, & \text{if } |K_x x+K_z z| \leqq 1\\ 1, & \text{if } K_x x+K_z z>1\end{cases}$$

where $\mu = 0.1$. The parameter matrices are given as follows:

$$A_{11}=\begin{bmatrix}-3 & 4\\ 0 & 0\end{bmatrix}\ A_{12}=\begin{bmatrix}0 & 0\\ 0.345 & 0\end{bmatrix}\ B_1=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$A_{21}=\begin{bmatrix}0 & -0.524\\ 0 & 0\end{bmatrix}\ A_{22}=\begin{bmatrix}-0.465 & 0.262\\ 0 & -1\end{bmatrix}\ B_2=\begin{bmatrix}0\\ 1\end{bmatrix}$$

$$K_x=[1\ \ 0.861] \quad K_z=[1.220\ \ 0.310].$$

The initial conditions are given as $x(0) = z(0) = [2.\ 3.]^T$.

This substitution of u into (23) and (24) yields a piecewise-linear model, with three regions: $S_1 = \{(x, z): K_x x + K_z z < -1\}$, $S_2 = \{(x, z): |K_x x + K_z z| \leqq 1\}$, and $S_3 = \{(x, z): K_x x + K_z z > 1\}$. The initial condition is in S_3.

The reduced-order models given in the form of (7) and (22) are used in finding the time response. Comparisons between these results and those obtained by time-integrating the full order model are shown in Figs. 1–4. Note that the approximation matches the actual response very closely, i.e., within an error of order $O(\mu)$. The computation time for the approximation was roughly one-third of that for the actual system. Furthermore, as the value of μ decreases, the approximation becomes more accurate and the relative computation time decreases due to the numerical stiffness in the actual system.

V. Summary

A singular perturbation technique is developed in this note which allows for a decoupling of a continuous piecewise-linear system into slow and fast subsystems. Under the assumption of asymptotic stability, the fast variable is found to decay in the boundary layer to its quasi-steady-state solution. This quasi-steady-state solution is given by a continuous implicit function of the slow variable. The solution is found using the finite step algorithm given in the note. Sufficient conditions for the approximation to

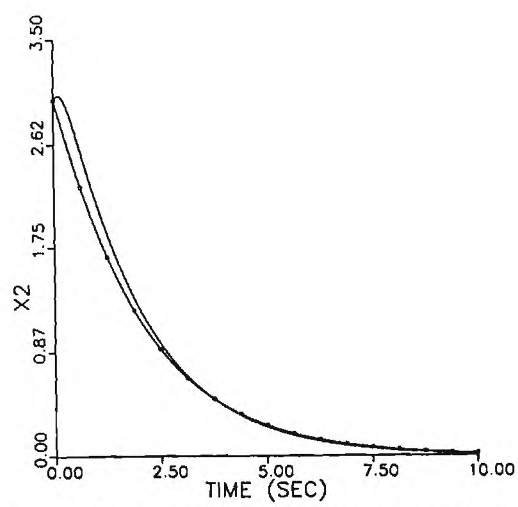

Fig. 2. Response of x_2 to initial condition for actual system (solid line) and approximated system.

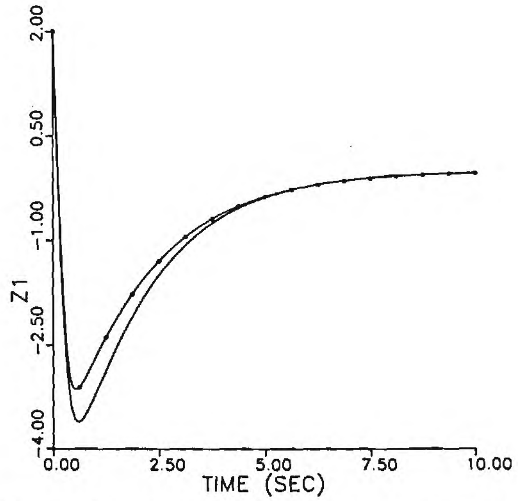

Fig. 3. Response of z_1 to initial condition for actual system (solid line) and approximated system.

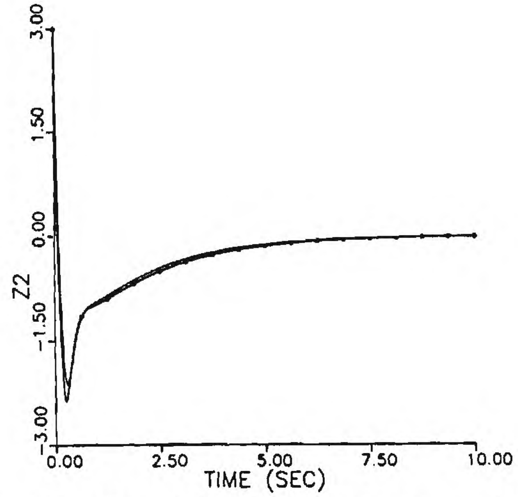

Fig. 4. Response of z_2 to initial condition for actual system (solid line) and approximated system.

be accurate to an order of $O(\mu)$ are given. The technique developed is successfully illustrated via a numerical example.

References

[1] P. V. Kokotovic, R. E. O'Malley, and P. Sannuti, "Singular perturbations and order reduction in control theory—An overview," *Automatica*, vol. 12, pp. 123–132, 1976.

[2] J. J. Levin, "The asymptotic behavior of the stable initial manifolds of a system of nonlinear differential equations," *Trans. Amer. Math. Soc.*, vol. 85, pp. 357–368, 1957.

[3] J. Katzenelson, "An algorithm for solving nonlinear resistive networks," *Bell Syst. Tech. J.*, vol. 44, pp. 1605–1620, 1965.

[4] T. Fujisawa and E. S. Kuh, "Piecewise-linear theory of nonlinear networks," *SIAM J. Appl. Math.*, vol. 22, pp. 307–328, Mar. 1972.

[5] L. O. Chua, "Efficient computer algorithms for piecewise-linear analysis of resistive networks," *IEEE Trans. Circuits Syst.*, vol. CAS-18, pp. 73–85, Jan. 1971.

[6] S. N. Stevens and P.-M. Lin, "Analysis of piecewise-linear resistive networks using complimentary pivot theory," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 429–441, May 1981.

[7] T. Ohtsuki, T. Fujisawa, and S. Kumagai, "Existence theorem and a solution algorithm for piecewise-linear resistor networks," *SIAM J. Math. Anal.*, vol. 8, pp. 69–99, Feb. 1977.

[8] S. M. Kang and L. O. Chua, "A global representation of multidimensional piecewise-linear functions with linear partitions," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 938–940, Nov. 1978.

[9] W. C. Rheinboldt and J. S. Vandergraft, "On piecewise affine mappings in R^n," *SIAM J. Appl. Math.*, vol. 29, pp. 680–689, Dec. 1975.

[10] V. C. Prasad and P. B. L. Gaur, "Homeomorphism of piecewise-linear resistive networks," *Proc. IEEE*, vol. 71, pp. 175–177, Jan. 1983.

[11] M. Kojima and R. Saigal, "On the relationship between conditions that insure a PL mapping is a homeomorphism," *Math. Operat. Res.*, vol. 5, pp. 101–109, Feb. 1980.

[12] S.-M. Lee and K.-S. Chao, "Multiple solutions of piecewise-linear resistive networks," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 84–89, Feb. 1984.

[13] Y. D. Landau, *Adaptive Control—The Model Reference Approach*. New York: Marcel-Dekker, 1979.

APPENDIX G

B. S. Heck and A. H. Haddad, "Extensions of Singular Perturbation Analysis in Piecewise Linear Systems", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 958-963, March 1988.

EXTENSIONS OF SINGULAR PERTURBATION ANALYSIS IN PIECEWISE-LINEAR SYSTEMS[1]

B.S. Heck and A.H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

This paper continues the analysis of singularly perturbed piecewise-linear systems. It provides a less restrictive sufficient condition for the validity of the singular perturbation analysis of such systems. The paper also considers the additional time-scale separation analysis required by the existence of sliding modes. Finally, the effect of random inputs on such systems is examined.

1. INTRODUCTION

This paper addresses problems in piecewise-linear systems which are singularly perturbed. Such systems are found in many applications including electrical circuits and flight controls. The piecewise-linearity may be due to nonlinear elements such as saturation or may result from a linearization about various operating points of a nonlinear plant. Singular perturbation theory is used to separate the system into reduced-order models, one containing the slow dynamics and one containing the fast dynamics. Standard singular perturbation techniques, however, are limited to systems which are smooth [1,2]. Recently, singular perturbation theory has been extended to certain types of piecewise-linear systems, i.e., those with continuous dynamics [3] and those with a scalar quantized control [4]. This paper extends these earlier results. These earlier papers [3,4] provided reduced-order models for the slow and fast dynamics and theorems showing that these models approximate the actual system within an appropriately small error. The theorems were based on geometric ideas and were restrictive in their applicability. The results are extended to include the occurrence of a sliding mode in the quantized control case. A new nongeometric criterion is introduced for using singular perturbation in the continuous piecewise-linear case. This criterion is easy to use and the proof is straightforward. Finally, the effect of random inputs is also considered.

The remainder of the paper is outlined as follows. Section 2 contains background information summarizing the results of [3] and [4] and providing physical insight into the restrictions required for using these results. theorems. Section 3 discusses the effect of a sliding mode occurring in the quantized control case. Section 4 contains a new criterion for applying singular perturbation theory. The random input analysis is contained in Section 5.

2. BACKGROUND MATERIAL

The two types of system analyzed in [3,4] are those which are continuous and those which are the result of a quantized control. Both types of systems may be expressed in the following form.

$$\dot{x} = f_1(x,z), \quad x(t_0)=x_0 \tag{1}$$

$$\mu\dot{z} = f_2(x,z), \quad z(t_0)=z_0 \tag{2}$$

where μ is a small positive parameter and f_1 and f_2 are piecewise-linear functions mapping from R^{n+m} to R^n and R^m, respectively. The functions are affine in specific regions of the state space (R^{n+m}) where a region is typically defined as an intersection of halfspaces. The i^{th} region is defined by the set $S_i=\{(x,z)\colon d_{i-1}<K_1x+K_2z\leq d_i\}$ where K_1 and K_2 are row vectors.

The systems with continuous dynamics as analyzed in [3] are those where the piecewise-linear functions, f_1 and f_2, are continuous. Such systems may be represented in the i^{th} region by the following "linear" system:

$$\dot{x} = A_{11}{}^{i}x + A_{12}{}^{i}z + w_1{}^{i} \tag{3}$$

$$\mu\dot{z} = A_{21}{}^{i}x + A_{22}{}^{i}z + w_2{}^{i} \tag{4}$$

The fast model yielding the boundary layer solution is found by introducing an expanded time-scale $\tau=(t-t_0)/\mu$. Equations (1) and (2) are expressed in the τ-time as

$$\frac{d\tilde{x}}{d\tau} = \mu f_1(\tilde{x},\tilde{z}), \quad \tilde{x}(0)= x_0 \tag{5}$$

$$\frac{d\tilde{z}}{d\tau} = f_2(\tilde{x},\tilde{z}), \quad \tilde{z}(0)= z_0 \tag{6}$$

where $\tilde{z}(\tau)=z(\mu\tau+t_0)$ and $\tilde{x}(\tau)=x(\mu\tau+t_0)$. The variable $\tilde{x}$ is found to remain constant with respect to τ, so $\tilde{x}(\tau)=x_0$. Equation (6) is then approximated as follows:

$$\frac{d\hat{z}}{d\tau} = \hat{f}(\hat{z}), \quad \hat{z}(0)=z_0 \tag{7}$$

where $\hat{f}(\hat{z})=f_2(x_0,\hat{z})$. The function $\hat{f}$ is a continuous piecewise-linear mapping from R^m into R^m. The state space in R^m is partitioned into regions where the function is affine; e.g., the i^{th} region is defined by the set $R_i=\{z\colon d_{i-1}<K_1x_0+K_2z\leq d_i\}$. The equilibrium point for (7) (i.e., the initial quasi-steady-state solution, $z_s(t_0)$) is found using the Katzenelson algorithm [3]. The approximation for the boundary layer solution, given by $\hat{z}-z_s(t_0)$, is found implicitly from the fast model defined in the i^{th} region of R^m as follows:

[1] This work is supported by the U.S. Air Force under contract AFOSR-87-0308.

$$\frac{d\hat{z}}{d\tau} = A_{21}{}^{i}x_0 + A_{22}{}^{i}\hat{z} + w_2{}^{i} \tag{8}$$

The reduced-order slow model of (3)-(4) for t outside of the boundary layer is given in the i^{th} region of R^n, $\{x_s\colon d_{i-1}<K_1x_s+K_2z_s\le d_i\}$, as follows:

$$\dot{x}_s = A_{11}{}^{i}x_s + A_{12}{}^{i}z_s + w_1{}^{i}, \quad x_s(t_0)=x_0 \tag{9}$$

where $\bar{z}_s$ is a continuous implicit function of x and is found using the Katzenelson algorithm.

Reduced-order models for systems with scalar quantized control input are developed in [4]. It was found that, without loss of generality, only those systems need to be considered which satisfy (1) and (2) with

$$f_1(x,z) = A_0x + B_0u \tag{10}$$

$$f_2(x,z) = A_2z + B_2u \tag{11}$$

$$u = Q(-K_1x-K_2z)$$

The quantizer function is defined as $Q(-K_1x-K_2z)=c_i$ for (x,z) in the i^{th} region. It is required that $c_i<c_{i+1}$, $d_0=-\infty$, $d_{n+1}=+\infty$ and that A_2 be invertible.

A fast model for this system approximating the actual solution in the boundary layer was developed in a manner similar to that for the continuous dynamics case described above. Using notation introduced previously, the fast model is given below.

$$\frac{d\hat{z}}{d\tau} = A_2\hat{z} + B_2u; \quad \hat{z}(0)=z_0 \tag{12}$$

$$u = Q(-K_1x_0-K_2\hat{z})$$

The boundary layer solution is given as $\hat{z}(\tau)-z_s$, where z_s is the equilibrium point of (12). Define $z_i=-A_2{}^{-1}B_2c_i$. Then z_s can be written as a mapping of K_1x_0, $z_s=f(K_1x_0)$, where $f(\xi)$ is as follows.

i) $f(\xi) = -A_2{}^{-1}B_2u$, if $K_2=0$ where $u=Q(-\xi)$,

ii) $f(\xi) = z_i$, if $K_2\ne 0$

and $d_i\le -K_2z_i-\xi<d_{i+1}$ for some i, (13)

iii) $f(\xi) = (\xi+d_{i+1})m$, if $K_2\ne 0$ and

$-d_{i+1}-K_2z_{i+1} < \xi \le -d_{i+1}-K_2z_i$ for some i

where $m = (z_{i+1}-z_i)/[K_2(z_i-z_{i+1})]$.

If there is no feedback from the fast variable, then case i) holds. Case ii) corresponds to an equilibrium point z_i lying inside its own region, i.e. $(x_0,z_i)\varepsilon S_i$. Case iii) corresponds to an equilibrium point lying on one of the boundaries between regions. For case iii), the resulting control switches rapidly between two values to maintain the equilibrium. It was shown in [4] that f is single-valued if $K_2A_2{}^{-1}B_2<0$; therefore, this assumption will be made in this paper. Note that f is a continuous function.

The quantized system given by (1)-(2) and (10)-(11) is approximated outside of the boundary layer by the solution to the following slow model:

$$\dot{x}_s = A_0x_s + B_0u; \quad x_s(t_0)=x_0 \tag{14}$$

$$u = Q(-K_1x_s-K_2z_s); \quad z_s = f(K_1x_s)$$

By the definition of f, it is seen that case iii) applies for (x_s,z_s) lying on a boundary hyperplane and case ii) applies otherwise.

The approximation errors for the reduced-order models (8), (9), (12) and (14) are shown under certain restrictions to be of order $O(\mu)$. In the fast model of (8), the error is due to the substitution of x_0 for x. The actual solution, $\tilde{z}$, is given by the system described in region R_i as

$$\frac{d\tilde{z}}{d\tau} = A_{21}{}^{i}(x_0+O(\mu)) + A_{22}{}^{i}\tilde{z} + w_2{}^{i} \tag{15}$$

$$R_i = \{\tilde{z}\colon d_{i-1}+O(\mu) < K_1x_0+K_2\tilde{z} \le d_i+O(\mu)\}.$$

The following theorems from [3] prove that the approximation errors are of order $O(\mu)$.

Theorem 1: Let the vector field near a boundary at $d_i=K_2\tilde{z}+K_1x_0+O(\mu)$ in the space R^m be given by

$$f(\tilde{z}) = A_{21}{}^{i}(x_0+O(\mu)) + A_{22}{}^{i}\tilde{z} + w_2{}^{i}$$

Assume the $f(\tilde{z})$ does not vanish near the boundary. If $f(\tilde{z})$ intersects the boundary with an angle of order $O(\mu^0)$, then the difference between the solutions of (8) and (15) is of order $O(\mu)$.

Theorem 2: Let the vector field near a boundary at $d_i=K_2\hat{z}+K_1x_0$ in the space R^m be given by

$$f(\hat{z}) = A_{21}{}^{i}x_0 + A_{22}{}^{i}\hat{z} + w_2{}^{i}$$

Assume that $f(\hat{z})$ does not vanish near the boundary. If $f(\hat{z})$ intersects the boundary with an angle of order $O(\mu^0)$, then the difference between the solutions of (8) and (15) is of order $O(\mu)$.

The gist of the proofs of these theorems is that if the solutions of (8) and (15) both exist a particular region of the state space, then the error between them is of order $O(\mu)$ due to the linearity. The problems that may arise at a boundary crossing are eliminated due to the restriction on the vector field. Thus, if one solution crosses into another region, the other solution must also cross into that region within a time delay of $O(\mu)$. The resulting error remains of order $O(\mu)$.

Theorems 1 and 2 may be directly applied to the quantized control system to show that the error in the fast model (12) is also of order $O(\mu)$ since continuity is not required in the proofs. The approximation errors in both slow models (9) and (14) are due to the fact that $z=z_s+O(\mu)$. This error is analogous to the error introduced into the boundary layer solutions; therefore, Theorems 1 and 2 are applicable. The main consideration for using the slow models is that the boundary layer solution must be stable so it is negligible outside of the boundary layer. A further consideration is that z_s must vary slowly with x_s so that the fast dynamics are not excited. This was shown for both models separately in [3] and [4].

The restriction in the hypothesis of Theorems 1 and 2 concerning the angle of intersection is hard to satisfy in many cases. For example, the angle of intersection described in Theorem 2 is found from the inner product of $f(\hat{z})$ and the normal to the surface, $n=K_2{}^{T}/\|K_2\|$. Hence, it is required that

$$K_2(A_{21}{}^i x_0 + A_{22}{}^i \hat{z} + w_2{}^i) = O(\mu^0) \qquad (16)$$

near the boundaries. Note that the boundary hyperplanes are parallel; all are given as translates of the null space of K_2 in R^m.

We now define a new variable y by

$$y = \hat{z} + (A_{22}{}^i)^{-1}(A_{21}{}^i x_0 + w_2{}^i).$$

Then the condition in (16) becomes

$$K_2 A_{22}{}^i y = O(\mu^0).$$

along the boundaries defined by

$$K_2 y = d_i'; \quad d_i' = d_i + K_2(A_{22}{}^i)^{-1}(A_{21}{}^i x_0 + w_2{}^i) + K_1 x_0$$

If $K_2 \neq 0$, and since $A_{22}{}^i$ has full rank, the condition will fail only in the $O(\mu)$ neighborhood about the intersection of the null space of $K_2 A_{22}{}^i$ and the boundary. Note that this intersection is an m-2 dimensional manifold on which the vector field is exactly tangent to the boundary. (If $A_{22}{}^i$ does not rotate the domain space, e.g. if $A_{22}{}^i = -I$, then there is no intersection.)

Hence, the use of Theorems 1 and 2 in showing that a particular system approximation is valid, almost requires knowing the solution beforehand. Unless A_{22} has special properties mentioned above, there exists at least one point of tangency on every boundary. If the vector field is continuous, then there exists only one point of tangency. If the vector field is discontinuous at the boundary, then there are two distinct points of tangency, one for each side of the boundary. The points in the space where the condition of the theorems fails form a set of measure $O(\mu)$ in the space. Whether the solution of the system is in this set depends on the initial conditions. However, note that this condition is sufficient but not necessary for the approximation error of the reduced-order models to be of order $O(\mu)$.

3. SLIDING MODE EQUATIONS

The previous results on singular perturbation of systems with quantized control do not account for the possibility of a sliding mode to exist on a switching boundary. Sliding modes may occur in any system in which the dynamical equations are discontinuous. Much research can be found on this topic under the more general title of variable structure systems, see for example [5]. The term "sliding mode" characterizes the behavior of a system when the vector fields on both sides of a switching boundary point towards the boundary. A representative point is directed towards the boundary from both sides and, therefore, is forced to move (or slide) along the boundary. Because the system is constrained to lie on a surface with smaller dimension than the space, a reduced-order system may be obtained. Often, the resulting reduced-order model has many properties such as robustness and invariance to disturbance which makes it attractive to control system designers [5].

In a physical system with discontinuous control, a representative point does not actually travel along the switching curve, rather, it "chatters" along the curve. The chatter is caused because an actuator cannot switch instantaneously. It may switch with a time-lag or may act as a first-order filter so that a representative point actually crosses the switching boundary into the other side before the control switches to direct it back again. In systems which are linear with respect to the control, the limiting behavior of chattering as the time-lag goes to zero is the sliding mode where the switching frequency goes to infinity [5]. For the purposes of this paper, the time-lag for switching is assumed to be of order $O(\varepsilon)$ where $\varepsilon \ll \mu$. In this way, the actuator dynamics are much faster than the fast system dynamics. (If this was not the case, then the original model (1)-(2) would be inadequate.) Thus, the system displays three time-scales, two of which (t and τ) are of interest. Therefore, setting $\varepsilon = 0$ yields the ideal sliding mode equations.

The previous theorems proving the that the slow and fast models approximate the actual singularly perturbed system with quantized control are not applicable when sliding occurs. These proofs relied on the time delay between boundary crossings of the actual solution and of the approximated solution to be of order $O(\mu)$. Each solution then spent a nonzero length of time in a particular region where the linearity properties kept the approximation error to be of order $O(\mu)$. When sliding occurs, the consecutive time spent in any one region is zero and the number of boundary crossings in any finite time interval is infinite. Therefore, the phenomenon of sliding must be handled separately.

The proof of the following theorem shows that if sliding occurs in the fast time-scale, then the approximation error remains of order $O(\mu)$. The case of sliding in the normal time-scale will follow as a consequence of this.

Theorem 3: Given the system in (5)-(6),(10)-(11) and the approximation in (12) where the vector fields on each side of a switching boundary intersect the boundary with an angle of $O(\mu^0)$, if either of the systems is sliding along the boundary and if $K_2 B_2$ is invertible then the approximation error, $\hat{z}(\tau)-\hat{z}(\tau)$, is of order $O(\mu)$,

Proof: It is clear from the proofs of Theorem 1 and 2 that if either the approximation or the actual system is sliding then the other system must also be sliding. Hence, it suffices to show that the solutions of the sliding modes of the two systems differ by $O(\mu)$. The method of equivalent control [5] will be used to find the sliding modes of the systems. Let the sliding surface for the approximation be given by $s = K_1 x_0 + K_2 \hat{z} - d_i = 0$. If the system is sliding, then $ds/d\tau = 0$.

$$\frac{ds}{d\tau} = K_2 \frac{d\hat{z}}{d\tau} = 0 \qquad (17)$$

The substitution of (12) into (17) yields

$$K_2(A_2 \hat{z} + B_2 u_{eq}) = 0$$

where the equivalent control, u_{eq}, can be solved as

$$u_{eq} = -(K_2 B_2)^{-1} K_2 A_2 \hat{z} \qquad (18)$$

We now substitute u_{eq} for u in (12) to obtain the sliding mode equations:

$$\frac{d\hat{x}}{d\tau} = 0 \tag{19}$$

$$\frac{d\hat{z}}{d\tau} = (A_2 - B_2(K_2B_2)^{-1}K_2A_2)\hat{z} \tag{20}$$

with the constraint that $K_1x_0+K_2\hat{z}-d_i=0$.

The sliding mode of the actual system sliding on the surface $s=K_1\bar{x}+K_2\bar{z}$ may be similarly obtained using (5)-(6) and (10)-(11). The equivalent control is found to be

$$u_{eq} = -(K_2B_2 + \mu K_1B_0)^{-1}(K_2A_2\bar{z} + \mu K_1A_0\bar{x}).$$

The substitution of u_{eq} for u in (10) and (11) yields the sliding mode equations:

$$\frac{d\bar{x}}{d\tau} = \mu(A_0 - \mu B_0(K_2B_2+\mu K_1B_0)^{-1}K_1A_0)\bar{x} - \mu B_0(K_2B_2+\mu K_1B_0)^{-1}K_2A_2)\bar{z} \tag{21}$$

$$\frac{d\bar{z}}{d\tau} = -\mu B_2(K_2B_2+\mu K_1B_0)^{-1}K_1A_0\bar{x} + (A_2 - B_2(K_2B_2+\mu K_1B_0)^{-1}K_2A_2)\bar{z} \tag{22}$$

with the constraint that $K_1\bar{x}+K_2\bar{z}-d_i=0$. Note that (21) and (22) along with its constraint are regular perturbations of the sliding mode equations for the approximate system (19) and (20) with its constraint. Hence, the error between the solutions is of order $O(\mu)$.

A sliding mode naturally occurs in the normal time-scale every time a boundary hyperplane is crossed. Prior to a boundary crossing in the t-time, the slow model has errors of order $O(\mu)$, and the quasi-steady-state solution, z_s, is given by case ii) of the definition of f in (13). When the solutions x_s and z_s hit the boundary, the conditions for using case iii) are satisfied to find z_s. As mentioned in Section 2, the control begins switching rapidly to maintain that value of z_s. In essence, the system satisfies the requirements for sliding in the τ-time but has reached the quasi-steady-state solution of z_s. This can be verified by noting that the value of z_s given in case iii) is an equilibrium point of the sliding mode equation in the fast time (20). Note that since the boundary layer solution was negligible prior to sliding and the switching occurs very quickly (on order of $O(\varepsilon)$), the boundary layer solution remains of order $O(\mu)$. Since z_s is continuous with respect to x_s, the conditions of case iii) in (13) are satisfied for a nonzero length of time in the t-time. Hence, the system must slide in the normal time-scale.

The sliding mode in the t-time is found from the quasi-steady-state equivalent control of the fast system. Replace $\hat{z}$ with z_s in (18) and substitute u_{eq} in (18) for u in (14) to yield

$$\dot{x}_s = A_0x_s - B_0(K_2B_2)^{-1}K_2A_2z_s \tag{23}$$

Since the solutions lie on a boundary, z_s as defined in case iii) of (13) may be substituted into (23). The resulting equation is the sliding mode in the normal time-scale,

$$\dot{x}_s = A_0x_s + \frac{B_0K_1x_s + B_0d_{i+1}}{K_2A_2{}^{-1}B_2} \tag{24}$$

valid on the $K_1x_s+K_2z_s=d_{i+1}$ surface.

4. APPROXIMATION ACCURACY

Theorems 1 and 2 are restrictive in their application due to the requirement that the vector field cannot cross a boundary tangentially. The following theorem for continuous systems provides a condition for the accuracy of the approximation that is easier to use and removes the above restriction. First, the system is reformulated and two preliminary results are given.

Let (3) and (4) be written as

$$\dot{x} = A_{11}x + A_{12}z + g_1(x,z) \tag{25}$$

$$\mu\dot{z} = A_{21}x + A_{22}z + g_2(x,z) \tag{26}$$

where $A_{11}=A_{11}{}^0$, $A_{12}=A_{12}{}^0$, $A_{21}=A_{21}{}^0$, $A_{22}=A_{22}{}^0$ (the parameter matrices in the i=0 region), and $g_1(x,z)$, $g_2(x,z)$ are piecewise-linear functions defined in region i as

$$g_1(x,z) = (A_{11}{}^i - A_{11}{}^0)x + (A_{12}{}^i - A_{12}{}^0)z + w_1{}^i$$

$$g_2(x,z) = (A_{21}{}^i - A_{21}{}^0)x + (A_{22}{}^i - A_{22}{}^0)z + w_2{}^i$$

<u>Lemma 1</u>: If A_{22} in equation (26) is stable, then there exists real positive numbers K and σ such that $\|e^{A_{22}t}\| \leq Ke^{-\sigma t}$. For proof, see [6].

<u>Lemma 2</u>: If a function, $g_2(x,z)$, is piecewise-linear, then it satisfies a Lipshitz condition, i.e., there exists a positive real number k such that

$$\|g_2(x,z)-g_2(\hat{x},\hat{z})\| \leq k\left\| \begin{bmatrix} x \\ z \end{bmatrix} - \begin{bmatrix} \hat{x} \\ \hat{z} \end{bmatrix} \right\|$$

For proof, see [7].
We now state the main result:

<u>Theorem 4</u>: For a continuous piecewise-linear singularly perturbed system, the error in the approximation given by the fast model (8) is of order $O(\mu)$ if k,K and σ as defined above satisfy $kK\leq\sigma$.

<u>Proof</u>: The actual system is given in the fast time-scale, $\tau=(t-t_0)/\mu$, by

$$\frac{dx}{d\tau} = \mu A_{11}x + \mu A_{12}z + \mu g_1(x,z)$$

$$\frac{dz}{d\tau} = A_{21}x + A_{22}z + g_2(x,z)$$

(For simplicity of notation, $\bar{z}$ is denoted as z and $\bar{x}$ as x.) The fast model approximation can be similarly given by

$$\frac{d\hat{x}}{d\tau} = 0$$

$$\frac{d\hat{z}}{d\tau} = A_{21}\hat{x} + A_{22}\hat{z} + g_2(\hat{x},\hat{z})$$

Let $\varphi_1(\tau) = x(\tau)-\hat{x}(\tau)$ and $\varphi_2(\tau) = z(\tau)-\hat{z}(\tau)$, then the following differential equation can be written.

$$\begin{bmatrix} d\varphi_1/d\tau \\ d\varphi_2/d\tau \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} + \begin{bmatrix} \mu(A_{11}x+A_{12}z+g_1(x,z)) \\ g_2(x,z) - g_2(\hat{x},\hat{z}) \end{bmatrix}$$

The following solutions are obtained.

$$\varphi_1(\tau) = \varphi_1(0) + \mu\int_0^\tau (A_{11}x(s)+A_{12}z(s)+g_1(x,z))ds$$

$$\varphi_2(\tau) = -A_{22}^{-1}(I-e^{A_{22}\tau})A_{21}\varphi_1(0) + e^{A_{22}\tau}\varphi_2(0)$$

$$+ \mu\int_0^\tau -A_{22}^{-1}(I-e^{A_{22}\tau})A_{21}(A_{11}x(s)+A_{12}z(s)+g_1(x,z))ds$$

$$+ \int_0^\tau e^{A_{22}(\tau-s)}(g_2(x,z) - g_2(\hat{x},\hat{z}))ds.$$

It is given that $\varphi_1(0)$ and $\varphi_2(0)$ are of order $O(\mu)$. For finite τ and bounded parameters, $\varphi_1(\tau)=\varphi_1(0)+O(\mu)=O(\mu)$. Similarly, the first integral in the second expression is $O(\mu)$. Using Lemma 1) from above, it can be shown that

$$\|\varphi_2(\tau)\| \le Ke^{-\sigma\tau}\|\varphi_2(0)\| +$$

$$K\int_0^\tau e^{-\sigma(\tau-s)}\|g_2(x,z)-g_2(\hat{x},\hat{z})\|ds + O(\mu).$$

From Lemma 2) and the fact that

$$\left\|\begin{matrix}\varphi_1\\ \varphi_2\end{matrix}\right\| \le \|\varphi_1\| + \|\varphi_2\|,$$

the following expression can be obtained:

$$e^{\sigma\tau}\|\varphi_2(\tau)\| \le K\|\varphi_2(0)\| + kK\int_0^\tau e^{\sigma s}[\|\varphi_1\|+\|\varphi_2\|]ds + O(\mu)$$

This is reduced using the results from [6] to

$$e^{\sigma\tau}\|\varphi_2(\tau)\| \le [K\|\varphi_2(0)\| + O(\mu)]e^{kK\tau},$$

which finally yields

$$\|\varphi_2(\tau)\| \le K\|\varphi_2(0)\|e^{-(\sigma-kK)\tau} + O(\mu).$$

Hence, if $kK\le\sigma$, then $\|\varphi_2(\tau)\|$ is of order $O(\mu)$.

The application of this theorem requires obtaining values for k, K and σ. Results from [7] can be used to find a minimum value for k.

$$k = \max_i\|A_{21}{}^i-A_{21},\ A_{22}{}^i-A_{22}\|$$

where $\|A\|$ is the maximum singular value of A. The following three methods may be used to find values for K and σ given $\|e^{A\tau}\| \le Ke^{-\sigma\tau}$ where A is stable.

1) If A is diagonalizable to Λ such that $A=M^{-1}\Lambda M$, then

$$\|e^{A\tau}\| \le \|M^{-1}\|\|M\|\|e^{\Lambda\tau}\| \le \|M^{-1}\|\|M\|e^{-\sigma\tau}$$

where $-\sigma$ is the largest real part of any eigenvalue and $K=\|M^{-1}\|\|M\|$.

2) Reference [8] shows how to obtain the following values.
Let $\beta(A) = \max\limits_j\left[\lambda_j\left(\frac{A+A^T}{2}\right)\right]$, where λ_j is an eigenvalue then $\sigma=-\beta(A)$ and $K=1$.

3) Let $B=TAT^{-1}$, then $K=\|T\|\|T^{-1}\|$ and $\sigma=-\beta(B)$ [9].

Thus, Theorem 4 may be applied without prior knowledge of the solution.

5. RANDOM INPUTS

The effect of a random input on the singularly perturbed continuous piecewise-linear model is now considered. The stochastic model may be represented by the following form.

$$dx = f_1(x,z)dt + g_1dW, \quad x(t_0)=x_0 \tag{27}$$

$$\mu dz = f_{21}(x)dt + f_{22}(x)zdt + \sqrt{\mu}g_2dW,\ z(t_0)=z_0 \tag{28}$$

where: x and z are scalar variables; W represents a Wiener process with variance parameter Q; g_1 and g_2 are constants; and f_1, f_{21} and f_{22} are continuous and piecewise-linear. As in the deterministic case, the state space is partitioned into regions where the functions are affine. Note that the noise input to the second equation is scaled to preserve the well-posedness of the fast time problem. See [10] for a discussion of this problem in linear singularly perturbed systems.

The behavior of the system in the τ time-scale is evaluated by expressing (27) and (28) as

$$d\tilde{x} = \mu f_1(\tilde{x},\tilde{z})d\tau + \sqrt{\mu}g_1dW, \quad \tilde{x}(0)=x_0 \tag{29}$$

$$d\tilde{z} = f_{21}(\tilde{x})d\tau + f_{22}(\tilde{x})\tilde{z}d\tau + g_2dW, \quad \tilde{z}(0)=z_0 \tag{30}$$

where W is now defined as a Wiener process in the τ-time. It is shown below that this system may be approximated by a reduced-order (fast) model of the following form:

$$d\hat{z} = f_{21}(x_0)d\tau + f_{22}(x_0)\hat{z}d\tau + g_2dW, \quad \hat{z}(0)=z_0 \tag{31}$$

Note that this system is completely linear.

The analysis of the fast system focuses on the propagations of the conditional joint probability density function and its corresponding characteristic function. Define the conditional joint probability density function as $p(x,z,\tau|x_0,z_0,0)$ and the characteristic function as

$$\Phi(w,v,\tau|x_0,z_0,0) = \iint_{-\infty}^{+\infty} e^{j(wx+vz)}p(x,z,\tau|x_0,z_0,0)dxdz$$

where

$$p(x,z,\tau|x_0,z_0,0) =$$

$$\frac{1}{(2\pi)^2}\iint_{-\infty}^{+\infty} e^{-j(wx+vz)}\Phi(w,v,\tau|x_0,z_0,0)dwdv$$

Because the system is piecewise-smooth, a Fokker-Plank equation which holds almost everywhere may be derived to obtain the conditional density function.

$$\frac{\partial p}{\partial\tau} = -\frac{\partial}{\partial z}[(f_{21}(x)+f_{22}(x)z)p] +$$

$$\frac{1}{2}g_2{}^2Q\frac{\partial^2}{\partial z^2}p + O(\mu^{\frac{1}{2}}) \qquad \text{a.e.} \tag{32}$$

where $p=p(x,z,\tau|x_0,z_0,0)$. The propagation of the characteristic function is derived from (32) as

$$\frac{\partial\Phi}{\partial\tau} = -\iint_{-\infty}^{+\infty} e^{j(wx+vz)}\frac{\partial}{\partial z}[(f_{21}(x)+f_{22}(x)z)p]dxdz$$

$$+ \frac{1}{2}g_2{}^2Q\iint_{-\infty}^{+\infty} e^{j(wx+vz)}\frac{\partial^2}{\partial z^2}pdxdz + O(\mu^{\frac{1}{2}}) \tag{33}$$

It is shown below that a solution to this equation is

$$\Phi(w,v,\tau|x_0,z_0,0)=e^{jwx_0}\hat{\Phi}(v,\tau|x_0,z_0,0) + O(\mu^{\frac{1}{2}}) \tag{34}$$

or, equivalently,

$$p(x,z,\tau|x_0,z_0,0)=\delta(x-x_0)\hat{p}(z,\tau|x_0,z_0,0) + O(\mu^{\frac{1}{2}})$$

where $\hat{p}(z,\tau|x_0,z_0,0)$ is the conditional probability density function of the fast model (31) and $\Phi(v,\tau|x_0,z_0,0)$ is its corresponding characteristic function.

The Fokker-Planck equation for the fast model (31) is

$$\frac{\partial\hat{p}}{\partial\tau} = -\frac{\partial}{\partial z}[(f_{21}(x_0)+f_{22}(x_0)z)\hat{p}] + \frac{1}{2}g_2{}^2Q\frac{\partial^2}{\partial z^2}\hat{p} \qquad \text{a.e.} \tag{35}$$

where $\hat{p} = \hat{p}(z,\tau|x_0,z_0,0)$. An equation for $\Phi(v,\tau|x_0,z_0,0)$ is obtained from (35) as

$$\frac{\partial\hat{\Phi}}{\partial\tau} = -\int_{-\infty}^{+\infty} e^{jvz}\frac{\partial}{\partial z}[(f_{21}(x_0)+f_{22}(x_0)z)\hat{p}]dz + \frac{1}{2}g_2{}^2Q\int_{-\infty}^{+\infty} e^{jvz}\frac{\partial^2}{\partial z^2}\hat{p}\, dxdz \tag{36}$$

The substitution of (34) into (33) yields an equation which is a regular perturbation of (36), thus proving (34). Hence, the fast model (31) is a valid approximation of the original system in the τ time-scale.

Suppose that the system is stable so that it reaches a steady-state in τ. The steady-state probability density function $\hat{p}_S(z|x_0)$ can be found from (35) to satisfy

$$(f_{21}(x_0)+f_{22}(x_0)z)\hat{p}_S = \frac{1}{2}g_2{}^2Q\frac{\partial}{\partial z}\hat{p}_S \qquad \text{a.e.} \tag{37}$$

Due to linearity, if the input is Gaussian, $\hat{p}_S$ is conditionally Gaussian. This steady-state model may be used to develop a reduced-order slow model valid in the normal time-scale.

6. SUMMARY

Previous results on the singular perturbation of piecewise-linear systems are extended in this paper. Sliding mode equations are developed in both the normal time- and the fast time-scales for the case of the quantized control. It is found that the occurrence of a sliding mode does not affect the validity of the time-scale separation procedure given in an earlier paper. A new, nongeometric theorem is given to prove that the approximations developed previously for the continuous dynamics case are accurate to within an error of order $O(\mu)$. This theorem is easy to apply and is less restrictive in its assumptions. However, because all the theorems provide sufficient but not necessary conditions, none supercedes the others. Finally, the effect of a random input on a particular continuous piecewise-linear system is analyzed. A reduced-order model approximating the system in the boundary layer is developed.

REFERENCES

[1] Kokotovic, P.V.,R.E. O'Malley, and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--an Overview," Automatica, vol. 12, pp. 123-132, March 1976.

[2] Levin, J.J., "The Asymptotic Behavior of the Stable Initial Manifolds of a System of Nonlinear Differential Equations," Trans. Am. Math. Soc.. vol. 85, pp. 357-368, 1957.

[3] Heck, B.S. and A.H. Haddad, "Singular Perturbation in Piecewise-Linear Systems," to appear in the IEEE Trans. Auto. Control.

[4] Heck, B.S. and A.H. Haddad, "On Quantized Control of Linear Singularly Perturbed Systems," to appear in Automatica.

[5] Utkin, V.I., "Variable Structure Systems with Sliding Modes," IEEE Trans. Auto. Control, vol. AC-22, pp. 212-222, April 1977.

[6] Coddington, E.A. and N. Levinson, Theory of Ordinary Differential Equations, Robert E. Krieger Publishing Co., Malabar, FL, pp. 315, 1985.

[7] Fujisawa, T. and E.S. Kuh, "Piecewise-Linear Theory of Nonlinear Networks," SIAM J. Appl. Math., vol. 22, pp. 307-328, March 1972.

[8] Doeser, C.M. and M. Vidyasagar, Feedback Systems: Input-Output Properties, Academic Press, NY, 1975.

[9] Ezzine, J., "Parameter-Perturbations in Digital Control Systems," Master's Thesis, University of Alabama in Huntsville, 1985.

[10] Khalil, H.K., A.H. Haddad and G.L. Blankenship, "Parameter Scaling and Well-Posedness of Stochastic Singularly Perturbed Control Systems," Proc. 12th Asilomar Conf. on Circuits, Systems and Computers, Pacific Grove, CA, Nov. 1978.

APPENDIX H

B. S. Heck and A. H. Haddad, "Singular Perturbation Theory for Piecewise Linear Systems with Random Inputs," Stochastic Analysis and Applications, to appear.

SINGULAR PERTURBATION THEORY FOR PIECEWISE-LINEAR SYSTEMS WITH RANDOM INPUTS

B.S. Heck

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

A. H. Haddad

Department of Electrical Engineering and Computer Science
2145 Sheridan Road
Northwestern University
Evanston, Illinois 60208

ABSTRACT

The effect of random inputs on a continuous piecewise-linear singularly perturbed system is investigated in this paper. Reduced-order models are developed for a second-order system (one fast and one slow variable) which has a random input. It is shown that the solutions of the reduced-order models approximate the actual solution with differences in probability density functions of order $O(\mu^{\frac{1}{2}})$ (in a distributional sense). For the special case of a system which is linear in the fast variable, it is shown that the mean-squared error between the approximate and actual solutions in the fast time scale is of order $O(\mu)$. An outline is provided for the extension of the results to the vector variable case.

1. INTRODUCTION

A system inherently possessing both fast and slow dynamics can often be simplified by using singular perturbation theory to separate the system into reduced-order models, one containing the fast dynamics and one containing the slow dynamics. The standard theory, however, is restricted to systems with smooth dynamics [1-3]. Recently, this theory has been extended for deterministic systems to piecewise-linear systems [4,5]. Piecewise-linear singularly perturbed systems appear in many applications including flight controls and electrical circuits. The piecewise-linearity may occur from a piecewise-linear element such as a saturation or dead zone or may occur as a result of a piecewise-linear approximation of a nonlinear system. It is desireable to extend singular perturbation theory to piecewise-linear systems with random inputs.

Reduced-order models for linear singularly perturbed systems with Gaussian random input have been developed in [6,7] for filtering, smoothing and control purposes. The filtering problem for smooth singularly perturbed nonlinear systems with wide-sense stationary random input is discussed in [8]. Reduced-order filters are designed for the smooth nonlinear system corresponding to the fast and slow dynamics. This paper extends the previous work on singular perturbation theory in piecewise-linear systems to the case of random inputs for possible use in filtering, smoothing and stochastic control. An example of this application is a singularly perturbed piecewise-linear flight control system which has random wind disturbances.

1.1 Problem Formulation

The system investigated in this paper is continuous and piecewise-linear with a random input. Since the resulting system is nonlinear, the model is written in terms of coupled Itô differential equations:

$$dx = f_1(x,z)dt + g_1 dW \qquad (1)$$

$$\mu dz = f_2(x,z)dt + \sqrt{\mu} g_2 dW \qquad (2)$$

where: x and z are scalar variables; W represents a Wiener process with variance parameter Q; g_1 and g_2 are constants; f_1 and f_2 are continuous and piecewise-linear functions. As in the deterministic case studied in [4], the state space is partitioned into regions where the functions are affine. Let the system be defined in the i^{th} region as follows:

$$dx = A_{11}{}^{i}xdt + A_{12}{}^{i}zdt + g_1dW \tag{3}$$

$$\mu dz = A_{21}{}^{i}xdt + A_{22}{}^{i}zdt + \sqrt{\mu}g_2dW \tag{4}$$

The superscripts simply denote the region number. For simplicity, the regions are restricted to be nonoverlapping, nonempty and parallel. By parallel, it is meant that the boundaries of the regions are parallel hyperplanes.

The random input to the fast subsystem is assumed to be scaled by $\sqrt{\mu}$ so that the well-posedness of the problem is preserved. It has been shown by Khalil et. al. [9] that the well-posedness is questionable unless the white noise input to the fast variable is scaled by a factor of order $O(\mu^{\alpha})$ where $0<\alpha\leq\frac{1}{2}$ or wide-band noise is used instead. The problem occurs with unscaled white noise because as $\mu\to0$ the bandwidth of the fast subsystem approaches infinitity, so that the fast variable acts like white noise in the normal time-scale. This is valid as an input to the slow model but not as a dynamic process itself. Scaled random inputs, however, do not exhibit this problem.

The outline of the paper is given as follows. In Section 2, the behavior of the system in the fast time scale is discussed. Also, a reduced-order model is developed which is valid in the fast-time scale. Similarly, a reduced-order model is developed in Section 3 to approximate the slow dynamics of the system with respect to the normal time-scale. The extension of the second-order analysis to higher order systems is outlined in Section 4. Concluding remarks are included in Section 5.

2. FAST SUBSYSTEM

The behavior of the system in the fast time-scale is investigated below. At each sample time, t_i, of the normal time scale, the fast subsystem may be evaluated. Define the expanded time variable by $\tau=(t-t_i)/\mu$ and restrict the samples so that

$t_{i+1}-t_i$ is large relative to μ. The original system in (1)-(2) is reformulated in terms of τ as follows

$$d\tilde{x} = \mu f_1(\tilde{x},\tilde{z})d\tau + \sqrt{\mu} g_1 d\tilde{W}; \qquad \tilde{x}(0) = x_0 \tag{5}$$

$$d\tilde{z} = f_2(\tilde{x},\tilde{z})d\tau + g_2 d\tilde{W}; \qquad \tilde{z}(0) = z_0 \tag{6}$$

where $\tilde{W}$ is now defined as a Wiener process in the τ time-scale given as $\tilde{W}(\tau)=\mu^{\frac{1}{2}}W(\mu\tau)$. It is shown that this system may be approximated by the solution to a reduced-order (fast) model of the following form:

$$d\hat{z} = f_2(x_0,\hat{z})d\tau + g_2 d\tilde{W}; \qquad \hat{z}(0) = z_0 \tag{7}$$

$$\hat{x}(\tau) = x_0$$

The proof that the resulting approximation error is of order $O(\mu)$ focuses on the propagations of the conditional joint probability density function and its corresponding characteristic function. Define the conditional joint probability density function for the solution of (5)-(6) as $p(\tilde{x},\tilde{z};\tau|x_0,z_0;0)$ and the characteristic function as

$$\Phi(v,w;\tau|x_0,z_0;0) = \iint_{-\infty}^{+\infty} e^{j(vx+wz)}\, p(x,z;\tau|x_0,z_0;0)dxdz \tag{8}$$

where

$$p(x,z;\tau|x_0,z_0;0) = \frac{1}{4\pi^2} \iint_{-\infty}^{-\infty} e^{-j(vx+wz)}\Phi(v,w;\tau|x_0,z_0;0)dwdv \tag{9}$$

Because the system is continuous and piecewise-smooth, a Fokker-Planck equation which holds almost everywhere may be derived to obtain the conditional joint probability density function:

$$\frac{\partial p}{\partial \tau} = - \frac{\partial}{\partial z} [f_2(x,z)p] + \frac{1}{2} g_2{}^2 Q \frac{\partial^2}{\partial z^2} p - \mu \frac{\partial}{\partial x} [f_1(x,z)p] + \tag{10}$$

$$\frac{1}{2} \mu g_1{}^2 Q \frac{\partial^2}{\partial x^2} p + \frac{1}{2} \mu^{\frac{1}{2}} g_1 g_2 Q \frac{\partial^2}{\partial x \partial z} p + \frac{1}{2} \mu^{\frac{1}{2}} g_1 g_2 Q \frac{\partial^2}{\partial z \partial x} p \quad \text{a.e.}$$

where $p = p(x,z;\tau|x_0,z_0;0)$. The equation holds everywhere except for the set of measure zero where the dervivative of f_1 and f_2 do not exist. The initial condition (in a distributional sense) and auxilary conditions are

$$p(x,z;0|x_0,z_0;0) = \delta(x-x_0)\delta(z-z_0);$$

$$p \geqq 0 \quad \text{and} \quad \iint_{-\infty}^{+\infty} p \, dxdz = 1 \tag{11}$$

(For a discussion of the derivation of the Fokker-Planck equation, see Wong [11].)

Examination of (10) shows that the propagation of p is relatively insensitive to the variation of x. Since this is a linear partial differential equation, the methods of Kato [12] can be used to show that the solution of (10) can be approximated by the solution of the following equation with errors in the solution of order $O(\mu^{\frac{1}{2}})$ (in a distributional sense).

$$\frac{\partial p_a}{\partial \tau} = - \frac{\partial}{\partial z} [f_2(x,z)p_a] + \frac{1}{2} g_2{}^2 Q \frac{\partial^2}{\partial z^2} p_a \qquad \text{a.e.} \tag{12}$$

where $p_a = p_a(x,z;\tau|x_0,z_0;0)$. The initial and auxilary conditions remain the same (again, the initial conditions are defined in a distributional sense):

$$p_a(x,z;0|x_0,z_0;0) = \delta(x-x_0)\delta(z-z_0);$$

$$p_a \geqq 0 \quad \text{and} \quad \iint_{-\infty}^{+\infty} p_a \, dxdz = 1 \tag{13}$$

Hence, $p = p_a + O(\mu^{\frac{1}{2}})$ in distribution.

To remove the consideration of the differential equation being defined almost everywhere, the propagation of the characteristic function is introduced. Denote the characteristic function of p_a as $\Phi_a(v,w;\tau|x_0,z_0;0)$ where a characteristic function is defined in equation (8). Then an expression yielding the propagation of $\Phi_a(v,w;\tau|x_0,z_0;0)$ can be found from (12) by first multiplying both sides of the equation by $e^{jvx+jwz}$ and then integrating with respect to x and z.

$$\frac{\partial\Phi_a}{\partial\tau} = \iint_{-\infty}^{+\infty} e^{jvx+jwz} \left[- \frac{\partial}{\partial z} [f_2(x,z)p_a] + \frac{1}{2} g_2{}^2Q \frac{\partial^2}{\partial z^2}p_a\right] dxdz \qquad (14)$$

The values of the derivatives can be assigned arbitrarily for points in the set of measure zero where the derivative is not defined. Since the righthand-side of (12) multiplied by $e^{jvx+jwz}$ differs from the integrand in (14) on a set of measure zero, the right-hand side of (14) is equal to the integration with respect to x and z of the right-hand side of (12) multiplied by $e^{jvx+jwz}$ (for proof, see [13]). The corresponding initial condition is $\Phi_a(v,w;0|x_0,z_0;0)=e^{jvx_0}e^{jwz_0}$ and the auxilary conditions correspond to (13), i.e., $\Phi_a(0,0;\tau|x_0,z_0;0)=1$.

Similar to the case for the actual solution, a Fokker-Planck equation can be derived to find the conditional probability density function for the approximation given in (7)

$$\frac{\partial\hat{p}}{\partial\tau} = - \frac{\partial}{\partial z} [f_2(x_0,z)\hat{p}] + \frac{1}{2} g_2{}^2Q \frac{\partial^2}{\partial z^2}\hat{p} \qquad \text{a.e.} \qquad (15)$$

where $\hat{p} = \hat{p}(z;\tau|x_0,z_0;0)$. This is subject to the following initial and auxilary conditions:

$$\hat{p}(z;\tau|x_0,z_0;0) = \delta(z-z_0); \ \hat{p} \geqq 0 \text{ and } \int_{-\infty}^{+\infty} \hat{p}\, dz = 1 \qquad (16)$$

Denote the characteristic function for $\hat{p}$ as $\hat{\Phi}(w;\tau|x_0,z_0;0)$. The propagation equation for $\hat{\Phi}(w;\tau|x_0,z_0;0)$ is found from (15) to be

$$\frac{\partial\hat{\Phi}}{\partial\tau} = \int_{-\infty}^{+\infty} e^{jwz} \left[- \frac{\partial}{\partial z} [f_2(x_0,z)\hat{p}] + \frac{1}{2} g_2{}^2Q \frac{\partial^2}{\partial z^2}\hat{p} \right] dz \qquad (17)$$

where the initial condition is $\hat{\Phi}(w;0|x_0,z_0;0)=e^{jwz_0}$.

A comment can be made about the steady-state value of $\hat{p}$. It is assumed that the system in (7) is stable so that a steady-state solution for $\hat{p}$ exists. It can be found by setting the time derivative to zero in the Fokker-Planck equation. The steady-state can then be solved from the resulting equation:

$$[f_2(x_0,z)\hat{p}_\infty] = \frac{1}{2} g_2{}^2Q \frac{\partial}{\partial z} \hat{p}_\infty \qquad \text{a.e.} \qquad (18)$$

where $\hat{p}_\infty = \hat{p}_\infty(z|x_0)$. Note that $\hat{p}_\infty$ can be considered as a function of x_0, but continuity of that function is not guaranteed.

It can now be shown that the joint probability density function given by the solution to (12) is equal to $p_a=\delta(x-x_0)\hat{p}$, or, equivalently, $\Phi_a=e^{jvx_0}\hat{\Phi}$. The expressions for p_a and Φ_a are substituted into (14) to yield

$$\frac{\partial\hat{\Phi}}{\partial\tau} e^{jvx_0} = \iint_{-\infty}^{+\infty} e^{jvx+jwz} \left[- \frac{\partial}{\partial z} [f_2(x,z)\delta(x-x_0)\hat{p}] \right.$$

$$\left. + \frac{1}{2} g_2{}^2Q \frac{\partial^2}{\partial z^2}[\delta(x-x_0)\hat{p}] \right] dxdz \qquad (19)$$

Integration with respect to x yields

$$\frac{\partial\hat{\Phi}}{\partial\tau} e^{jvx_0} = e^{jvx_0} \int_{-\infty}^{+\infty} e^{jwz} \left[- \frac{\partial}{\partial z} [f_2(x_0,z)\hat{p}] + \frac{1}{2} g_2{}^2Q \frac{\partial^2}{\partial z^2}\hat{p} \right] dz \quad (20)$$

Since the pair $(\hat{p},\hat{\Phi})$ is a solution to (17) it must also be a solution to (19). Therefore, $p_a = \delta(x-x_0)\hat{p}$ and $\Phi_a = e^{jvx_0}\hat{\Phi}$.

Finally, the assertion that the probability density function of the solution to the approximate model differs from that of the true solution by factor of order $O(\mu^{\frac{1}{2}})$ is proven in a distributional sense. Since $p = p_a + O(\mu^{\frac{1}{2}})$ (in distribution), the results of the preceeding paragraph imply that $p = \delta(x-x_0)\hat{p} + O(\mu^{\frac{1}{2}})$ (in distribution). Correspondingly, $\Phi = e^{jvx_0}\hat{\Phi} + O(\mu)$. Hence, the statistical moments of the true solution and the approximation differ only by an error of $O(\mu^{\frac{1}{2}})$.

2.1 Systems Linear in z

It can be further shown that for a system which is linear in z, the mean-squared error between the actual solution and the approximate solution is of order $O(\mu)$. A continuous piecewise-linear system that is linear in z has the following form:

$$d\tilde{x} = \mu f_{11}(\tilde{x})d\tau + \mu A_{12}\tilde{z}d\tau + \sqrt{\mu}g_1 d\tilde{W} \tag{21}$$

$$d\tilde{z} = f_{21}(\tilde{x})d\tau + A_{22}\tilde{z}d\tau + g_2 d\tilde{W} \tag{22}$$

where f_{11} and f_{21} are continuous piecewise-linear functions; A_{12}, A_{22}, g_1 and g_2 are constants and $\tilde{W}$ is a Wiener process defined in τ with variance parameter Q. This is simply a subset of systems of the general form given in (5)-(6). Note that the requirement that A_{12} and A_{22} be constant is a consequence of the continuity of the system. Also, stability of the fast model is required in this analysis, hence, A_{22} is stable. The process is ill-defined if A_{22} is not stable.

Examination of (21) shows that $\tilde{x}$ stays relatively constant with respect to τ and can be approximated by x_0. The approximation for $\tilde{z}$ is given by the solution to the following equation

$$d\hat{z} = f_{21}(x_0)d\tau + A_{22}\hat{z}d\tau + g_2 d\tilde{W} \tag{23}$$

To show that the mean-squared error between $\tilde{z}$ and $\hat{z}$ is of order $O(\mu)$, define the approximation error as $\varphi(\tau) = \tilde{z} - \hat{z}$. Then an equation for φ is given by

$$d\varphi = (f_{21}(\tilde{x})-f_{21}(x_0))d\tau + A_{22}\varphi d\tau; \quad \varphi(0) = 0 \qquad (24)$$

The solution to (24) due to linearity is given by

$$\varphi(\tau) = \int_0^\tau e^{A_{22}(\tau-\sigma)}[f_{21}(\tilde{x})-f_{21}(x_0)]\, d\sigma \qquad (25)$$

An upper bound for φ can be found by noting that f_{21} satisfies a Lipschitz condition; hence, there exists a positive constant $k<\infty$ such that:

$$\varphi(\tau) \leqq \int_0^\tau e^{A_{22}(\tau-\sigma)} \; k\|\tilde{x}-x_0\| \, d\sigma \qquad (26)$$

The mean-squared error of φ is found from (25) to have an upper bound as follows:

$$E\{\varphi^2(\tau)\} \leqq$$

$$\int_0^\tau\int_0^\tau e^{A_{22}(\tau-\sigma)} \, e^{A_{22}(\tau-\Theta)} \; k \; E\|[\tilde{x}(\sigma) - x_0][\tilde{x}(\Theta) - x_0\| \, d\sigma d\Theta \qquad (27)$$

Since A_{22} is stable and the quantity $\tilde{x}(\tau_1)-x_0$ is of order $O(\mu)$ for $0\leqq\tau_1\leqq\tau$, the integrand is found to be of order $O(\mu)$. Hence, $E\{\varphi^2(\tau)\}$ is of order $O(\mu)$.

3. SLOW SUBSYSTEM

The slow dynamics of the system (1)-(2) can be approximated in distribution by the solution of the following model:

$$dx_s = f_1(x_s,z_s)dt + g_1dW; \qquad x_s(t_0) = x_0 \qquad (28)$$

$$0 = f_2(x_s,z_s)$$

Note that z_s is found from x_s using the Katzenelson algorithm given in reference [4]. This algorithm is computationally efficient for solving algebraic piecewise-linear expressions. The approximation is validated below by showing that the true joint probability density function of x and z differs from that of x_s and z_s by a factor of $O(\mu^{\frac{1}{2}})$ (in distribution). This approximation is shown to be valid outside of the initial boundary layer as long as the solution does not cross into another region of the state space. A boundary layer may need to be evaluated after each time the solution crosses a boundary between regions.

The Fokker-Planck equation yielding the joint probability density function of the actual solution given in (1)-(2), $p(x,z;t|x_0,z_0;t_0)$, can be derived using an approach similar to one found in [11]. The Chapman-Kolmogorov equation is the starting point.

$$p(x,z;t+\Delta|x_0,z_0;t) =$$

$$\iint_{-\infty}^{+\infty} p(x,z;t+\Delta|x_1,z_1;t)p(x_1,z_1;t|x_0,z_0;t_0)\ dx_1dz_1 \qquad (29)$$

An expression for $p(x,z;t+\Delta|x_1,z_1;t)$ is found using the characteristic function. Define the characteristic function as

$$\Phi(v,w;t+\Delta|x_1,z_1;t) = \iint_{-\infty}^{+\infty} e^{jvx+jwz}\ p(x,z;t+\Delta|x_1,z_1;t)\ dxdz \qquad (30)$$

where

$$p(x,z;t+\Delta|x_1,z_1;t) = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} e^{-jvx-jwz}\ \Phi(v,w;t+\Delta|x_1,z_1;t)\ dvdw \ (31)$$

Expand the following term in a Taylor series about $x=x_1$.

$$e^{-jv(x_1-x)} = 1 - jv(x_1-x) + \frac{v^2}{2}(x_1-x)^2 + \text{h.o.t.} \qquad (32)$$

This series is substituted into (30) to yield:

$$\Phi(v,w;t+\Delta|x_1,z_1;t) \approx \iint_{-\infty}^{+\infty} e^{jwz}[1-jv(x_1-x)+\frac{v^2}{2}(x_1-x)^2\ ]$$

$$\times \quad p(x,z;t+\Delta|x_1,z_1;t)e^{jvx_1}\,dxdz \qquad (33)$$

Integration with respect to x yields

$$\Phi(v,w;t+\Delta|x_1,z_1,t) =$$

$$\int_{-\infty}^{+\infty} e^{jwz}\ e^{jvx_1}\ [1-jvf_1(x_1,z_1)\Delta + \frac{v^2}{2}g_1{}^2Q\Delta]p(z;t+\Delta|x_1,z_1;t)]\ dz \quad (34)$$

To solve for $p(z;t+\Delta|x_1,z_1;t)$, a boundary layer following t may be evaluated. Define a new time scale by $\tau=\Delta/\mu$ and let $\tilde{z}(\tau)=z(\mu\tau+t)$. Then it is found using the derivation in Section 2 that an $O(\mu^{\frac{1}{2}})$ approximation for the conditional probability density function of $\tilde{z}$, $\tilde{p}(z;\tau|x_1,z_1;t)$, is given by the solution to the following Fokker-Planck equation.

$$\frac{\partial\hat{p}}{\partial\tau} = -\frac{\partial}{\partial z}[f_2(x_1,z)\hat{p}] + \frac{1}{2}Qg_2{}^2\frac{\partial^2\hat{p}}{\partial z^2} \qquad \text{a.e.} \qquad (35)$$

where $\hat{p}=\hat{p}(z;\tau|x_1,z_1;t)$ and $\hat{p}(z;0|x_1,z_1;0)=\delta(z-z_0)$. Assuming that the system is stable, the probability density function reaches a steady-state in τ denoted as $\hat{p}(z|x_1)$. For a small value of Δ, x stays relatively constant so that $p(z,t+\Delta|x_1,z_1,t)$ is approximated (in distribution) by

$$p(z,t+\Delta|x_1,z_1;t) \approx \hat{p}(z|x_1) + O(\mu) \tag{36}$$

Restrictions on the $O(\mu)$ term arise due to the fact that both $p(z;t+\Delta|x_1,z_1;t)$ and $\hat{p}(z|x_1)$ are probability density functions so they must satisfy certain conditions. One problem with this analysis is that $\hat{p}(z|x_1)$ may not depend continuously on x_1 for those x_1 which lie on a boundary between regions in the state space. If $\hat{p}$ is not continuous with repect to x_1, then the conditional moments of z may not be continuous either. The fast dynamics may then be excited sufficiently so that the slow model approximation is not valid when a boundary is crossed. Since proving continuity of $\hat{p}$ with respect to x_1 may be difficult, it may suffice for many practical applications to show only that the mean and covariance of z are continuous as a function of x_1 if the higher order moments are negligible.

Once an expression in (36) is obtained, it can be substituted into (34) and the resulting expression then substituted into (31) to yield an expression for $p(x,z;t+\Delta|x_1,z_1;t)$:

$$\begin{aligned} p(x,z;t+\Delta|x_1,z_1;t) &= \\ &\frac{1}{4\pi^2}\iiint_{-\infty}^{+\infty} e^{-jvx-jwz}\, e^{jvx_1+jwy}[1-jvf_1(x_1,z_1)\Delta+\frac{v^2}{2}g_1{}^2Q\Delta] \\ &\times \hat{p}(y|x_1)\, dydvdw \;+\; O(\mu) \end{aligned} \tag{37}$$

This is then substituted into the Chapman-Kolmogorov equation (28) to yield

$$\begin{aligned} p(x,z;t+\Delta|x_0,z_0;t) &= \\ &\frac{1}{4\pi^2}\iint_{-\infty}^{+\infty}\iint_{-\infty}^{+\infty} e^{-jvx-jwz}\, e^{jvx_1}\,[1-jvf_1(x_1,z_1)\Delta+\frac{v^2}{2}g_1{}^2Q\Delta] \\ \times &\int_{-\infty}^{+\infty} e^{jwy}\hat{p}(y|x_1)p(x_1,z_1;t|x_0,z_0;t)\, dydvdwdx_1dz_1 + O(\mu) \end{aligned} \tag{38}$$

Note that the $O(\mu)$ term is placed outside of the integrals since all of the integrals correspond to taking expectation or transformation. The expected value of the perturbed quantity is a perturbation of the expected value of the quantity.

Since the integrand in equation (38) is continuous, the order of integration may be interchanged. Integration with respect to w yields

$$p(x,z;t+\Delta|x_0,z_0;t) =$$

$$\frac{1}{2\pi}\int\cdots\int_{-\infty}^{+\infty} e^{-jvx}\, e^{jvx_1}\, [1-jvf_1(x_1,z_1)\Delta+ \frac{v^2}{2} g_1{}^2Q\Delta]$$

$$\times \quad \hat{p}(y|x_1)p(x_1,z_1;t|x_0,z_0;t)\ \delta(z-y)\ dydvdx_1dz_1 + O(\mu) \qquad (39)$$

Integration with respect to y yields the following:

$$p(x,z;t+\Delta|x_0,z_0;t) =$$

$$\frac{1}{2\pi}\iint_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-jvx}\, e^{jvx_1}\, [1-jvf_1(x_1,z_1)\Delta+ \frac{v^2}{2} g_1{}^2Q\Delta]$$

$$\times\ \hat{p}(z|x_1)p(x_1,z_1;t|x_0,z_0;t)\ dvdx_1dz_1 + O(\mu) \qquad (40)$$

Integration with respect to z_1 yields

$$p(x,z;t+\Delta|x_0,z_0;t) = \frac{1}{2\pi}\iint_{-\infty}^{+\infty} e^{-jvx}\, e^{jvx_1}\, [1-jvf_1(x_1,\overline{z}_1)\Delta+ \frac{v^2}{2} g_2{}^2Q\Delta]$$

$$\times\ \hat{p}(z|x_1)p(x_1;t|x_0,z_0;t)\ dvdx_1 \ + O(\mu) \qquad (41)$$

where $\overline{z}_1$ is defined by

$$0 = f_2(x_1,\overline{z}_1) \qquad (42)$$

To obtain $p(x;t+\Delta|x_0,z_0;t)$, integrate (41) with respect to z on both sides to yield

$$p(x;t+\Delta|x_0,z_0;t) = \frac{1}{2\pi}\iint_{-\infty}^{+\infty} e^{-jvx}\, e^{jvx_1}\, [1-jvf_1(x_1,\bar{z}_1)\Delta + \frac{v^2}{2} g_1{}^2Q\Delta]$$

$$\times\ p(x_1;t|x_0,z_0;t)\ dvdx_1 \ + O(\mu) \qquad (43)$$

The expression in (43) is evaluated with the following result:

$$p(x;t+\Delta|x_0,z_0;t) = p(x;t|x_0,z_0;t) - \frac{\partial}{\partial x} f_1(x,\bar{z})p(x;t|x_0,z_0;t)\Delta$$

$$+ \frac{1}{2} g_1{}^2Q \frac{\partial^2}{\partial x^2}p(x;t|x_0,z_0;t)\Delta + O(\mu) \quad \text{a.e.} \qquad (44)$$

where $\bar{z}$ is defined by $0=f_2(x,\bar{z})$. As $\Delta\to 0$, this expression becomes the usual Fokker-Planck equation where $p=p(x,t|x_0,z_0;t_0)$:

$$\frac{\partial p}{\partial t} = - \frac{\partial}{\partial x} [f_1(x,\bar{z})p] + \frac{1}{2} g_1{}^2Q \frac{\partial^2 p}{\partial x^2} + O(\mu) \quad \text{a.e.} \qquad (45)$$

The initial condition (in a distributional sense) and auxiliary conditions are

$$p(x;t_0|x_0,z_0;t_0) = \delta(x-x_0);\ p\geq 0;\ \int_{-\infty}^{+\infty} pdx = 1 \qquad (46)$$

The Fokker-Planck equation of the slow approximation is given by

$$\frac{\partial p_s}{\partial t} = - \frac{\partial}{\partial x}[f_1(x_s,z_s)p_s] + \frac{1}{2} g_1{}^2\ Q\ \frac{\partial^2 p_s}{\partial x^2} \qquad \text{a.e.} \qquad (47)$$

where $p_s=p_s(x;t|x_0;t_0)$ and $p_s(x;t_0|x_0;t_0)=\delta(x-x_0)$ (in a distributional sense). Since (45) is a regular perturbation of

(47) satisfying the same initial and auxilary conditions, the solutions are found to differ by $O(\mu)$, i.e.,

$$p(x;t|x_0;t_0) = p_s(x_s;t|x_0;t_0) + O(\mu) \tag{48}$$

Hence, for statistical purposes, x can be approximated by x_s with approximation errors in the probability density functions of order $O(\mu)$ in distribution.

4. EXTENSION TO VECTOR VARIABLES

The results of Sections 2 and 3 are directly extendable to the vector variable case. The fast subsystem approximation is given by equation (7) where $\hat{z}$ and x_0 are vectors variables ($\hat{z} \varepsilon R^r$ and $x_0 \varepsilon R^m$) and f_2 and g_2 are vectors of appropriate length. Similarly, the slow subsystem is given by equation (28) where x_s and z_s ($x_s \varepsilon R^m$ and $z \varepsilon R^r$) are vector variables and f_1, f_2, g_1 are vector valued functions of appropriate length.

It is still possible to show that the errors between the probability density functions of the fast subsystem and the solution are of order $O(\mu^{\frac{1}{2}})$. The Fokker-Planck equation for the true probability density function is generalized from the previous case as follows:

$$\begin{aligned}\frac{\partial p}{\partial \tau} = &- \sum_{i=1}^{r} \frac{\partial}{\partial z_i}[f_2(x,z)p] + \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{r} (g_2{}^TQg_2)_{i,j} \frac{\partial^2}{\partial z_i \partial z_j} p \\ &- \mu \sum_{i=1}^{m} \frac{\partial}{\partial x_i} [f_1(x,z)p] + \mu \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (g_1{}^TQg_1)_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} p \\ &+ \mu^{\frac{1}{2}} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{r} (g_1{}^TQg_2)_{i,j} \frac{\partial^2}{\partial x_i \partial z_j} p \\ &+ \mu^{\frac{1}{2}} \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{m} (g_2{}^TQg_1)_{i,j} \frac{\partial^2}{\partial z_i \partial x_j} p\end{aligned} \tag{49}$$

where $p=p(x,z;\tau|x_0,z_0;0)$ and x_i and z_j are components of the $\tilde{x}$ and $\tilde{z}$ vectors. The propagation of p is relatively insensitive to variation in any of the components of the x vector. Hence, following the previous analysis, it can be easily seen that the probability density function of the fast subsystem approximates that of the true solution. For systems linear in z, the analysis in Section 2.1 is extended in a straighforward manner.

Most of the analysis in Section 3 for the slow subsystem involves the derivation of the Fokker-Planck equation. Since the Fokker-Planck equation has been derived for the vector case in [11], these steps generalize accordingly. The only difference between the analysis in Section 3 and previous work is the solution for $p(z;t+\Delta|x_1,z_1;t)$ in (34). This can be approximated by the steady-state of the probability density function for the fast subsystem found from the corresponding Fokker-Planck equation.

5. SUMMARY

Reduced-order models are developed in this paper which approximate the original system both in the fast time scale and in the slow time scale. It is shown that the approximations are valid in terms of the statistical information of the true solution. It is further shown for systems that are linear in the fast variable, that the fast subsystem approximates the true solution with a mean-squared error of order $O(\mu)$.

ACKNOWLEDGEMENT

This research was completed while both authors were at the Georgia Institute of Technology. The research was supported, in part, by the U.S. Air Force under Grant AFOSR-87-0308.

REFERENCES

[1] P.V. Kokotovic, R.E. O'Malley and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--an Overview," Automatica, vol. 12, March 1976, pp. 123-132.

[2] P.V. Kokotovic, H.K. Khalil and J. O'Reilly, Singular Perturbation Methods in Control: Analysis and Design, Academic Press, London, 1986.

[3] J.J. Levin, "The Asymptotic Behavior of the Stable Initial Manifolds of a System of Nonlinear Differential Equations," Trans. Am. Math. Soc., vol 85, 1957, pp. 357-368.

[4] B.S. Heck and A.H. Haddad, "Singular Perturbation Theory for Piecewise-Linear Systems," IEEE Trans. Auto. Control, vol. AC-34, Jan. 1989, pp. 87-90.

[5] B.S. Heck and A.H. Haddad, "Singular Perturbation Analysis for Linear Systems with Quantized Control," Automatica, vol. 24, Nov. 1988, pp. 755-764.

[6] A.H. Haddad and P.V. Kokotovic, "Stochastic Control of Linear Singularly Perturbed Systems," IEEE Trans. Auto. Control, vol. AC-22, Oct. 1977, pp. 815-821.

[7] A.H. Haddad, "Linear Filtering of Singularly Perturbed Systems," IEEE Trans. Auto. Control, vol. AC-21, August 1976, pp. 515-519.

[8] M. Ansary and H. Khalil, "Reduced-Order Modeling of Nonlinear Singularly Perturbed Systems Driven by Wide-Band Noise," Proc. Twenty-first CDC, Orlando, Dec. 1982, pp. 1090-1094.

[9] V.D. Razvig, "Reduction of Stochastic Differential Equations with Small Parameters and Stochastic Integrals," Int. J. Control, vol. 28, 1978, pp. 707-720.

[10] H.K. Khalil, A.H. Haddad and G.L. Blankenship, "Parameter Scaling and Well-Posedness of Stochastic Singularly Perturbed Control Systems," Proc. 12th Asilomar Conf. on Circuits, Systems and Computers, Pacific Grove, CA, Nov. 1978.

[11] E. Wong and B. Hajek, Stochastic Processes in Engineering Systems, Springer-Verlag, New York, 1985.

[12] T. Kato, Perturbation Theory for Linear Operators, Springer-Verlag, New York, 1976.

[13] H.L. Royden, Real Analysis, Macmillan Publishing Co., Inc., New York, 1968, pp. 87.

APPENDIX I

B. S. Heck and A. H. Haddad, "Singular Perturbation Analysis for Linear Systems with Vector Quantized Control", Proc. 1989 American Control Conference, Pittsburgh, PA, pp. 2178-2183, June 1989.

FA10 - 11:15

SINGULAR PERTURBATION ANALYSIS FOR LINEAR SYSTEMS WITH VECTOR QUANTIZED CONTROL

Bonnie S. Heck
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Abraham H. Haddad
Department of EE/CS
Northwestern University
Evanston, Illinois 60208

Abstract

In this paper, the analysis for a singularly perturbed linear system with quantization in the feedback loop is performed. It is found that the system has variable structure and can exhibit sliding behavior on the switching surfaces. Because the system is nonsmooth and standard singular perturbation techniques are not applicable, a new technique is developed for a two-input case to obtain the boundary layer solution and the outer solution. A discussion of the approximation error is included. The technique developed is successfully illustrated on a numerical example.

1. Introduction

Singular perturbation theory is an asymptotic approximation scheme used to simplify systems which contain both fast and slow dynamics. These types of systems, termed "numerically stiff," are often difficult to analyze numerically due to ill-conditioning in the system matrices. Singular perturbation theory removes the numerical problem by separating the system into reduced-order models, one containing the fast dynamics and one containing the slow dynamics. This theory has received considerable attension in the past thirty years (see the surveys given in References [1-3]). However, the common restriction placed on systems for using singular perturbation theory is that the system dynamics must be smooth [1-3].

In many systems, the actuators supply inputs with discrete rather than continuous values, i.e. the input is quantized. Examples of these types of actuators include relays, stepper motors, and certain types of hydraulic and pneumatic devices [4,5]. The resulting control is discontinuous with respect to the state variable, hence the system is nonsmooth and standard singular perturbation techniques are not applicable. The basic theory of singular perturbation is extended in this paper to the case of a quantizer existing in a two-dimensional feedback loop. The scalar case was developed separately in Reference [6]. Note that the discontinuous control causes the system to be a variable structure system.

Intuitively, discontinuities in the control would seem to excite the fast dynamics in the same way as would a step input. However, it is found that under very mild restrictions, the slow system slides along the switching surface instead of crossing through it. Therefore, there are no jumps in the quasi-steady-state solution which would cause the fast dynamics to respond with a step response outside of the initial boundary layer.

1.1 Problem Formulation

The system under consideration is assumed to be linear and time-invariant. It was determined in [6] that, for the scalar quantized control case, the system may be transformed into decoupled coordinates and the singular perturbation analysis performed on the decoupled coordinates. For the purposes of this paper, it is also assumed that the system may be transformed to decoupled coordinates and is represented by:

$$\dot{\xi} = A_0\xi + B_0u, \qquad \xi(0) = \xi_0 \tag{1}$$

$$\mu\dot{\eta} = A_2\eta + B_2u, \qquad \eta(0) = \eta_0 \tag{2}$$

where $\xi \in R^p$, $\eta \in R^r$, $u \in R^2$, $\mu > 0$ is small and A_2 is Hurwitz. Define the control vector to be

$$u = \begin{bmatrix} q_1(-K_{11}\xi-K_{12}\eta) \\ q_2(-K_{21}\xi-K_{22}\eta) \end{bmatrix} \tag{3}$$

where K_{11}, K_{12}, K_{21} and K_{22} are row vectors and q_1 and q_2 are quantizer functions defined as follows.

$$\begin{aligned} q_1(x) &= c_{1,i} \text{ for } d_{1,i} \le x < d_{1,i+1}; \; i=1,\ldots,n \\ q_2(x) &= c_{2,j} \text{ for } d_{2,j} \le x < d_{2,j+1}; \; j=1,\ldots,k \end{aligned} \tag{4}$$

The parameters are specified such that $c_{1,i} < c_{1,i+1}$, $c_{2,i} < c_{2,i+1}$, $d_{1,i} < d_{1,i+1}$, $d_{2,i} < d_{2,i+1}$, $d_{1,1} = -\infty$, $d_{2,1} = -\infty$, $d_{1,n+1} = +\infty$ and $d_{2,k+1} = +\infty$.

This system is a variable structure system with nk possible linear subsystems. The state space (R^{p+r}) can be partitioned into nk nonoverlapping regions defined by

$$\begin{aligned} S_{ij} = \{(\xi,\eta)\text{: } & q_1(-K_{11}\xi-K_{12}\eta) = c_{1,i} \text{ and} \\ & q_2(-K_{21}\xi-K_{22}\eta) = c_{2,j}\} \end{aligned} \tag{5}$$

The boundary between two regions is a convex partition of a hyperplane defined by $-K_{11}\xi-K_{12}\eta = d_{1,i}$ for some i or $-K_{21}\xi-K_{22}\eta = d_{2,j}$ for some j. Note that if $K_{11} \neq K_{21}$ and $K_{12} \neq K_{22}$ then each of the regions (except those of S_{1j} $\forall j$, S_{i1} $\forall i$, S_{nj} $\forall j$, and S_{ik} $\forall i$) is bordered by two sets of parallel hyperplanes.

For the purpose of this paper, it is assumed that K_{12} and K_{22} are not of order $O(\mu)$. If either of the quantities was of order $O(\mu)$ then the slow manifold would be nearly discontinuous. In particular, the quasi-steady-state solution would be a discontinuous function of the slow variable. Hence, the fast dynamics would have to be evaluated after each switch in the corresponding control component.

Reduced-order models of the system described above are developed in this paper using a singular perturbation approach. The boundary layer approximation is given in Section 2. Section 3 presents the outer solution approximation. A numerical example is given in Section 4 and concluding remarks are given in Section 5.

2. Boundary Layer Solution

The fast dynamics are most prominent during the initial boundary layer and can be separated from the slow dynamics by the introduction of an expanded time-scale $\tau=(t-t_0)/\mu$. It can be easily shown that ξ stays relatively constant with respect to τ; hence, ξ is approximated by ξ_0. The approximation for η is given by the solutiōn to the following equation.

$$\frac{d\hat{\eta}}{d\tau} = A_2\hat{\eta} + B_2u_f; \qquad \hat{\eta}(0) = \eta_0 \tag{6}$$

$$u_f = \begin{bmatrix} q_1(-K_{11}\xi_0-K_{12}\hat{\eta}) \\ q_2(-K_{21}\xi_0-K_{22}\hat{\eta}) \end{bmatrix}$$

This reduced-order model is also a variable structure system with nk possible linear subsystems. Hence, the reduced-order state space (R^r) can be partitioned into nk nonoverlapping regions for which the system is linear. Analogously with the full-order model, the regions R_{ij} are defined as:

$$R_{ij} = \{\hat{\eta}: q_1(-K_{11}\xi_0-K_{12}\hat{\eta})=c_{1,i} \text{ and } q_2(-K_{21}\xi_0-K_{22}\hat{\eta})=c_{2,j}\} \tag{7}$$

Correspondingly, the boundaries between regions are hyperplanes defined by $-K_{11}\xi_0-K_{12}\hat{\eta} = d_{1,i}$ for some i or $-K_{21}\xi_0-K_{22}\hat{\eta} = d_{2,j}$ for some j.

It is assumed that the system in (6) is asymptotically stable to one equilibrium point. Stability in "ordinary" smooth system can be shown by use of Lyapunov's second method. However, in variable structure systems the Lyapunov function is generally discontinuous and, hence, not everywhere differentiable. Paden and Sastry introduced a generalized Lyapunov theorem in [7] which is suitable for discontinuous functions. Such a method may be useful in determining asymptotic stability of this system. The equilibrium point of (6) (also defined as quasi-steady-state solution at t=0, $\eta_s(0)$) is derived below and a discussion of the approximation error follows.

2.1 Evaluation of the Equilibrium Point

There are three basic positions for the equilibrium point of (6): in the interior of a region, on single boundary hyperplane, or on an intersection between two boundary hyperplanes. The first two cases are treated very similarly to the scalar quantized control analysis discussed in Reference [6]. The last case is more complicated and can be solved uniquely only for limited types of systems. The derivation of the equilibrium point as a function of $K_{11}\xi_0$ and $K_{22}\xi_0$, $f(K_{11}\xi_0,K_{22}\xi_0)$, is shown in the next three subsections for the three possible positions.

2.1.1 Interior Position. The piecewise-linearity of the system in (6) is utilized in determining the system behavior; i.e., for $\hat{\eta}$ in the R_{ij} region of the state space, the system is given by the following description.

$$\frac{d\hat{\eta}}{d\tau} = A_2\hat{\eta} + B_2\begin{bmatrix} c_{1,i} \\ c_{2,j} \end{bmatrix} \tag{8}$$

The behavior in this region is governed by the position of the following point which is termed the regional equilibrium point.

$$\eta_{ij} = -A_2^{-1} B_2\begin{bmatrix} c_{1,i} \\ c_{2,j} \end{bmatrix} \tag{9}$$

If η_{ij} lies in R_{ij} (i.e., satisfies (7)), then it is a local equilibrium point. If η_{ij} does not lie in R_{ij}, then points in R_{ij} are directed out of the region.

2.1.2 Single Boundary Position. An equilibrium point may lie on a boundary between two regions if trajectories from the two bordering regions head toward the boundary. To find the conditions for such an occurrence, the specific example of an equilibrium point existing on the boundary between R_{ij} and $R_{i+1,j}$ is examined. The control $u_2=c_{2,j}$ is constant across this boundary, but u_1 switches between $c_{1,i}$ and $c_{1,i+1}$. Suppose the following condition is satisfied.

$$d_{1,i+1} + K_{12}\eta_{ij} \le -K_{11}\xi_0 < d_{1,i+1} + K_{12}\eta_{i+1,j} \tag{10}$$

where $\eta_{i+1,j}$ and η_{ij} are regional equilibrium points for $R_{i+1,j}$ and R_{ij} respectively. This condition states that each of the regional equilibrium points of the two bordering regions lies across the boundary hyperplane from its associated region. Hence, trajectories in $R_{i+1,j}$ and in R_{ij} move toward that hyperplane. If the regions were parallel then this would be a sufficient condition for the equilibrium point to lie on that boundary hyperplane. However, with nonparallel regions it is possible for the representative point to move (or slide) along the hyperplane until it reaches a position where the hyperplane no longer borders R_{ij} and $R_{i+1,j}$. Hence, to find the equilibrium point, first assume that the regions are parallel and find the equilibrium point to such a system. Then, if that candidate equilibrium point lies on the part of the hyperplane that borders the actual regions, it is a true equilibrium point.

The equivalent control method which was developed for use in variable structure systems [8] is used to find the candidate equilibrium point. If representative points from both sides of the boundary are directed toward the boundary, then the control u_1 starts switching very quickly between $c_{1,i}$ and $c_{1,i+1}$ while $u_2=c_{2,j}$ remains constant. The system filters out the high frequency leaving only the low frequency average component. The average value of u_1 as $\tau\to+\infty$ is the equivalent control at the equilibrium point, u_e. The corresponding candidate equilibrium point, η_s, is found to be:

$$\eta_s = \frac{(K_{11}\xi_0+d_{1,i+1})(\eta_{ij}-\eta_{i+1,j})}{K_{12}(\eta_{i+1,j}-\eta_{ij})} + \frac{(K_{12}\eta_{i+1,j})\eta_{ij}-(K_{12}\eta_{ij})\eta_{i+1,j}}{K_{12}(\eta_{i+1,j}-\eta_{ij})} \tag{11}$$

This is a true equilibrium point if it lies on the part of the hyperplane that borders R_{ij} and $R_{i+1,j}$. This corresponds to satisfying the following condition:

$$d_{2j} + K_{22}\eta_s \le -K_{21}\xi_0 < d_{2,j+1} + K_{22}\eta_s \tag{12}$$

Hence, if the conditions in equation (10) and (12) are satisfied, then the equilibrium point is given by equation (11). These conditions are illustrated in the second-order example given in Figure 1. Note that the equilibrium point lies on the intersection of the line connecting η_{ij} and $\eta_{i+1,j}$ and the hyperplane.

The equilibrium point existing on a boundary between R_{ij} and $R_{i,j+1}$ is found using a dual argument. The conditions in (10) and (12) correspond to the followin conditions, respectively.

$$d_{2,j+1}+K_{22}\eta_{ij} \le -K_{21}\xi_0 < d_{2,j+1}+K_{22}\eta_{i,j+1} \quad \text{and} \tag{13}$$
$$d_{1i} + K_{12}\eta_s \le -K_{11}\xi_0 < d_{1,i+1} + K_{12}\eta_s \tag{14}$$

where the equilibrium point is given by:

$$\eta_s = \frac{(K_{21}\xi_0+d_{2,j+1})(\eta_{ij}-\eta_{i,j+1})}{K_{22}(\eta_{i,j+1} - \eta_{ij})} + \frac{(K_{22}\eta_{i,j+1})\eta_{ij} - (K_{22}\eta_{ij})\eta_{i,j+1}}{K_{22}(\eta_{i,j+1} - \eta_{ij})} \tag{15}$$

Therefore, if the conditions in (13) and (14) are satisfied, then the equilibrium point is given by (15).

2.1.3 Boundary Intersection Position. Another possible position for the equilibrium point is on the intersection between two boundary hyperplanes. To find the conditions for this occurrence it is known that neither the conditions for (11) or for (15) to be the equilibrium point can be satisfied since it is assumed that there exists only one equilibrium point. There are two separate additional conditions on the system each of which yield an equilibrium point on the intersection of boundaries.

$$\begin{aligned} &d_{1,i+1} + K_{12}\eta_{ij} \le -K_{11}\xi_0 < d_{1,i+1} + K_{12}\eta_{i+1,j+1} \\ &\text{and} \\ &d_{2,j+1} + K_{22}\eta_{i+1,j} \le -K_{21}\xi_0 < d_{2,j+1} + K_{22}\eta_{i,j+1} \end{aligned} \tag{16}$$

or

$$\begin{aligned} &d_{1,i+1} + K_{12}\eta_{i,j+1} \le -K_{11}\xi_0 < d_{1,i+1} + K_{12}\eta_{i+1,j} \\ &\text{and} \\ &d_{2,j+1} + K_{22}\eta_{ij} \le -K_{21}\xi_0 < d_{2,j+1} + K_{22}\eta_{i+1,j+1} \end{aligned} \tag{17}$$

A second-order example of a system satisfying the condition in (16) is given in Figure 2.

The equilibrium point for system (6) is not, in general, found uniquely when either the condition in (16) or (17) is satisfied. By definition of the quantizer function, for the equilibrium point to exist on the intersection between two boundary hyperplanes, then η_s must be a solution to the following equation.

$$\begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix} \eta = \begin{bmatrix} -d_{1,i+1} - K_{11}\xi_0 \\ -d_{2,j+1} - K_{21}\xi_0 \end{bmatrix} \quad \text{for some i,j} \tag{18}$$

For a second-order system with an invertible $[K_{12}^T, K_{22}^T]^T$ matrix, a unique solution for η_s is found to be given by

$$\eta_s = \begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}^{-1} \begin{bmatrix} -d_{1,i+1} - K_{11}\xi_0 \\ -d_{2,j+1} - K_{21}\xi_0 \end{bmatrix} \tag{19}$$

If it can be determined that the system is sliding on the intersection of the hyperplanes, which is the surface defined by the solution η of equation (18), then a unique solution for η_s may exist for larger order systems. Using notation common to variable structure system theory, define an affine functional s as:

$$s = -K\eta + v \tag{20}$$

where K and v are determined from (18) such that the solution of $s(\eta)=0$ is the switching surface. A sliding mode exists on that surface if $s^T\dot{s}<0$ [8]. The equilibrium point of the sliding mode can be found by first transforming the system into regular form, then, finding the equilibrium point for the sliding mode equation in the new coordinates and, finally, transforming the equilibrium point back to the original coordinates. The equilibrium point of a sliding mode existing on the intersection of the two boundary hyperplanes is given below (for details of the derivation, see [9]).

$$\eta_s = -(S_1 - S_2(KS_2)^{-1}KS_1)A_s^{-1}T_1A_2S_2(KS_2)^{-1}v + S_2(KS_2)^{-1}v \tag{21}$$

where: $A_s = T_1A_2S_1 + T_1A_2S_2(KS_2)^{-1}(-KS_1)$, $T_1=U_2^T$, $S_1=U_2$, $S_2=U_1$ and U_1 and U_2 are obtained from a singular value decomposition of B_2, i.e. $B_2=[U_1, U_2]\Sigma V^T$.

2.1.4 Function for the Equilibrium Point. In summary, the equilibrium point η_s for the system in (6) can be found as a mapping of the variables $K_{11}\xi_0$ and $K_{21}\xi_0$, $\eta_s=f(K_{11}\xi_0,K_{21}\xi_0)$, from equations (9), (11), (15), (19), and (21) depending on the conditions that are satisfied. The mapping is a function (i.e., single-valued) since it is assumed that there exists an unique equilibrium point. The function is piecewise-linear since each of the function definitions is linear. Continuity is not guaranteed and may need to be evaluated on a case by case basis. However, using the methods in [6] for the scalar control case, it is straightforward to show that the function is continuous for a second-order system. To demonstrate the use of this function, an example is given below.

Example: Let a system be given by

$$\frac{d\eta}{d\tau} = \begin{bmatrix} -8 & 4 \\ 0 & -4 \end{bmatrix} \eta + \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} u \tag{22}$$

The control selected is bang-bang with components:

$$\begin{aligned} u_1 &= \text{sgn}\, s_1; \quad s_1 = -[1, 2]\eta - \beta \\ u_2 &= \text{sgn}\, s_2; \quad s_2 = -[-2, 2]\eta - \gamma \end{aligned} \tag{23}$$

where the sgn function is defined below:

$$\text{sgn}\, s = \begin{cases} 1 & \text{if } s \ge 0 \\ -1 & \text{if } s < 0 \end{cases} \tag{24}$$

where $s_1(\eta)=0$ and $s_2(\eta)=0$ define the switching hyperplanes and β and γ are parameters. To correspond to the previous notation, let

$$K_{12}=[1, 2]; \quad K_{22}=[-2, 2]; \quad \beta = K_{11}\xi_0; \quad \gamma = K_{21}\xi_0;$$

$$A_2 = \begin{bmatrix} -8 & 4 \\ 0 & -4 \end{bmatrix}; \quad B_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \tag{25}$$

$$c_{1,1} = c_{2,1} = -1, \quad c_{1,2} = c_{2,2} = 1, \quad d_{1,2} = d_{2,2} = 0$$

Define the regions as

$$\begin{aligned} R_{11} &= \{\eta: u_1=-1 \text{ and } u_2=-1\} = \{\eta: \beta>-K_{12}\eta \text{ and } \gamma>-K_{22}\eta\} \\ R_{12} &= \{\eta: u_1=-1 \text{ and } u_2=1\} = \{\eta: \beta>-K_{12}\eta \text{ and } \gamma\le-K_{22}\eta\} \\ R_{21} &= \{\eta: u_1=1 \text{ and } u_2=-1\} = \{\eta: \beta\le-K_{12}\eta \text{ and } \gamma>-K_{22}\eta\} \\ R_{22} &= \{\eta: u_1=1 \text{ and } u_2=1\} = \{\eta: \beta\le-K_{12}\eta \text{ and } \gamma\le-K_{22}\eta\} \end{aligned}$$

The regional equilibrium points are found to be

$$\eta_{11} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}; \quad \eta_{12} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \quad \eta_{21} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}; \quad \eta_{22} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{26}$$

The equilibrium point can be found as a function of β and γ using the function definitions given in equations (9), (11), (15), (19) and (21). For a regional equilibrium point to lie inside its associated region, one of the following conditions must hold:

a) if $\beta>3$ and $\gamma>0$ then $\eta_{11}\in R_{11}$ and $\eta_s=\eta_{11}$
b) if $\beta>-2$ and $\gamma\leq-2$ then $\eta_{12}\in R_{12}$ and $\eta_s=\eta_{12}$
c) if $\beta\leq2$ and $\gamma>2$ then $\eta_{21}\in R_{21}$ and $\eta_s=\eta_{21}$
d) if $\beta\leq-3$ and $\gamma\leq0$ then $\eta_{22}\in R_{22}$ and $\eta_s=\eta_{22}$

Note that these are mutually exclusive conditions.

For an equilibrium point to lie on a boundary, either conditions (10) and (12) or (13) and (14) are satisfied. The following conditions and corresponding equilibrium points are given for η_s to lie on the boundaries between R_{11} and R_{21}, R_{12} and R_{22}, R_{11} and R_{12}, and R_{21} and R_{22}, respectively.

e) if $3\geq\beta>2$ and $6<2\beta+\gamma$ then $\eta_s = \beta\begin{bmatrix}-1\\0\end{bmatrix}+\begin{bmatrix}2\\-1\end{bmatrix}$

f) if $-2\geq\beta>-3$ and $-6\geq2\beta+\gamma$ then $\eta_s = \beta\begin{bmatrix}-1\\0\end{bmatrix}+\begin{bmatrix}-2\\1\end{bmatrix}$

g) if $0\geq\gamma>-2$ and $3<\beta-\gamma(5/2)$ then $\eta_s = \gamma\begin{bmatrix}-\frac{1}{2}\\-1\end{bmatrix}+\begin{bmatrix}-1\\-1\end{bmatrix}$

h) if $2\geq\gamma>0$ and $-3\geq\beta-\gamma(5/2)$ then $\eta_s = \gamma\begin{bmatrix}-\frac{1}{2}\\-1\end{bmatrix}+\begin{bmatrix}1\\1\end{bmatrix}$

If the equilibrium point of (22) lies on the intersection between s_1 and s_2, then the following must be true.

i) if the conditions in a)-h) are not satisfied and
$3\geq\beta>-3$ and $2\geq\gamma>-2$ then

$$\eta_s = \begin{bmatrix}-1/3 & 1/3\\-1/3 & -1/6\end{bmatrix}\begin{bmatrix}\beta\\\gamma\end{bmatrix}$$

The domain space of the mapping $\eta_s=f(\beta,\gamma)$ can be partitioned into ten nonoverlapping regions each corresponding to a function definition a)-i). It can be shown easily that this mapping is a function since none of the regions overlap. Also, the function is continuous. This can be shown easily by noting that the function is continuous within each partition of the β-γ space. It is straightforward to show that on any boundary between two partitions the two function definitions are equal.

2.2 Boundary Layer Approximation Error

The errors introduced by approximating the true solution by the solution of (6) are due entirely to the assumption that $\xi\approx\xi_0$ in the boundary layer. As in the other piecewise-linear systems discussed in [6,10], the approximation errors are of order $O(\mu)$ for the time intervals when both the actual solution and the approximate solution exist within the same region of the R^{p+r} state space due to linearity. When a single boundary hyperplane is crossed, then the previous results on the approximation errors in the scalar quantized control case are applicable. That is, if the vector field does not intersect the boundary with an angle of order $O(\mu)$, then the approximation error remains of order $O(\mu)$. However, if the solution crosses a boundary hyperplane within an $O(\mu)$ neighborhood of an intersection between boundary hyperplanes, this result cannot be used.

There are certain problems introduced by allowing intersections of switching boundaries to exist in the system definition. If the actual solution crosses a boundary near an intersection, then the approximation may not cross into the same region. From that point, there is no guarantee that the approximation error remains of order $O(\mu)$. A consolation in this is that if the system is not sliding on switching surface, then the chances of hitting a boundary within an $O(\mu)$ neighborhood of the intersection for an arbitrary initial condition is of order $O(\mu)$. The exception to this is when the equilibrium point lies on the intersection. In that case, the solution must eventually travel into the $O(\mu)$ neighborhood about that point and will cross the boundaries there. However, if the function f which defines the equilibrium point is continuous, then $O(\mu)$ approximation errors in ξ result in $O(\mu)$ errors between the equilibrium point, η_s, and the actual value of $\eta(\tau)$ as $\tau\rightarrow+\infty$. Therefore, if the solutions differed by an amount of order $O(\mu)$ prior to entering the small neighborhood about η_s, then continuity of f implies that the error will remain of order $O(\mu)$.

3. Outer Solution

The outer solution is found by neglecting the fast dynamics. It is assumed that the fast variable, η, reaches a quasi-steady-state value within the initial boundary layer. This initial quasi-steady-state value is, of course, the equilibrium point of the fast subsystem (6). The quasi-steady-state solution, $\eta_s(t)$, cannot be found by simply setting $\mu=0$ in equation (2) as is done in standard singular perturbation techniques, because the solution η_s to the resulting equation is undefined for values of (ξ_s,η_s) that lie on a switching boundary in the state space. Instead, η_s, is found as an equilibrium point of the fast subsystem using the function defined in Section 2.1, i.e. $\eta_s(t) = f(K_{11}\xi_s(t),K_{21}\xi_s(t))$.

Similarly, the control in the slow time-scale, u_s, cannot be obtained by simply substituting ξ_s and η_s for ξ and η in the original definition of the control (3) because the control is undefined for values of (ξ_s,η_s) that lie on a switching boundary. In ordinary systems with discontinuous control, the system chatters along the sliding surface causing the control to switch at a very high frequency. The average value of the control (i.e. the equivalent control) determines the motion of the system along the sliding surface. In this application, however, the system does not chatter as seen from the definition of η_s. Thus, the slow control must be given as the equivalent control for those values of (ξ_s,η_s) that lie on a switching boundary. Alternately, it can be given as the final value of the equivalent control in the fast time-scale.

$$u_s = -(B_2{}^TB_2)^{-1}B_2{}^TA_2\eta_s \qquad (27)$$

Note that the use of the psuedo-inverse is justified by the consistancy of the equation, i.e., η_s was derived from the equivalent control.

Thus, the solution of the system (1)-(2) can be approximated outside of the boundary layer by the solution to the following system:

$$\dot{\xi}_s = A_0\xi_s + B_0u_s;\ \ \xi_s(0) = \xi_0 \qquad (28)$$

$$\eta_s = f(K_{11}\xi_s,K_{21}\xi_s)$$

where u_s is given in equation (27) and the function f is defined from (9),(11),(15),(19) and (21). As stated previously, f is piecewise-linear and may be continuous; hence, the slow manifold defined by the function f is piecewise-linear and may be continuous.

The continuity of the slow manifold is required for the approximation errors given by $\eta(t)-\eta_s(t)$ and $\xi(t)-\xi_s(t)$ to be of order $O(\mu)$ for the time outside of the initial boundary-layer. If the function defining the slow manifold is continuous, it satisfies a Lipschitz condition since it is piecewise-linear [11]. Therefore, a sufficient condition for the approximation error to be of order $O(\mu)$ is that the Lipschitz constant be bounded as $\mu \to 0$, i.e. it cannot be of order $O(1/\mu)$. If it was of order $O(1/\mu)$, then the equilibrium point for the fast dynamics would change too quickly in the t-time scale thereby invalidating the separation in time scales between t and τ. The resulting behavior would require evaluation of the fast dynamics after each switch in the control, and the slow model approximation in (28) would be valid only in the time-intervals between switches.

4. Numerical Example

An example of the approximation method for the two-input quantized control system is demonstrated here. The system is given in the form of equations (1)-(2). The control is selected to be bang-bang with components

$$\begin{aligned} u_1 &= \operatorname{sgn} s_1; \quad s_1 = -K_{11}\xi - K_{12}\eta \\ u_2 &= \operatorname{sgn} s_2; \quad s_2 = -K_{21}\xi - K_{22}\eta \end{aligned} \tag{29}$$

The parameter matrices are given as follows:

$$\begin{aligned} A_0 &= \begin{bmatrix} -3 & 1 & 0 \\ -1 & -3 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad B_0 = \begin{bmatrix} 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \\ A_2 &= \begin{bmatrix} -8 & 4 \\ 0 & -4 \end{bmatrix} \quad B_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \end{aligned} \tag{30}$$

$$\begin{aligned} K_{11} &= [2\ \ 1\ \ 1]; \; K_{12} = [1\ \ 2]; \\ K_{21} &= [1\ \ 1\ \ 2]; \; K_{22} = [-2\ \ 2] \end{aligned}$$

To correspond to the previous notation, the parameters of the quantizer functions are defined as:

$$c_{1,1} = c_{2,1} = -1; \; c_{1,2} = c_{2,2} = 1; \; d_{1,2} = d_{2,2} = 0$$

The initial conditions are $\xi_0 = [-1.5, 1, -0.75]^T$ and $\eta_0 = [1, 2]^T$ and $\mu=0.1$.

Note that the fast subsystem is the same as given in the example described in Section 2.1.4. Therefore, the boundary layer approximation is found as the solution of (22). The equilibrium point of this system, $\eta_s(0)$, is found as a function of $K_{11}\xi_0$ and $K_{21}\xi_0$ using the function definitions a)-i) listed in Section 2.1.4 where $\beta= K_{11}\xi_0$ and $\gamma=K_{21}\xi_0$. With the given initial conditions on ξ, the equilibrium point is found to be $\eta_s(0)=[-2.75, -2]^T$. This corresponds to an equilibrium point existing on the boundary between R_{12} and R_{22} so that as $\hat{\eta}$ approaches $\eta_s(0)$, the control u_1 begins switching very rapidly while $u_2=1$ remains constant. It can be shown that with the given initial conditions on η, the fast subsystem will slide in the τ-time scale on the switching surface defined by $s_1=0$.

The outer solution is found from the slow model in (28). The quasi-steady-state solution is found from the function definitions a)-i) listed in Section 2.1.4 where $\beta=K_{11}\xi_s$ and $\gamma=K_{21}\xi_s$. The initial conditions for this example are such that the system starts out sliding on the $s_1=0$ surface in the normal time-scale.

Comparison between the time-integration of the actual system and the approximate system are shown for representative states in Figures 3-4. The errors between the trajectories are of order $O(\mu)$. In the approximate solution, the boundary layer correction $(\hat{\eta}-\eta_s(0))$ is added to the outer solution for $0 \le t \le 0.2$, beyond which it is negligible. Both the approximation and the true solution are asymptotically stable to the origin. Therefore, both systems are found to slide on the intersection of the switching surfaces defined by $s_1=0$ and $s_2=0$. The computation time for obtaining the actual solution was roughly 18 times longer than that for obtaining the approximate solution for. As in all singular perturbation approaches, as μ decreases, the approximation becomes more accurate and the relative computational time-savings greatly increases.

5. Summary

This paper presents the analysis of a singularly perturbed two-input quantized control system. The discontinuities in the control occur in the state space on single boundary surfaces as well as on intersections of boundaries. This latter occurrence is precisely what makes the analysis of the two-input case so much more complicated than that of the scalar case. In spite of the complications, reduced-order models are developed which yield the outer solution and the boundary layer solution. As part of the slow model, a function is derived which solves for the quasi-steady-state solution in terms of the slow variable. This function is known to be continuous when the fast subsystem is second-order. Other cases may need to be evaluated numerically. As in the scalar case, the system may possess a sliding mode in the fast time as well as the slow time scales. The results of a numerical simulation show that the approximation method described in this paper can yield very accurate results with very good computational time-savings.

It appears that the extension of singular perurbation theory to the general multiple input quantized control is a very complex problem. In particular, finding a function which solves for the quasi-steady-state solution is very tedious and, in fact, may not be possible. Hence, in general, singular perturbation theory does not simplify the analysis. However, it may be the only alternative if the original system is too numerically stiff to be solved any other way. A discussion of the multiple input case is contained in [9].

Acknowledgement

This work was supported, in part, by the U.S. Air Force under contract AFOSR-87-0308.

References

[1] P.V. Kokotovic, R.E. O'Malley and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--an Overview," Automatica, vol. 12, pp. 123-132, March 1976.

[2] V.R. Saksena, J. O'Reilly and P.V. Kokotovic, "Singular Perturbations and Time-Scale Methods in Control Theory: Survey 1976-1983," Automatica, vol. 20, pp. 273-293, May 1984.

[3] P.V. Kokotovic, H.K. Khalil and J. O'Reilly, Singular Perturbation Methods in Control: Analysis and Design, Academic Press, London, 1986.

[4] T.J Harned, "Making Stepper Motors Behave," Machine Design, vol. 57, pp. 54-56, Sept. 1985.

[5] Parker Pneutronics, July 1984 Bulletin, Pepperall, MA.

[6] B.S. Heck and A.H. Haddad, "Singular Perturbation Analysis for Linear Systems with Scalar Quantized Control," Automatica, vol. 24, pp. 755-764, Nov. 1988.

[7] B.E. Paden and S.S. Sastry, "A Calculus for Computing Filipov's Differential Inclusion with Application to the Variable Structure Control of Robot Manipulators," Proc. IEEE Conf. on Decision and Control, Athens, Greece, pp. 578-582, Dec. 1986.

[8] V.I. Utkin, "Variable Structure Systems with Sliding Modes," IEEE Trans. Auto. Control, vol. AC-22, pp. 212-222, April 1977.

[9] B.S. Heck, On Singular Perturbation Theory for Piecewise-Linear Systems, Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, August 1988.

[10] B.S. Heck and A.H. Haddad, "Singular Perturbation in Piecewise-Linear Systems," IEEE Trans. Auto. Control, vol. AC-34, pp. 87-90, Jan. 1988.

[11] T. Fujisawa and E.S. Kuh, "Piecewise-linear Theory of Nonlinear Networks," SIAM J. Appl. Math., vol. 22, pp. 307-328, March 1972.

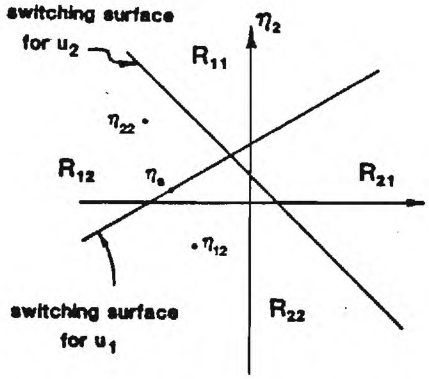

Figure 1: Graphical example of condition for equilibrium point, η_e, to lie on a boundary between $R_{i,j}$ and $R_{i+1,j}$.

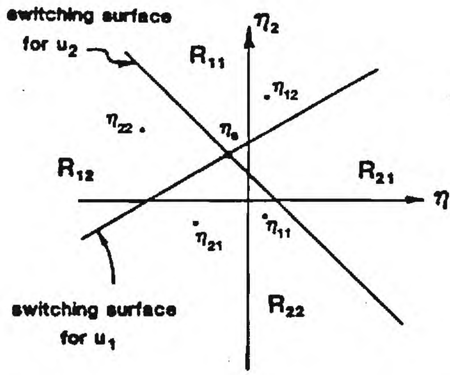

Figure 2: Graphical example of condition for equilibrium point, η_e, to lie on intersection of boundaries.

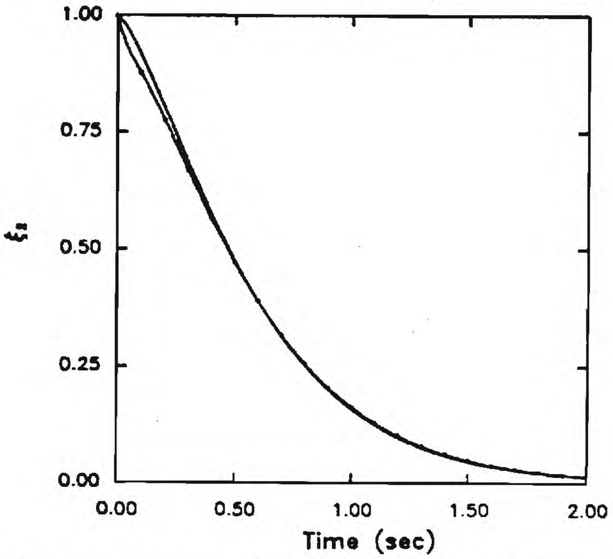

Figure 3: Response of ξ_2 to initial condition for actual system (solid line) and approximate system.

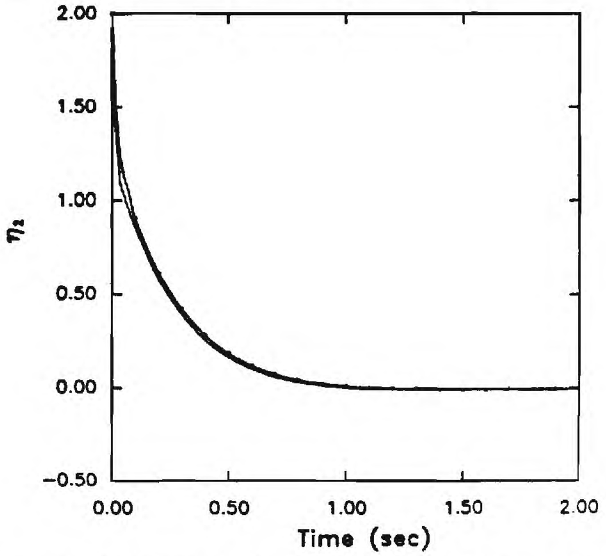

Figure 4: Response of η_2 to initial condition for actual system (solid line) and approximate system.

APPENDIX J

M. A. Ingram and A. H. Haddad, "Optimal and Suboptimal Filtering for Linear Systems Driven by Self-Excited Poisson Processes", Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, pp. 426-435, October 1987.

OPTIMAL AND SUBOPTIMAL FILTERING FOR LINEAR SYSTEMS DRIVEN BY SELF-EXCITED POISSON PROCESSES

MARY ANN INGRAM AND ABRAHAM H. HADDAD
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

ABSTRACT

Stochastic differential equations for the conditional density function and moments are presented for a linear system which is excited by a marked Poisson process whose rate depends on the state of the system and which is observed in white Gaussian noise. The set of optimal filtering equations is infinite dimensional, therefore, any practical filter is suboptimal. A suboptimal filter is developed for the case of unmarked Poisson excitation. This suboptimal filter estimates the Poisson process via a combined sequential estimation and detection scheme based on the criterion of maximum a posteriori (MAP) probability. An example computation is presented.

1. INTRODUCTION

This paper examines the issue of state estimation for a linear system which is driven by a Poisson process whose rate parameter depends on the state of the system. The input process is described as "self-excited" since its rate function can be specified given the past history of the input process.

The model of a dynamic system driven by a Poisson process with a state dependent rate is motivated by several practical situations. In aircraft maneuvers, the pilot's discrete application of controls is sometimes modeled as a Poisson input process. It is reasonable to expect that the rate of the control actions is dependent on the state of the aircraft. Another example is the tracking of a light source with a photon detector. The rate of photon arrivals certainly depends on the state of the tracking system, notably the tracking error angle.

The most general system considered in this paper is described by the following scalar equations:

$$dx_t = a_t x_t dt + b_t d\eta_t \tag{1}$$

$$y_t = \frac{dz_t}{dt} = c_t x_t + \frac{dw_t}{dt} \tag{2}$$

where η_t is a marked Poisson process whose marks (i.e., the amplitudes of the jumps) $\{u_i\}$ are a sequence of mutually independent, identically distributed random variables with density $p_u(u)$. The incident rate f η_t is a memoryless function of the state, $\mu(x_t)$. The process w_t is a rownian motion with diffusion V_t.

The objective is to estimate x_t given the history of the observation rocess, either y_s or z_s, for $s < t$. In Section 2, an expression for the inimium mean-squared error (MMSE) estimate is derived, and shown to be mpractical. Good suboptimal approximations to the MMSE estimate are esirable, but are not pursued here. Instead, in Section 3, the maximum a posteriori (MAP) criterion is used to derive a practical filter for x_t.

2. OPTIMAL FILTER EQUATIONS

This section derives the expression for the stochastic partial differential equation satisfied by $p_{t|t}(x)$, the conditional density function of x_t given $Z_t \overset{\Delta}{=} \{z_s; s < t\}$, based on a filtering theorem for white Gaussian observation noise. Furthermore, recursive equations are obtained for the central moments of this density function. The procedure used here is similar to the one used by Kwakernaak [1] to analyze a linear time invariant (LTI) system driven by an unmarked Poisson process with a constant rate.

First, the filtering theorem stated in Kwakernaak [1] is summarized for the special case of a scalar system with independent observation noise.

Filtering Theorem [1]: Let Q_t, $t > t_o$, be the semi-martingale defined by

$$dQ_t = R_t dt + dM_t \qquad t > t_o \tag{3}$$

where M_t is a martingale with respect to a growing family of σ-fields F_t, $t > t_o$, and where R_t is a process adapted to F. Let z_t, $t > t_o$, be the semi-martingale process

$$dz_t = h_t dt + dw_t \qquad t > t_o \tag{4}$$

where h is another process adapted to F, and w_t is a Brownian motion independent of F, such that $E(dw^2) = V_t dt$, $V_t > 0$ for $t > t_o$. Define Z_t as the growing family of σ-fields generated by the process z_t. For an arbitrary process ξ_t, define $\hat{\xi}_t \overset{\Delta}{=} E(\xi_t | Z_t)$. Then $\hat{Q}_t$ satisfies the dynamic equation

$$d\hat{Q}_t = \hat{R}_t dt + [\widehat{Q_t h_t} - \hat{Q}_t \hat{h}_t] V_t^{-1} [dz_t - \hat{h}_t dt] \tag{5}$$

The filtering theorem will be applied to $Q_t = e^{i\nu x_t}$, for x_t as defined in (1). However, the differential rule for filtered Poisson processes must first be used to obtain dQ_t. The rule may be found in Snyder [2, p. 200], and is also a special case of the differential rule for discontinuous semimartingales [1,3].

<u>Differential Rule [2]</u>: For an appropriately smooth function $Q(x_t)$ and for x_t defined in (1), the rule is

$$dQ(x_t) = a_t x_t \left(\frac{\partial Q(x_t)}{\partial x_t}\right)dt + \int_U \left[Q(x_t + b_t u) - Q(x_t)\right]K(dt,du) \qquad (6)$$

where the last integral is a counting integral [2, p. 195], evaluated over the mark space U, with respect to the Poisson counting measure $K(dt,du)$. $K(\Delta t,A)$ is the number of jumps of η_t during the interval Δt with marks in the set $A \subsetneq U$.

Equation (6) may be put in the form of (3) by letting

$$dM_t = \int_U \left[Q(x_t + b_t u) - Q(x_t)\right]\left[K(dt,du) - \mu(x_t)p_u(u)dtdu\right] \qquad (7)$$

and taking $R_t dt$ as the remainder. The substitution of R_t into (5) yields

$$d\widehat{e^{i\nu x_t}} = \Big(i\nu \widehat{a_t x_t e^{i\nu x_t}} + \widehat{e^{i\nu x_t}[e^{i\nu b_t u} - 1]\mu(x_t)}\Big)dt$$

$$+ \left[\widehat{c_t x_t e^{i\nu x_t}} - c\hat{x}_t \widehat{e^{j\nu x_t}}\right]V_t^{-1}\left[dz_t - c_t \hat{x}_t dt\right] . \qquad (8)$$

Let $\theta_t = b_t u$ (recall u is the mark variable) and $p_{\theta_t}(\cdot)$ be the probability density function for θ_t. If it is assumed that the conditional density function $p_{t|t}(x)$ exists, then taking the inverse Fourier transform of each term of (8) yields

$$dp_{t|t}(x) = Lp_{t|t}(x)dt + V_t^{-1}c_t(x-\hat{x}_t)p_{t|t}(x)\left[dz_t - c_t\hat{x}_t dt\right] \qquad (9)$$

where L is the linear operator given by

$$Lp(x) = -\frac{\partial}{\partial x}\left[a_t x p(x)\right] + \left(p_{\theta_t}(x) * \left[\mu(x)p(x)\right]\right) - \mu(x)p(x) \qquad (10)$$

where "*" denotes convolution. As in Kwakernaak's case, equation (9) is the same as the Kushner equation for systems driven by Brownian motion, except for the definition of L.

Equations (9) and (10) can be used to derive stochastic differential equations for $\hat{x}_t$ and the n^{th} conditional central moments $P_{n,t} = E[(x_t-\hat{x}_t)^n|Z_t]$ as follows:

$$P_{n,t} = E[(x_t-\hat{x}_t)^n|Z_t] \qquad n = 1,2,... \tag{11}$$

$$d\hat{x}_t = a_t\hat{x}_t dt + b_t E(u)\widehat{\mu(x_t)}dt + V_t^{-1}c_t P_{2,t}[dz_t - c_t\hat{x}_t dt]$$

$$\begin{aligned} dP_{n,t} = {} & na_t P_{n,t}dt + \sum_{k=1}^{n} \binom{n}{k} b_t^k E(u^k)\,\widehat{(x_t-\hat{x}_t)^{n-k}\mu(x_t)}dt - nb_t E(u)\widehat{\mu(x_t)}P_{n-1,t}dt \\ & + V_t^{-1}c_t[P_{n+1,t} - nP_{2,t}P_{n-1,t}][dz_t - c_t\hat{x}_t dt] \\ & + nV_t^{-1}c_t^2 P_{2,t}[\tfrac{n-1}{2}P_{2,t}P_{n-2,t} - P_{n,t}]dt \qquad n = 2,3,... \end{aligned} \tag{12}$$

Equations (11) and (12) represent an infinite set of coupled stochastic differential equations. Thus, an exact mean-squared error optimal filter is impossible to implement. Furthermore, in Kwakernaak's opinion, simple truncation of the moment equations (for the constant rate case) leads to unstable filters and generally poor results. Hence, approximate suboptimal filtering techniques are required, and are under investigation. This paper considers an alternative approach which uses a different error criterion, and is treated in the next section.

3. A MAP APPROACH

For this analysis, it is assumed that the driving process n_t is a counting process, i.e., it has only unit jumps. Furthermore, it is assumed that the system being driven is linear time-invariant, that is, $a_t = a$ and $b_t = b$ in equation (1). Thus, it is clear that knowledge of the jump times implies knowledge of x_t. The approach followed in this section is to obtain MAP estimates of the number N_T of jumps in n_t and the jump times $\underline{\tau}_{N_T} = [\tau_1,\tau_2,...,\tau_{N_T}]$ on the interval [0,T), given the observations Y_T =

$\{y_s; s < T\}$. The state estimate at time T, denoted $\tilde{x}_T$, is then constructed by the appropriate superposition of impulse responses. The approach is made into a practical sequential algorithm by using time discretization and a finite time window.

This is an extension of the work of Au and Haddad [3] wherein the approach outlined above was taken for marked Poisson driving processes which have constant known rates.

The MAP estimates $\tilde{N}_T$ and $\tilde{\underline{\tau}}_M$ satisfy

$$(\tilde{N}_T, \tilde{\underline{\tau}}_M) = \arg\left(\max_{\substack{0 < N^* < M \\ \underline{\tau}_M^* \epsilon R_+^M}} \left[\ln p_{N_T, \underline{\tau}_M | Y_T, N_T < M}(N^*, \underline{\tau}_M^*)\right]\right) \tag{13}$$

where the argument of the logarithm is a joint a posteriori probability density function. M is an integer chosen large enough so that $\Pr[N_T > M]$ is negligible. The condition $N_T < M$ ensures that $\underline{\tau}_M$ includes enough jump times to construct $\tilde{x}_T$.

The log of the density function in (13) can be replaced by the following expression without changing the result:

$$\ln \frac{1}{2V_T} \int_0^T \left[2y_t - \sum_{i=0}^{N^*} h(t, \tau_i^*)\right]\left[\sum_{j=0}^{N^*} h(t, \tau_j^*)\right]dt$$

$$+ \ln p_{\underline{\tau}_M | N_T, N_T < M}(\underline{\tau}_M^* | N_T = N^*, N_T < M) + \ln \Pr[N_T = N^* | N_T < M] \ . \tag{14}$$

The first term is recognized as the log likelihood function, wherein $h(t, \tau_j^*)$ represents the response of the system at time t to an impulse at time τ_j^*. For brevity, $h(t, \tau_o^*)$ is defined as the unforced response due to a known initial condition x_o.

The next objective is to simplify the expressions of the second and third terms of (14). Note that the event $N_T = N^*$ is also the event $\tau_{N^*} < T < \tau_{N^*+1}$. Therefore, the probability density in the second term can be rewritten as

$$p_{\underline{\tau}_M | N_T, N_T < M}(\underline{\tau}_M^* | N_T = N^*, N_T < M) =$$

$$\begin{cases} \dfrac{p_{\underline{\tau}_M | N_T < M}(\underline{\tau}_M^* | N_T < M)}{\Pr[\tau_{N^*} < T, \tau_{N^*+1} > T | N^* < M]} & \text{for } 0 < \tau_1^* < \dots < \tau_{N^*}^* < T \\ & \text{and } T < \tau_{N^*+1}^* < \tau_M^* \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Since $\ln 0 = -\infty$, it is reasonable to restrict the region over which the expression in (13) is maximized to the region of support of (15). Under this restriction, the third term of (14) cancels the denominator of the nonzero part of (15).

The remaining term to simplify is the numerator of the nonzero part of (15). It is noted that the event $N_T < M$ is also the event $\tau_{M+1} > T$. Thus, the term of interest may be expressed as a marginal density function:

$$p_{\underline{\tau}_M | N_T < M}\left(\underline{\tau}_M^* | N_T < M\right) = \int_T^\infty p_{\underline{\tau}_{M+1} | \tau_{M+1} > T}\left(\underline{\tau}_{M+1}^* | \tau_{M+1} > T\right) d\tau_{M+1}^* \, . \qquad (16)$$

It is noted that the region of support of the integrand is over the "wedge" $0 < \tau_1 < \ldots < \tau_M < \tau_{M+1}$ minus the half space $\tau_{M+1} < T$. Therefore, (16) can be rewritten as:

$$p_{\underline{\tau}_M | N_T < M}\left(\underline{\tau}_M^* | N_T < M\right) = \int_{\mathrm{Max}\left(\tau_M^*, T\right)}^\infty \frac{p_{\underline{\tau}_{M+1}}\left(\underline{\tau}_{M+1}^*\right)}{\Pr\left(\tau_{M+1} > T\right)} d\tau_{M+1}^* \qquad (17)$$

The unconditional density in the integrand of (17) is a special case of the density considered by Snyder [2, p. 248] for a self-exciting point process. For this special case, the density can be expressed as:

$$p_{\underline{\tau}_{M+1}}\left(\underline{\tau}_{M+1}^*\right) = \begin{cases} \prod_{i=1}^{M+1} \frac{\partial}{\partial \tau_i^*} \; -\exp \int_{\tau_{i-1}^*}^{\tau_i^*} -\mu\left[\overline{x}_t\left(\underline{\tau}_{i-1}^*\right)\right] dt & \text{for } 0 < \tau_1^* < \ldots < \tau_{M+1}^* \\ 0 & \text{otherwise} \end{cases} \qquad (18)$$

with

$$\overline{x}_t\left(\underline{\tau}_{i-1}^*\right) = x_o e^{at} u_1(t) + \sum_{j=1}^{i-1} b \exp\left[a\left(t - \tau_j^*\right)\right] u_1\left(t - \tau_j^*\right)$$

where u_1 is the unit step function. In words, $\overline{x}_t\left(\underline{\tau}_{i-1}^*\right)$ is the value of the state assuming that η_t has had jumps only at times $\tau_1^*, \ldots, \tau_{i-1}^*$. Let $\overline{x}_t\left(\underline{\tau}_o^*\right)$ denote the unforced value of the state.

Substitution of (18) into (17) is straightforward due to the product form of (18) and yields

$$p_{\underline{\tau}_M | N_T < M}\left(\underline{\tau}_M^* | N_T < M\right) = \frac{p_{\underline{\tau}_M}\left(\underline{\tau}_M^*\right)}{\Pr\left(\tau_M > T\right)} \exp \int_{\tau_M^*}^{\mathrm{Max}\left(\tau_M^*, T\right)} -\mu\left[\overline{x}_t\left(\underline{\tau}_M^*\right)\right] dt \qquad (19)$$

where it has been assumed that there exists some $\alpha > 0$ such that $\mu[x] > \alpha$ for all x, thus making

$$\exp \int_{\tau_M^*}^{\infty} - \mu[\overline{x}_t] dt = 0 \, .$$

It is noted that $p_{\underline{\tau}_M}(\cdot)$ is defined by replacing $M + 1$ by M in (18). Evaluation of the derivatives yields:

$$p_{\underline{\tau}_M}(\underline{\tau}_M^*) = \begin{cases} \prod_{i=1}^{M} \mu[\overline{x}_{\tau_i^*}(\underline{\tau}_{i-1}^*)] \exp \int_{\tau_{i-1}^*}^{\tau_i^*} - \mu[\overline{x}_t(\underline{\tau}_{i-1}^*)]dt & \text{for } 0 < \tau_1^* < \ldots < \tau_M^* \\ 0 & \text{otherwise} \end{cases} \qquad (20)$$

The combination of equations (14), (15), (19), and (20) results in a new MAP equation:

$$(\tilde{N}_T, \underline{\tilde{\tau}}_M) =$$

$$\arg \left\{ \underset{0 < N^* < M}{\text{Max}} \left[\underset{\substack{\underline{\tau}^* \epsilon R_+^M \\ \tau_1^* < \ldots < \tau_{N^*}^* < T \\ T < \tau_{N^*+1}^* < \ldots < \tau_M^*}}{\text{Max}} \frac{1}{2V_T} \int_0^T [2y_t - \overline{x}_t(\underline{\tau}_M^*)][\overline{x}_t(\underline{\tau}_M^*)]dt + \ell n \left(\prod_{i=1}^{M} \mu[\overline{x}_{\tau_i^*}(\underline{\tau}_{i-1}^*)] \right) - \int_0^{\text{Max}(\tau_M^*, T)} \mu[\overline{x}_t(\underline{\tau}_M^*)]dt \right] \right\} \qquad (21)$$

where the maximization is to be performed in two steps, first over the $\underline{\tau}_M^*$'s for fixed N*, and second over the N*'s.

4. SEQUENTIAL MAP APPROXIMATION

The MAP equation (21) derived in the previous section is now approximated as a sequential algorithm. In this approximation, the observations are processed in subintervals each of length Δ, which is chosen such that the probability of having two or more jumps in each interval is negligibly small. Each subinterval of observations is used to detect a jump in the subinterval and to estimate the jump time, as well as to update the estimates of past jump times.

In order to reduce computational complexity of the algorithm, estimates further than L subintervals away from the new subinterval are not updated and considered "finalized." The selection of L represents a tradeoff between performance and complexity. Thus, observations in the k^th subinterval $[(K-1)\Delta, K\Delta)$ are used to update estimates in the "window" $[(K-L)\Delta, K\Delta)$. $\tilde{N}_{(K-L)\Delta}$ represents the number of finalized estimates of jump times.

Equation (21) is next modified so that maximization is performed only over jump times occurring after the time $(K-L)\Delta$. Any additive terms which depend solely on finalized estimates are dropped. For brevity, let $\tilde{N}_F$ =

$\tilde{N}_{(K-L)\Delta}$ ("F" for finalized). Furthermore, redefine $\underline{\tau}_L$ as $[\tau_{\tilde{N}_F+1},\dots,\tau_{\tilde{N}_F+L}]$, and redefine $\bar{x}_t(\underline{\tau}^*_L)$ as the state assuming that jumps have occurred only at the finalized times and at the proposed times $\underline{\tau}^*_L$. The modified (approximate) version of (21) is:

$$(\tilde{N}_{K\Delta},\tilde{\underline{\tau}}_L) = \arg\left\{ \max_{\tilde{N}_F < N^* < \tilde{N}_F + L} \left[\max_{\substack{\tau_{\tilde{N}_F} < \tau^*_{\tilde{N}_F+1} < \dots < \tau^*_{N^*} < K\Delta \\ K\Delta < \tau^*_{N^*+1} < \dots < \tau^*_{\tilde{N}_F+L}}} \frac{1}{2V_T}\int_{(K-L)\Delta}^{K\Delta} [2y_t - \bar{x}(\underline{\tau}^*_L)][\bar{x}(\tau^*_L)]dt \right.\right.$$
$$\left.\left. + \ln\left(\int_{i=\tilde{N}_F+1}^{\tilde{N}_F+L} \mu[\bar{x}_{\tau^*_i}(\underline{\tau}^*_{i-1})]\right) - \int_{(K-L)\Delta}^{\mathrm{Max}(\tau^*_{\tilde{N}_F+L},K\Delta)} \mu[\bar{x}_t(\underline{\tau}^*_L)]dt \right]\right\} \tag{22}$$

There is a remaining difficulty with the maximization over the τ^*'s in (22). Assume that this maximization is being performed for a given, fixed N^*. Furthermore, assume a discretized domain, i.e., a subset of equally spaced discrete values in R^L. The discretization implies that the expression in (22) is evaluated over a finite number of values for the τ^*'s between $\tau_{\tilde{N}_F}$ and $K\Delta$, but there are still an infinite number of values to check for the τ^*'s above $K\Delta$. Maximizing over these "future" jump times is equivalent to maximizing the joint a priori probability density for these jump times.

The constant rate case $(\mu[x_\varepsilon] = \mu_o)$ presents no difficulty, because the joint a priori density function for the jump times after $K\Delta$ has its maximum at $\tau^*_{N^*+1} = \tau^*_{N^*+2} = \dots = \tau^*_{\tilde{N}_F+L} = K\Delta$. It is easily shown that the same is true for stable first order systems and rate functions $\mu[x]$ which monotonically increase with $|x|$. However, for more general LTI systems and rate functions, finding the maximum of the a priori joint density is apparently not as easy. This matter is currently under investigation.

5. EXAMPLE

Figures 1 and 2 display simulation results based on the algorithm of Section 4. The parameters are (see equations 1 and 2) $a_t = -5$, $b_t = 2$, and $c_t = 1$. The rate or intensity, $\mu(x_t)$, of the counting process η_t, takes only two values: $\mu(x_t) = 2$ for $|x_t| < 1$ and $\mu(x_t) = 4$ for $|x_t| > 1$. Figure 1 contains the state trajectory. The rate takes its high value when the trajectory is above the dashed line and the low value otherwise.

For estimation, $\Delta = 0.03125$ sec. This yields an approximate upper bound for $\Pr[n_{t+\Delta} - n_t > 1]$ of $4\Delta = 0.125$. The observation noise samples have a standard deviation $(\sqrt{V_t})$ of 0.15. The estimation/detection window is $L = 4$. Estimation results are shown in figure 2. Some errors may be observed at $t \cong 2$ and $3 < t < 4$. It is noted that for $\sqrt{V_T} = 0.1$, all of the jumps were correctly detected (to the order of the simulation sample period) and for $\sqrt{V_T} = 0.2$, several more false detections occurred in the region $0.5 < t < 1.5$.

6. CONCLUSIONS

The state estimation problem has been considered for a linear system observed in additive white Gaussian noise, where the system is driven by a Poisson process with a state dependent rate. It is no surprise that the minimum mean-squared estimator is infinite dimensional, since the same is true for the simpler constant rate case. However, it is expected that the form of the equations will suggest a good suboptimal approximation in the future. An implementable estimator was developed based on maximum a posteriori (MAP) estimates of the number and times of the jumps in the driving process. However, the feasibility of this scheme has been shown only for certain LTI systems and rate functions. Further investigation is needed to enlarge the apparently limited applicability of this MAP approach.

ACKNOWLEDGEMENTS

This research is supported by the U.S. Air Force under Contracts F08635-84-C-0273 with the Armament Laboratory and AFOSR-87-0308.

REFERENCES

[1] H. Kwakernaak, "Filtering for systems excited by Poisson white noise," Eds.: A. Bensoussan, J. L. Lions, Control Theory, Numerical Methods, and Computer Systems, Springer Lecture Notes in Economics and Mathematical Systems, vol. 107, Berlin, pp. 468-492, 1975.

[2] D. L. Snyder, Random Point Processes, Wiley, New York, 1975.

[3] S. P. Au and A. H. Haddad, "Suboptimal sequential estimation-detection scheme for Poisson driven linear systems," Inform. Sci., vol. 16, pp. 95-113, 1978.

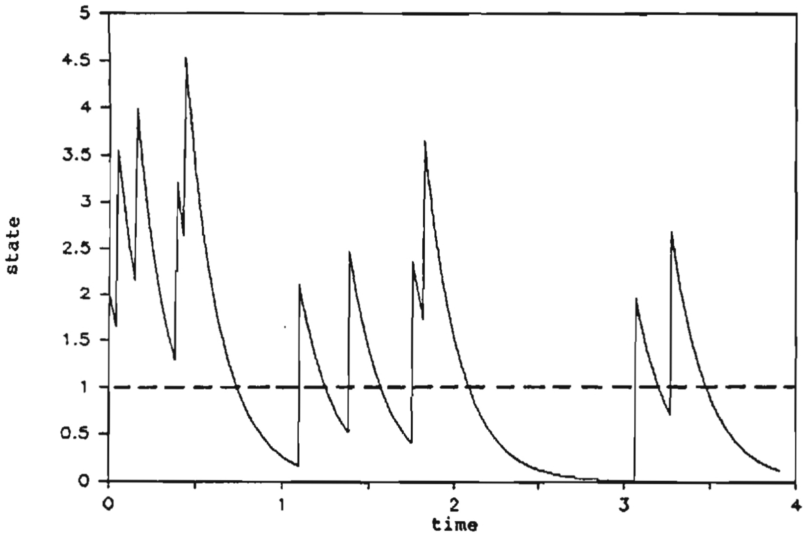

Figure 1. Example state trajectory

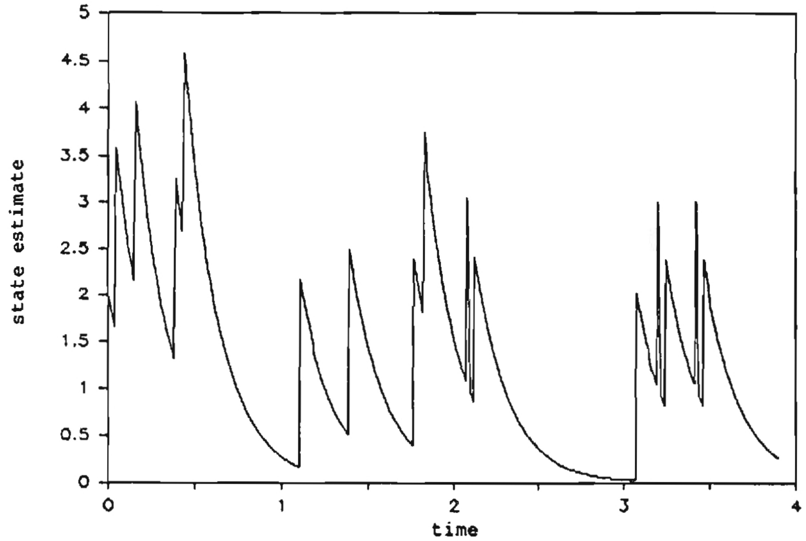

Figure 2. Estimate output

APPENDIX K

M. A. Ingram and A. H. Haddad, "A Sequential Detection Approach to State Estimation of Linear Systems Driven by Self-Excited Point Processes", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 2334-2335, Dec. 1988.

on Decision and Control
Austin, Texas • December 1988

A Sequential Detection Approach to State Estimation of Linear Systems Driven by Self-Excited Point Processes*

M. A. Ingram

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250

A. H. Haddad

Department of EE/CS
Northwestern University
Evanston, Illinois 60208

Abstract

A sequential detection scheme is used to determine the approximate occurrence times of impulses in a self-excited point process which drives a scalar linear system. Observations of the state are corrupted by additive white Gaussian noise. The state estimate is constructed based on the detected impulses.

1 Introduction

Linear systems driven by a combination of a marked (randomly weighted) impulse process and a white Gaussian noise process have been used as models for maneuvering targets [1], switching environments [2], and seismic signals in oil exploration [3]. Impulsive input processes with state-dependent statistics are applicable if a system is prone to a high disturbance rate in some regions of the state space and a low rate in other regions. As preparation for the analysis of the complex model described above, a simpler problem has been addressed in which the only disturbance is a self-excited point process [9] with constant marks. The process is described as self-excited because its instantaneous average rate is a function of the state of the system being driven.

The optimal mean square error filter involves an infinite set of coupled stochastic differential equations [4,6]. In cases where the instantaneous average rate of the input process is constant and is low relative to the bandwidth of the system, various truncations of the optimal filter have performed poorly [4,8].

These performance reports have prompted investigation into other approaches [8,6,5] which resemble a maximum a posteriori (MAP) approach. These approaches share the basic goal of determining the number, N, and the arrival times, $\underline{\tau}_N$, of input pulses in a time interval, using observations over that time interval and a priori statistics. Au [8] and Kwakernaak [5] make the additional assumption of random marks. However, it will be demonstrated that an inherent difficulty with this problem is preserved in the constant mark case. In all approaches but Kwakernaak's, the solutions to this fixed-interval smoothing problem were transformed into fixed-lag smoothing algorithms by allowing the interval to become a moving window. A smoothed estimate of the state is constructed by superimposing responses to the detected impulses as they are left behind by the time window.

In order to discuss the problem further, some definitions are needed. Let $N_{0,T}$ be the number of input impulses in the interval $[0,T)$. Let $\underline{\tau}_n = [\tau_1, \tau_2, \ldots, \tau_n]$ represent the first n consecutive arrival times of the impulses. Assume observations of a scalar system:

$$y_t = x_t(\underline{\tau}_{N_{0,T}}) + w_t \quad t \in [0,T) \tag{1}$$

$$x_t(\underline{\tau}_{N_{0,T}}) = \sum_{i=1}^{N_{0,T}} h(t-\tau_i) + \Phi(t,0)x_0 \tag{2}$$

where w_t is Gaussian white noise with spectral height σ^2 , $h(t)$ is the impulse response of the system, and $\Phi(t,0)x_0$ is the unforced response. Let $\{\mathcal{Y}_{0,T}\}$ represent the observations over the interval $[0,T)$. The instantaneous average rate of input impulses is defined to be $\mu[x_t]$, where μ is a positive, bounded function.

The difficulty referred to above is how to properly use the a priori statistics in this problem. The procedure is straightforward when the goal is to produce a MAP estimate of the consecutive arrival times in an interval $[0,T)$, given that there are exactly n arrival times. The MAP estimate $\hat{\underline{\tau}}_n$ maximizes the quantity

$$\Lambda\{\mathcal{Y}_{0,T} \mid \underline{\tau}_n\} p(\underline{\tau}_n \mid N_{0,T} = n) \tag{3}$$

where Λ is the likelihood functional

$$\exp\left(\frac{1}{2\sigma^2}\int_0^T [2y_t - x_t(\underline{\tau}_n)][x_t(\underline{\tau}_n)]dt\right) \tag{4}$$

and $p(\underline{\tau}_n \mid N_{0,T} = n)$ is the joint probability density function (pdf) of the first n occurence times in $[0,T)$, given that $N_{0,T} = n$. It is noted that for the state-dependent rate case, this pdf is not generally differentiable with respect to $\underline{\tau}_n$.

It is also straightforward to produce a MAP estimate of $N_{0,T}$, that is, a minimum probability of error detection of $N_{0,T}$. The MAP estimate $\hat{N}_{0,T}$ maximizes

$$E\{\Lambda\{\mathcal{Y}_{0,T} \mid \underline{\tau}_n\} \mid N_{0,T} = n\} \Pr\{N_{0,T} = n\}. \tag{5}$$

where, in the constant rate case, the second factor is a unimodal function of n, with its peak at $E\{N_{0,T}\} = \lambda T$.

In the problem at hand, however, neither $\hat{\underline{\tau}}_n$ or $\hat{N}_{0,T}$ alone will suffice. The two procedures must somehow be merged. The issue is more pressing when the model includes random marks which can take very small values. In that case, the maximum likelihood approach yields unreasonably large values of N as many small impulse responses are made to fit the observation noise. Any estimator of N must sufficiently penalize large values.

In each of the three approaches mentioned above, a single expression is maximized to determine both the number and times (also marks in [8] and [5]) of impulses in an interval. Au [8] used the likelihood ratio weighted by Snyder's sample function density (sfd) [9]:

$$f(\underline{\tau}_n, \underline{u}_n; n) = p(\underline{\tau}_n \mid N_{0,T} = n) p(\underline{u}_n \mid N_{0,T} = n) \Pr\{N_{0,T} = n\} \tag{6}$$

where $\underline{u}_n$ denotes the random marks. It is noted that the sfd is not a joint pdf, since the dimension of its domain depends on the last argument. For independently and identically distributed (iid) marks, the natural log of the sfd simplifies to

$$n \ln \lambda - \lambda T + \sum_{i=1}^{n} \ln p(u_i). \tag{7}$$

which, as a function of n, does not share the characteristics of $\ln \Pr\{N_{0,T} = n\}$; in particular, it does not necessarily penalize large n. In the algorithm, extreme values of n are prevented by limiting the rate of change in the collection of n's.

In treating the state-dependent rate case, Ingram and Haddad [6] replaced the sfd in Au's approach with the joint pdf of $\underline{\tau}_M$ and $N_{0,T}$, where M is chosen so that $\Pr\{N_{0,T} > M\} \ll 1$. Use of an actual pdf might seem appropriate for a MAP approach. However, it was observed that for the constant rate case, this pdf is constant with respect to both $\underline{\tau}_M$ and n when $n < M$. Thus for the constant rate case, the criterion is simply maximum likelihood with an upper bound on n.

*Supported by the U. S. Air Force under Grant AFOSR–87–0308

Kwakernaak [5] applied Rissanen's [7] shortest data description method to this problem. The resulting procedure is the same as Au's except that the expression in (6) is augmented with the factors $\Pi_{i=1}^{n} \eta_i \delta_i$ where η_i and δ_i are resolutions for digitizing the ith mark and arrival time, respectively. The resolutions are chosen to minimize the augmented expression, which is interpreted as the symbol length needed to encode the data n, $\underline{t}_n$, and $\underline{u}_n$. This method is optimal for the data length criterion and penalizes high values of n. However, it is not readily applicable when the input is a self-excited point process. This is due to the complexity of the expressions and the required differentiability of the sfd in the assignments of η_i and δ_i.

2 The Sequential Algorithm

The approach taken in this paper to the problem of estimating the self-excited input point process has two steps. The first step is to compute the MAP estimates $\hat{\underline{t}}_n$ for every n such that $0 < n \leq M$, where M is chosen as above. The likelihood functional for the $n = 0$ case is also computed.

The second step is to approximate the minimum probability of error detection of $N_{0,T}$ by replacing the averaged likelihood functional in (5) with the likelihood functional evaluated at $\hat{\underline{t}}_n$, i.e., by maximizing

$$\Lambda\{\underline{y}_{0,T} \mid \hat{\underline{t}}_n, x_0\} \Pr\{N_{0,T} = n \mid x_0\} \qquad (8)$$

over $0 \leq n \leq M$ to get $\hat{N}_{0,T}$. The final estimate of input impulse times is $\hat{\underline{t}}_{\hat{N}_{0,T}}$. Both terms in (8) are conditioned on x_0 because of the state dependence of the rate.

The main reason this approach was chosen is its explicit dependence in both steps on a priori statistics. This dependence was highly desirable, given the rather elaborate input model that has been assumed. The reason why equation (5) was not used is mainly due to implementation difficulties. Specifically, the likelihood functional can take on very large values for high signal-to-noise; this causes high sensitivity to approximation errors in the numerical integration needed to perform the expectation. Even if this sensitivity problem did not exist, the multiple integration would not be desirable because it is very time consuming.

Equation (8) is relatively easy to implement. The first factor is a byproduct of the first step and the second factor, $\Pr\{N_{0,T} = n \mid x_0\}$, can be computed off-line for the desired range of values for x_0. The second factor naturally imposes a penalty on high n and also noticeably depends on x_0.

In the sequential algorithm, this procedure is performed over the interval $[A, A+T]$, where A is periodically incremented. The smoothed state estimate $\hat{x}_A$ takes the place of x_0.

The algorithm has been tested on the following example. Let

$$\begin{aligned} h(t) &= 2e^{-5t}u(t) \\ \mu[x_t] &= \begin{cases} 1 & |x_t| \leq 1 \\ 4 & |x_t| > 1 \end{cases} \end{aligned}$$

Some values of $\Pr\{N_{0,T} = n \mid x_0\}$, computed for the example and $T = 0.1875$ are shown in the table. For this interval size, $\Pr\{N_{0,T} > 3\} \leq 0.0073$, so it is sufficient to consider values of n up to $M = 3$. It is observed that as x_0 increases, the probability weight gradually shifts away from $n = 0$, but the probabilities of $n = 1$ and $n = 3$ still differ by an order of magnitude. For $x_0 \geq 2.6$, the distribution for $0 \leq n \leq 3$ is unchanged because the high initial condition ensures that $\mu[x_t] = 4$ over the whole interval. For $\sigma^2 = 0.01$, all pulses except the first are detected within the time resolution of the simulation. As the noise strength grows from $\sigma^2 = 0.04$, errors begin to occur. Additional tests are being performed and a theoretical performance analysis is under investigation.

x_0	$n=0$	$n=1$	$n=2$	$n=3$
0.0	0.829	0.121	0.039	0.010
0.2	0.829	0.120	0.040	0.010
0.4	0.829	0.119	0.040	0.010
0.6	0.829	0.119	0.040	0.010
0.8	0.829	0.119	0.040	0.010
1.0	0.829	0.119	0.040	0.010
1.2	0.744	0.159	0.075	0.022
1.4	0.679	0.195	0.094	0.027
1.6	0.623	0.231	0.109	0.030
1.8	0.581	0.259	0.119	0.032
2.0	0.547	0.287	0.126	0.033
2.2	0.517	0.309	0.130	0.033
2.4	0.491	0.336	0.133	0.033
2.6	0.472	0.354	0.133	0.033

Table 1: Values of $\Pr\{N_{0,T} = n \mid x_0\}$ for the example system.

References

[1] R. L. Moose, "An adaptive state estimation solution to the maneuvering target problem," *IEEE Trans. Automatic Control*, Vol. AC-20, No. 3, June 1975.

[2] G. A. Ackerson and K. S. Fu, "On state estimation in switching environments," *IEEE Trans. Automatic Control*, Vol. AC-15, No. 1, February 1970.

[3] J. M. Mendel, "White noise estimators for seismic data processing in oil exploration," *IEEE Trans. Automatic Control*, Vol. AC-22, No. 5, October 1977.

[4] H. Kwakernaak, "Filtering for systems excited by Poisson white noise," Eds.: A. Bensoussan, J. L. Lions, *Control Theory, Numerical Methods, and Computer Systems*, Springer Lecture Notes in Economics and Mathematical Systems, Vol. 107, Berlin, pp. 468–492, 1975.

[5] H. Kwakernaak, "Estimation of pulse heights and arrival times," *Automatica*, Vol. 16, pp. 367–377, 1980.

[6] M. A. Ingram and A. H. Haddad, "Optimal and suboptimal filtering for linear systems driven by self-excited Poisson processes," *Proc. Twenty-Fifth Annual Allerton Conference on Communication, Control, and Computing*, Sept. 30– Oct. 2, 1987, Monticello, Illinois, pp. 426–435.

[7] J. Rissanen, "Modeling by shortest data description," *Automatica*, Vol. 14, pp. 465–471, 1978.

[8] S. P. Au, "State estimation for linear systems driven simultaneously by Wiener and Poisson processes," Ph.D. Thesis, University of Illinios at Urbana-Champaign, 1979.

[9] D. L. Snyder, *Random Point Processes*, Wiley, New York, 1975.

APPENDIX L

M. A. Ingram and A. H. Haddad, "On Linear Systems Driven by Self-Excited Point Processes", Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, pp. 937-938, October 1988.

On Linear Systems Driven by Self-Excited Point Processes*

Mary Ann Ingram
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250

Abraham H. Haddad
Department of EE/CS
Northwestern University
Evanston, Illinois 60208

ABSTRACT

In this work, mean-square continuity is proved for the state of a linear system disturbed by a point process with a state-dependent rate and random marks. The cross-correlation property and the linear optimal filter are derived for the case of zero mean marks.

SUMMARY

The model treated in this summary is a continuous linear time invariant system driven by a self-excited, marked point process. The term "self-excited" implies that the instantaneous average jump rate or intensity of the point process depends on the history of the process. Thus, self-excitation is one kind of time-correlation. In particular, the jump rate is specified as a memoryless function of the system state. The term "marked" describes a point process with random jump amplitudes (marks).

One possible application of this model is in the tracking of maneuvering targets. The jump process represents the commanded acceleration of the vehicle being tracked. The state-dependency of the rate represents a relation between the rate of acceleration jumps and the position and velocity of the vehicle. Another possible application is in state estimation for systems subject to abrupt failures, such as the onset of biases in sensors or actuators. Here, state-dependency of the rate may represent an increased vulnerability to failures under conditions of high heat, speed, or electrical current.

The state process is given by the following stochastic differential equation

$$dx_t = Ax_t dt + BdM_t \quad t \geq 0$$

where $x_t \in \Re^n$ is the state, with initial condition x_0, $A \in \Re^n \times \Re^n$, $B \in \Re^n$, and $M_t \in \Re$ is a piecewise constant random process to be defined below.

Let N_t denote the number of jumps in M_t. Let $\{u_1, u_2, \ldots, u_{N_t}\}$ be the consecutive jump heights or marks of M_t. Then M_t may be expressed

$$M_t = \sum_{i=1}^{N_t} u_i.$$

The marks are assumed to be independently and identically distributed with mean $\overline{u}$ and mean square value $\overline{u^2}$. We define the stochastic intensity [1] or instantaneous average rate of N_t to be $\mu[x_t]$, where μ is a scalar valued, positive, and bounded function of the state x_t of the system. It follows that $N_t - \int_0^t \mu[x_s]ds$ is a martingale with respect to the σ-algebra S_t generated by x_t. The observation model is given by

$$dz_t = x_t dt + dw_t$$

where w_t is an n-vector of Wiener processes with $E\{dw_t dw_t\} = Idt$.

There has been a fair amount of work concerning systems driven by compound Poisson processes, that is, independent increment point processes with random marks. The contribu-

*Supported by the U. S. Air Force under Grant AFOSR-87-0308

tions include representations and properties [2], and mean-square optimal, linear optimal, and suboptimal state estimation [3,4]. Self-excited and more general point processes have received attention [2,1], mainly as models for point process observations of dynamic systems.

Martingale theory has been applied successfully in the characterization of point processes [1] as well as in nonlinear filtering theory. Thus, it was desirable and instructive to use it in proving the following propositions.

Proposition 1 *If there exists a constant K such that*

$$\mu[x] < K < +\infty$$

for all $x \in \Re^n$, then x_t is mean-square continuous, that is,

$$\lim_{s \to t} E\{\|x_t - x_s\|^2\} = 0.$$

where $\| \cdot \|$ denotes the Euclidean distance in $\Re^n$.

Proposition 2 *If the mean of the marks is zero, i.e., if $\overline{u} = 0$, then x_t is of the separable class* [5], *which implies that for any nonlinear, scalar valued function $g(\cdot)$, there exists a constant vector C such that*

$$E\{x_t g(x_\tau)\} = E\{x_t x'_\tau\}C.$$

given that the appropriate expectations exist.

The innovations approach of Kailath [6] may be used to derive the linear filter. The mean-square continuity of x_t is sufficient to prove that the innovations process $v_t = z_t - \int_{t_0}^t \hat{x}_s \, ds$, where $\hat{x}_t$ is the filter output, is a process of orthogonal increments. When the marks have zero mean, the resulting filter is

$$\begin{aligned} d\hat{x}_t &= A\hat{x}_t dt + P_t dv_t \\ \dot{P}_t &= AP_t + P_t A' + BB' \, \overline{u^2} \, \overline{\mu[x_t]} - P_t P_t \end{aligned}$$

Although P_t will be difficult to compute due to the a priori expectation $\overline{\mu[x_t]}$, the computation can be done off-line. We are currently investigating methods of computation and the filter expression for the case of nonzero mean marks.

References

[1] P. Brémaud, *Point Processes and Queues, Martingale Dynamics*, Springer-Verlag, New York, 1981.

[2] D. L. Snyder, *Random Point Processes*, Wiley, New York, 1975.

[3] S. P. Au and A. H. Haddad, "Suboptimal sequential estimation-detection scheme for Poisson driven linear systems," *Inform. Sci.*, Vol. 16, pp. 95-113, 1978.

[4] H. Kwakernaak, "Filtering for systems excited by Poisson white noise," Eds.: A. Bensoussan, J. L. Lions, *Control Theory, Numerical Methods, and Computer Systems*, Springer Lecture Notes in Economics and Mathematical Systems, Vol. 107, Berlin, pp. 468–492, 1975.

[5] A. H. Haddad, "Dynamical representation of markov processes of the separable class," *IEEE Trans. Information Theory*, Vol. IT-16, No. 5, September 1970.

[6] T. Kailath, "A note on least squares estimation by the innovations method," *SIAM J. Contr.*, Vol. 10, No. 3, pp. 477-486, August 1972.

APPENDIX M

M. A. Ingram and A. H. Haddad, "A Linear System Driven by a Jump Process with a State-Dependent Rate: Properties and Linea Estimators", submitted for publications.

A Linear System Driven by a Jump Process with a State-Dependent Rate: Properties and Linear Estimators[1]

Mary Ann Ingram
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250

Abraham H. Haddad
Department of EE/CS
Northwestern University
Evanston, Illinois 60208

ABSTRACT

Modeling issues and the minimum mean squared error linear filter and smoother are studied for a linear system disturbed by a jump process with a state-dependent rate and random jump heights. The jump process is defined in terms of martingale processes. Martingale techniques are used to derive certain properties and second order statistics of the jump and state processes. It is shown that the linear filter and smoother are practical only for the case of zero-mean jump heights.

1 Introduction

Linear systems with random impulsive forcing functions have been used to model dynamic systems subject to abrupt failures or bias changes [1] and manuevering targets [2], as well as many other physical situations [3, chapt. 4]. State estimation for such systems from noisy observations has been an active research area for many years. The input process is often described as either an independent increment compound Poisson process [4,5,6,7] or a discrete-time semi-Markov process [2,8]. In both cases, the minimum mean squared error (MMSE) filter is not implementable. Thus researchers have considered various approximations to the MMSE filter as well as linear

[1] Supported by the U. S. Air Force under Grant AFOSR–87–0308

optimal filters and schemes which involve maximum a posteriori (MAP) or MAP-like criteria [6,8].

This note treats a linear system driven by an extended version of the compound Poisson process. The extension results from allowing the instantaneous average rate of the input impulses to depend on the state of the linear system. In the abrupt failure application, the state dependency is motivated by the idea that a system maybe more prone to failures when it is under some degree of "stress," as defined by a region of the state space. In the target tracking application, the likelihood of a pilot to give an acceleration command may depend on his speed and position, again reflected by the state of the system. Apparently, this input model has not been previously considered in the context of state estimation. A related model, a linear system with Markov jump parameters where the jump rate is state-dependent, has been considered for optimal control [9].

The objectives of this note are to determine various properties and quantities of the process of interest which are relevant to MMSE estimation, and to derive the MMSE linear filter and fixed-lag smoother. The note is summarized as follows. Section 2 contains a methodical development of the properties of the state process. The development is based on a semimartingale representation of the counting process which underlies the jump process. It is this semimartingale representation which precisely describes the state-dependency of the system disturbance. We build up from the counting process to the jump process, and finally to several representations of the state process. The state process is proved to be square integrable and mean square continuous. Section 3 contains a discussion on linear estimators for the state process, given observations in additive white Gaussian noise. A recursive form for the filter and fixed-lag smoother follow easily when the jump heights have zero mean. In the general case, however, a recursive form of the linear filter is not obtained. Some observations are made from the general filter expression and the form of the a priori covariance equations.

2 The State Process

In this section we define the state process and discuss its properties. Several representations are considered, and the properties of square integrability and mean square continuity are proved. We recall that these properties are true for the constant rate case, so it should not be surprising that they follow for the state-dependent rate case when the rate function is uniformly bounded. The definitions and proofs, however, do require some care.

The state process is given by the following stochastic differential equation

$$dx_t = Ax_t dt + BdM_t \quad t \geq 0 \tag{1}$$

where x_t is an n-dimensional state vector, with initial condition x_0. The $n \times n$ constant matrix A is assumed to be such that the solution of $\dot{x} = Ax$ is exponentially stable. The scalar process M_t is a piecewise constant random process. The jumps of M_t occur at an instantaneous average rate which depends on x_t.

2.1 The Input Process

The process M_t is known as a jump process. Its definition depends on a sequence of ordered pairs $\{(\tau_1, u_1), (\tau_2, u_2), ...\}$, where $\tau_i > 0$ is the time of the i^{th} jump and u_i is the jump height. This sequence of ordered pairs is known as a point process, and the u's are the marks of the point process. The τ sequence may be equivalently represented by the counting process N_t, which is the number of jumps prior to time t. Thus the jump process M_t may be expressed

$$M_t = \sum_{i=1}^{N_t} u_i. \tag{2}$$

In this note, the marks are assumed to be independent and identically distributed (iid) with probability density function (pdf) $p_u(u)$, mean $\overline{u}$, and mean square value $\overline{u^2}$. Also, u_j is independent of $\{N_s, M_s; 0 < s \leq t\}$ for $j > N_t$.

The state dependency of the rate is made precise in the definition of the counting process N_t. Because martingale theory is known to be quite powerful in the analysis of point processes on the real line [10,11] and it is fundamental to nonlinear filtering theory [12], N_t will be defined in terms of martingales.

Let $\mathcal{S}_t$ be the smallest σ-algebra containing the histories of both the state x_t and the counting process N_t, i.e. $\mathcal{S}_t = \sigma\{x_s, N_s; s \leq t\}$. The reason N_t is explicit in the definition of $\mathcal{S}_t$ is because if a mark u can take the value of zero, then a jump in N_t may not coincide with a jump in x_t. $\mathcal{S}_t$ is assumed to possess the "usual" properties of completeness and right continuity [13]. Consider the process $\lambda_t = \mu[x_t]$, $\mu : \mathcal{R}^n \rightarrow \mathcal{R}_+$, where μ is such that $0 < \mu[x] \leq K < +\infty$ for all $x \in \mathcal{R}^n$ and some constant K. We specify λ_t to be a stochastic intensity with respect to $\mathcal{S}_t$, or simply an $\mathcal{S}_t$-intensity, for N_t [11, p. 27]. Note that the definition of $\mathcal{S}_t$-stochastic intensity requires that N_t be measurable with respect to $\mathcal{S}_t$, hence the need for N_t in the definition of $\mathcal{S}_t$. An informal interpretation of λ_t is that on the infinitesimal interval $[t, t + dt)$, N_t acts like a Poisson process with rate parameter λ_t, or that $\Pr\{dN_t = 1 \mid \mathcal{S}_t\} = \lambda_t dt$.

By definition of $\mathcal{S}_t$-stochastic intensity,

$$D_t = N_t - \int_0^t \lambda_s ds \tag{3}$$

is an $\mathcal{S}_t$-martingale. The integral $\int_0^t \lambda_s ds$ is also known as the unique predictable compensator for N_t with respect to $\mathcal{S}_t$ [10, p. 59].

It is known that a counting process N_t may have more than one stochastic intensity with respect to a given growing σ-algebra, but there is only one stochastic intensity which is predictable [11, p. 30]. If $\mu[x_t]$ has left and right limits, then $\tilde{\lambda}_t = \mu[x_{t-}]$ is predictable [14, p. 46]. However, predictability is not required for the results in this paper; any λ_t differing from $\tilde{\lambda}_t$ on a set of Lebesgue measure zero may be used.

The description of a point process using a stochastic intensity is a relatively modern approach to the modeling of point processes. One of the classic approaches is to define a point process by the joint probability distribution of its jump times and marks and by the distribution of N_t. In his book [3], Snyder reviews distributional descriptions for many classes of point processes on the real line. One class of processes, the class of "marked self-excited point processes," includes the process of interest in this note [3, p.467]. The term self-excited means that the present and future statistics of the process depend on its own past. It is noted that the specification of $\lambda_t = \mu[x_t]$ as a stochastic intensity of N_t is consistent with the distributional characterization of a marked self-excited point process where the self-excitation is through state dependency. The connection is made through a theorem stated by Brémaud [11, p. 61]. We note in passing that the distributional description is useful in certain maximum a posteriori (MAP) approaches to this state estimation problem [15,16].

The definition of N_t will now be generalized to include random jump heights or marks. This procedure will lead to a decomposition of M_t similar to the decomposition in equation (3). The decompositon of M_t will, in turn, lead to a useful decomposition of x_t.

Let the mark sequence $\{u_n, n \geq 1\}$ take its values in the measure space $(U, \mathcal{U})$. The idea of the counting process N_t may be generalized to a counting measure $p((0,t] \times A)$, which is the number of jumps in M_t that have marks (or jump heights) in the set $A \in \mathcal{U}$ [11, p. 234]. It follows that $N_t = p((0,t] \times U)$ and

$$M_t = \int_U up((0,t] \times du). \tag{4}$$

If N_t has the stochastic intensity λ_t and the future marks are iid and independent of S_t, then it follows that the stochastic intensity of $p((0,t] \times A)$, is $\lambda_t \Pr\{A\}$. A heuristic argument is given below.

$$\begin{aligned}
\Pr\{p(dt \times A) = 1 \mid S_t\} &= \Pr\{dM_t \in A \mid dN_t = 1, S_t\} \Pr\{dN_t = 1 \mid S_t\} \\
&= \Pr\{u_{N_t+1} \in A \mid dN_t = 1\}\lambda_t dt \\
&= \int_A p_u(u) du \lambda_t dt.
\end{aligned}$$

A corollary of Brémaud [11, p. 235] then implies that

$$\begin{aligned}
R_t &= \int_0^t \int_U u\,[p(ds \times du) - \lambda_s p_u(u) ds du] \\
&= M_t - \int_0^t \bar{u} \lambda_s ds
\end{aligned} \tag{5}$$

is an S_t-martingale and $Q_t = \int_0^t \bar{u}\lambda_s ds$ is the unique predictable compensator of M_t. It is noted that (5) is also the unique decomposition with respect to $\tilde{S}_t = \sigma\{x_s; 0 \leq s \leq t\}$. This is because dM_t is conditionally independent of $\{N_s; 0 \leq s \leq t\}$ given x_t. It is further noted that R_t is an orthogonal increment process (its formal derivative is white noise) since all martingales have orthogonal increments.

In the appendix, R_t is shown to be an L^2-martingale with quadratic variance

$$\langle R, R \rangle_t = \int_0^t \overline{u^2} \mu[x_s] ds \tag{6}$$

which implies that $R_t^2 - \langle R, R \rangle_t$ is an $\mathcal{S}_t$-martingale [17, p. 115]. It follows that

$$\begin{aligned} E\{R_t R_\tau\} &= E\{\langle R, R \rangle_{t \wedge \tau}\} \\ &= \int_0^{t \wedge \tau} \overline{u^2}\, \overline{\mu[x_s]} ds. \end{aligned} \tag{7}$$

where $t \wedge \tau \equiv \min(t, \tau)$. The expressions in (6) and (7) will be employed below in a bound on state variance and in the linear filter expression, respectively.

2.2 Representations and Properties

The state process has several representations. These representations will be reviewed and then one will be used to prove the square integrability and mean square continuity of x_t.

The expressions in (1), (2), and (3) constitute one representation for x_t. It is also possible to give an augmented state equation with an independent-increment excitation. Let $\tilde{M}_t$ be a compound Poisson process with iid $p_u(u)$-distributed marks and a *unity* jump rate. It follows from a theorem and lemma of Brémaud [11, p. 41], that if λ_t is uniformly lower bounded away from zero, then x_t may also be represented by the following equations:

$$\begin{aligned} dx_t &= Ax_t + Bd\tilde{M}_{r_t} \\ dr_t &= \mu[x_t]dt \end{aligned} \tag{8}$$

where the augmented variable performs time scaling of the input process $\tilde{M}_t$. While the representation in (8) has not been useful in analysis, it is useful for computer simulations.

It is easily observed that x_t is a Markov process with a stationary transition function, since the statistics of dM_t in equation (1) are completely determined by x_t. More specifically, x_t is in the class of Piecewise-deterministic Markov processes, a class described by Davis [18] that "covers virtually all non-diffusion applications."

When a Markov process has a stationary transition function and is continous in probability (which is implied by mean square continuity, proved below), then its transition function is uniquely determined by its differential generator $\mathcal{A}$ [19, p. 184]. The operator $\mathcal{A}$ can be used to derive differential equations which propagate the covariance of x_t as well as other expected values associated with x_t. Given a continuously differentiable function f, an expression for $\mathcal{A}f(x_t)$ may be found by simplifying the general formula in Davis [18]:

$$\mathcal{A}f(x_t) = \frac{\partial f}{\partial x}Ax_t + \mu[x_t]\left[\int_U f(x_t + Bu)p_u(u)\, du - f(x_t)\right]. \tag{9}$$

A property of the generator is that

$$f(x_t) - f(x_s) - \int_s^t \mathcal{A}f(x_\tau)\, d\tau \tag{10}$$

is a martingale for $t \geq s$ [18].

The adjoint operator $\mathcal{A}^*$ yields the following evolution equation for the pdf of x_t, assuming the density exists.

$$\begin{aligned}\frac{\partial p_t(x)}{\partial t} &= \mathcal{A}^* p_t(x) \\ &= \frac{\partial}{\partial x}\left[-Axp_t(x)\right] + \int_{\mathcal{R}^n} p_t(x-y)\mu[x-y]p_{Bu}(y)\,dy - p_t(x)\mu[x] \quad (11)\end{aligned}$$

where $p_{Bu}(\cdot)$ is the pdf for the random vector Bu. It is observed that $\mathcal{A}^*$ is not a local operator because of the shifts in the convolution integral. The shifts also imply that the steady-state equation is classified as a differential delay equation.

Taking f in equation (10) to be the identity yields the $\mathcal{S}_t$-semimartingale decompositon of x_t, which is useful for deriving the MMSE nonlinear filter. This decomposition can also be deduced directly from (1) and (5) and is given by

$$\begin{aligned} x_t &= x_0 + G_t + H_t \quad (12) \\ G_t &= \int_0^t Ax_s + B\bar{u}\mu[x_s]\,ds \\ H_t &= \int_0^t BdR_s. \end{aligned}$$

where H_t is the martingale by an important property of stochastic integrals [12] and G_t is predictable since it is continuous.

The final representation to be considered is a decomposition of the superposition integral. Let $\Phi(t)$ be the state transition matrix corresponding to the plant matrix A. Then x_t can be represented as

$$\begin{aligned} x_t &= \Phi(t)x_0 + \int_0^t \Phi(t-s)BdM_s \quad (13) \\ &= \Phi(t)x_0 + \Phi(t)V_t + \Phi(t)I_t \quad (14) \\ V_t &= \int_0^t \Phi(-s)B\bar{u}\mu[x_s]ds \quad (15) \\ I_t &= \int_0^t \Phi(-s)BdR_s \quad (16) \end{aligned}$$

Here $\Phi(t)$ is factored out of the integrals to enable the semimartingale decomposition involving I_t and V_t. This representation rather than (12) is used to prove square-integrability and mean-square-continuity of x_t because the integrand of V_t is a uniformly bounded function of x_s, while the integrand of G_t is not. These properties are proved below.

The process x_t is square integrable iff [17]

$$\max_{t \in \mathcal{R}^+} E\{\|x_t\|^2\} < +\infty.$$

By the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned} E\{\|x_t\|^2\} &\leq E\{A^2\} + E\{B^2\} + E\{C^2\} \\ &\quad + 2\left[(E\{A^2\}E\{B^2\})^{\frac{1}{2}} + (E\{A^2\}E\{C^2\})^{\frac{1}{2}} + (E\{B^2\}E\{C^2\})^{\frac{1}{2}}\right] \quad (17) \end{aligned}$$

where

$$\begin{aligned} A &= \|\Phi(t)x_0\| \\ B &= \|\Phi(t)V_t\| \\ C &= \|\Phi(t)I_t\| \end{aligned}$$

and where $\|x\|$ is the Euclidean norm for $x \in \mathcal{R}^n$ and $\|\Phi\|$ is the matrix norm induced by $\|x\|$. The equation $\dot{x}_t = Ax_t$ is assumed to be exponentially stable, which implies that there exist positive constants γ and η such that $\|\Phi(t)\| \leq \gamma e^{-\eta t}$ [20]. Also, for random x_0, we assume $E\{\|x_0\|^2\} < +\infty$ and recall that $\mu[x] \leq K$ for all x. Hence, the following inequalities are implied:

$$E\{\|\Phi(t)x_0\|^2\} \leq \|\Phi(t)\|^2 E\{\|x_0\|^2\} \quad (18)$$

$$\begin{aligned} &E\{\|\Phi(t)V_t\|^2\} \\ &\quad = E\left\{\int_0^t \int_0^t \bar{u}^2 \mu[x_s]\mu[x_\tau] B^T \Phi(t-s)^T \Phi(t-\tau) B ds d\tau\right\} \\ &\quad \leq \int_0^t \int_0^t \bar{u}^2 K^2 \left| B^T \Phi(t-s)^T \Phi(t-\tau) B \right| ds d\tau \\ &\quad \leq \bar{u}^2 K^2 \|B\|^2 \int_0^t \|\Phi(t-s)\| ds \int_0^t \|\Phi(t-\tau)\| d\tau \quad (19) \end{aligned}$$

One of the properties of an L^2-martingale R_t is that if C_t is a bounded predictable process, then the stochastic integral $\psi_t = \int_0^t C_s dR_s$ is again an L^2-martingale with predictable variation

$$\langle \psi, \psi \rangle_t = \int_0^t C_s^2 d\langle R, R \rangle_s,$$

and $E\{\psi_t^2\} = E\{\langle\psi,\psi\rangle_t\}$ [17]. By a straightforward generalization to the vector case of (16), it follows that

$$\begin{aligned} & E\{\|\Phi(t)I_t\|^2\} \\ &= \text{Trace}\, E\left\{\int_0^t \Phi(t-s)B\overline{u^2}\mu[x_s]B^T\Phi(t-s)^T ds\right\} \\ &\le E\left\{\int_0^t \overline{u^2}\mu[x_s]\|\Phi(t-s)\|^2\|B\|^2 ds\right\} \\ &\le \overline{u^2}K\|B\|^2\int_0^t \|\Phi(t-s)\|^2 ds \qquad (20) \end{aligned}$$

Because $\dot{x} = Ax$ is exponentially stable, all of the integrals of $\Phi(t-s)$ are bounded by a constant [20]. Thus the bound on $E\{\|x_t\|^2\}$ does not depend on time, and x_t is square integrable.

Next x_t is shown to be mean square continuous. This property justifies the use of innovations in deriving the MMSE linear estimators. In order to show mean square continuity we must show

$$\lim_{w\to t} E\{\|x_t - x_w\|^2\} = 0.$$

Let $w < t$. The representation in equation (13) may be used to write

$$x_t - x_w = [\Phi(t-w) - I]x_w + \int_w^t \Phi(t-s)BdM_s.$$

Steps parallel to (14) and (17) yield the inequality:

$$\|x_t - x_w\|^2 \le \|\Phi(t-w) - I\|^2\|x_w\|^2 + \|\Phi(t-w)V_{w,t}\|^2 + \|\Phi(t-w)I_{w,t}\|^2$$

where $V_{w,t}$ and $I_{w,t}$ are defined the same as in (15) and (16), respectively, except the lower integration limit of 0 is replaced by w. Continuation of the same procedures yields inequalities nearly identical to (19) and (20), differing only in the lower integration limit. Further simplification is possible by using the inequalities below [20]. Let $\alpha = \|A\|$.

$$\begin{aligned} \|\Phi(t-w) - I\| &\le \alpha\int_0^{t-w} \|\Phi(s)\|ds \\ &\le \frac{\alpha\gamma}{\lambda}\left[1 - e^{-\lambda(t-w)}\right] \\ \int_w^t \|\Phi(t-s)\|ds &\le \frac{\gamma}{\lambda}\left[1 - e^{-\lambda(t-w)}\right] \\ \int_w^t \|\Phi(t-s)\|^2 ds &\le \frac{\gamma^2}{\lambda}\left[1 - e^{-2\lambda(t-w)}\right] \end{aligned}$$

Substitution of these inequalities leads to the expression

$$\begin{aligned} E\{\|x_t - x_w\|^2\} \leq{} & \frac{\alpha\gamma}{\lambda} E\{\|x_w\|^2\} \left[1 - \mathrm{e}^{-2\lambda(t-w)}\right] \\ & + \bar{u}^2 K^2 \|B\|^2 \frac{\gamma^2}{\lambda^2} \left[1 - \mathrm{e}^{-\lambda(t-w)}\right]^2 + \overline{u^2} K \|B\|^2 \frac{\gamma^2}{\lambda} \left[1 - \mathrm{e}^{-2\lambda(t-w)}\right]. \end{aligned}$$

The boundedness of $E\{\|x_w\|^2\}$ thus implies $E\{\|x_t - x_w\|^2\} \rightarrow 0$ as $t \rightarrow w$.

3 The Linear Filter and Smoother

For the estimation problem, we assume the m-dimensional observation process to have the following form:

$$dy_t = Cx_t dt + dv_t \tag{21}$$

where the matrix C is such that A and C yield a completely observable system. The observation noise v_t is an m-dimensional Wiener vector, independent of x_t, with $E\{v_t v_s^T\} = \int_0^{t\wedge s} \Psi_\tau d\tau$.

Because x_t is mean square continuous, it belongs to the Hilbert space spanned by all mean square continuous random processes. Thus its optimal linear filter exists as the projection of x_t onto the growing subspace generated by the observation process y_t. But the *practicality* of the filter is not guaranteed. An interesting characteristic of the process x_t described in Section 2.2 is that it appears to be on the "borderline" of the set of processes for which recursive, finite-dimensional filters exist. This is because in the case of zero mean marks, the linear filter expression is simple and familiar, whereas for the case of non-zero mean marks, a recursive, finite-dimensional filter does not seem to be possible. In this section, we give the linear filter and fixed-lag smoother for the case of zero mean marks, and discuss the difficulties associated with the case of non-zero mean marks.

3.1 Zero-Mean Marks

When the marks have zero mean, i.e. when $\bar{u} = 0$, the input jump process M_t is a martingale and hence has orthogonal increments. The filtering problem is classified by Kailath [21] as the Stratonovich-Kalman-Bucy (SKB) problem, for which the filter equations are well known. Let $\hat{x}_t$ be the optimal linear estimate of x_t, $\tilde{x}_t = x_t - \hat{x}_t$, and $P(t) = E\{\tilde{x}_t\tilde{x}_t^T\}$. The linear filter equations are:

$$\begin{aligned} d\hat{x}_t &= A\hat{x}_t dt + P(t)C^T\Psi_t^{-1}d\nu_t \\ \dot{P}(t) &= AP(t) + P(t)A^T - P(t)C^T\Psi_t^{-1}CP(t) + B\overline{u^2}\;\overline{\mu[x_t]}B^T. \end{aligned} \tag{22}$$

where $\nu_t = y_t - \int_0^t C\hat{x}_s ds$ is the innovations process. The only unusual characteristic in these equations is the a priori expectation $\overline{\mu[x_t]} = E\{\mu[x_t]\}$.

Recall that the rate function μ is necessarily nonlinear, since it must be positive. We have found that for a scalar system and the simple function

$$\mu[x] = \begin{cases} k_1 & |x| < a \\ k_2 & |x| \geq a \end{cases}$$

for some $a > 0$, $\overline{\mu[x_t]}$ can be well approximated by numerically propagating the pdf according to equation (11) and computing $E\{\mu[x_t]\}$. We also note that for scalar systems with certain mark distributions, it is possible to derive the steady state pdf.

The fixed-lag smoother for the zero mean mark case can be derived using Kailath's procedure [23], except that martingale properties are invoked in the computation of the error covariance $P(s,t) = E\{\tilde{x}_s\tilde{x}_t^T\}$, and differentials are used instead of derivatives where appropriate. Let $\hat{x}_{t|t+\Delta}$ denote the optimal linear estimate of x_t given the observations up to $t + \Delta$, with $\Delta > 0$. The equations that constitute the smoother are

$$d\hat{x}_{t|t+\Delta} = d\hat{x}_t + P(t)d\xi_t + dP(t)\xi_t,$$

the equations in (22), and

$$d\xi_t = -[A + P(t)C^T\Psi_t^{-1}C]^T\xi_t dt - C^T\Psi_t^{-1}d\nu_t + \tilde{\Phi}(t+\Delta,t)^T C^T \Psi_{t+\Delta}^{-1} d\nu_{t+\Delta}$$

where $\tilde{\Phi}(t,s)$ is the state transition matrix associated with the plant matrix $\tilde{A}_t = A - P(t)C^T\Psi_t^{-1}C$ and which maps from time t to time s. Let $\Sigma_{t|t+\Delta}$ denote the error covariance of $\hat{x}_{t|t+\Delta}$. The reduction in error covariance due to the lag is given [23] by

$$\Sigma_{t|t+\Delta} - P(t) = P(t)\left(\int_t^{t+\Delta} \tilde{\Phi}(s,t)^T C^T \Psi_s^{-1} C \tilde{\Phi}(s,t) ds\right) P(t).$$

3.2 Nonzero-Mean Marks

The linear filtering problem becomes more complex when the marks of the input process have a nonzero mean. The complexity derives from the fact that the compensator (i.e. the non-martingale part) of the jump process becomes nonzero and random. Two approaches were used on this problem. The objective of the first approach was to derive the filter directly using the innovations method. The objective of the second approach was to find the Gaussian process with the same autocovariance, and then write the optimal filter for the Gaussian process. The merit of the innovations approach is that it produces a filter expression, which may be simplified as much as possible. The autocovariance approach is useful because it implies an interesting interpretation of the compensator. In both approaches, the difficulty arises in covariance equations involving $\mu[x_t]$. Both approaches are summarized below.

For the innovations method, it is useful to consider the perturbation of x_t from its mean $\bar{x}_t = E\{x_t\}$, denoted by $\delta x_t = x_t - \bar{x}_t$. An expression for $\bar{x}_t$ may be found by taking the expectation of both sides of (13). A representation for δx_t is then

$$\delta x_t = \Phi(t-s)\delta x_s + \int_s^t \Phi(t-r)B\overline{u}\delta\mu_r \, dr + \int_s^t \Phi(t-r)BdR_r \qquad (23)$$

where $\delta\mu_t = \mu[x_t] - \overline{\mu[x_t]}$ and R_t is defined in (5).

The projection form of the optimal linear filter is [21]

$$\widehat{\delta x}_t = \int_0^t E\{\delta x_t \tilde{x}_s^T\} C^T \Psi_s^{-1} d\nu_s.$$

Substitution of (23) and interchange of integration order in the $\delta\mu$ term yields

$$d\left(\widehat{\delta x}_t\right) = A\widehat{\delta x}_t dt + B\overline{u}\widehat{\delta\mu}_t dt + P(t)C^T\Psi_t^{-1}d\nu_t. \qquad (24)$$

Application of the orthogonality principle and some algebra yields the error covariance equation

$$\begin{aligned}\dot{P}(t) = {} & AP(t) + P(t)A^T + B\overline{u}P_{\mu x}^T(t) + P_{\mu x}(t)\overline{u}B^T \\ & + B\overline{u^2}\,\overline{\mu[x_t]}B^T - P(t)C^T\Psi_t^{-1}CP(t),\end{aligned} \qquad (25)$$

where

$$P_{\mu x}(t) = C_{\mu x}(t) - E\{\widehat{\delta\mu}_t \widehat{\delta x}_t\}$$

and $C_{\mu x}(t)$ is the covariance of $\mu[x_t]$ and x_t.

It is observed from equation (24) that for a recursive filter to exist, there must be a recursion for $\widehat{\delta\mu}_t$. However, there is reasonable doubt that such a recursion exists or is worth the effort to derive, since the dynamics of $\mu[x_t]$

are very complex. The presence of $\mu[x_t]$ as a factor in (9) indicates that the differential equations for any expectation involving $\mu[x_t]$ will depend on higher order moments of $\mu[x_t]$.

In the second approach, the differential generator is applied to $f(x_t) = x_{i,t}x_{j,t}$ ($x_{i,t}$ is the i^{th} component of x_t) to get a differential equation for $C_{xx}(t,t) = E\{\delta x_t\ \delta x_t^T\}$. The result is

$$\begin{aligned}\dot{C}_{xx}(t,t) &= AC_{xx}(t,t) + C_{xx}(t,t)A^T \\ &+ B\overline{u}C_{\mu x}^T(t,t) + C_{\mu x}(t,t)\overline{u}B^T + \overline{\mu[x_t]}BB^T\overline{u^2}.\end{aligned} \tag{26}$$

In addition, it follows easily that

$$\frac{\partial}{\partial t}C_{xx}(t,s) = AC_{xx}(t,s) + B\overline{u}C_{\mu x}^T(t,s). \tag{27}$$

Now consider the Gaussian process x_t:

$$dx_t = Ax_tdt + B(\theta_tdt + dw_t^{(1)})$$

where $w^{(1)}$ is a scalar Wiener process, and θ_t is a scalar colored Gaussian noise which satisfies

$$d\theta_t = F\theta_tdt + Gdw_t^{(2)}$$

where $w^{(2)}$ is also a Wiener process. The covariance of x_t, when separated out of the covariance expression for the augmented state $[x_t^T, \theta_t]$, matches (26) and (27) if θ is replaced by $\overline{u}\mu[x_t]$ and the diffusion of $w(1)$ is assumed to be $\overline{u^2}\ \mu[x_t]$. There is also a match between the filter equations for the Gaussian process and the equations (24) and (25), when the appropriate notational substitutions are made.

These similarities imply that the compensator of the input jump process plays the role of the 'colored part' of the input noise. However, for the reasons given earlier, the evolution equation for $C_{\mu x}(t,t)$ is much more complex than for $C_{\theta\theta}(t,t)$ in the Gaussian case.

4 Conclusions

Martingale techniques have enabled a rigorous and complete derivation of certain representations and second-order statistics for the state process of interest. The linear filter and fixed-lag smoother were given for the case of zero-mean marks in the system disturbance. In the nonzero-mean case, the optimal linear filter did not seem to have a finite, recursive implementation. However, the form of the filter expression suggests that if a recursive linear filter exists for the compensator of an arbitrary jump process disturbance, then a recursive linear filter may exist for the state process.

Appendix

The following proposition is similar to a Lemma of Segall [14, p. 85], which proves that a counting process with a continuous compensator is locally square integrable. Here, the same property is proved for a jump process whose underlying counting process has an absolutely continuous compensator (i.e. an intensity) and whose marks are iid and mean square bounded.

Proposition: Let M_t be a jump process whose counting process N_t has an S_t-intensity such that $\lambda_t \leq K$ for all t. Let the jumps or marks of M_t be independent random variables with mean value $\overline{u}$ and mean square value $\overline{u^2}$. It follows from Section 2.1 that $R_t = M_t - \int_0^t \overline{u}\lambda_s \, ds$ is an S_t-martingale. Below it is proven that R_t is locally square integrable, and further, that R_t is an L^2-martingale.

Proof: By definition, the S_t-martingale R_t is locally square integrable if there exists a family of S_t stopping times $(T_n, n \geq 0)$, satisfying the properties $T_n \leq T_{n+1}$ and

$$\lim_{n \to +\infty} T_n = +\infty \quad \text{(a.s.)}, \tag{28}$$

such that for each n,

$$\max_{0 \leq t} E\{R^2_{t \wedge T_n}\} < +\infty.$$

where $t \wedge T_n = \min(t, T_n)$.

Define T_n as the time of the nth jump of N_t. Since $\{\omega : T_n(\omega) \leq t\} = \{\omega : N_t(\omega) > n\}$, each T_n qualifies as an S_t-stopping time.

To prove equation (28), we require only that $E\{N_t\} < +\infty$ for each $t \geq 0$, which follows from the bound on λ_t. Define the stopping time T as follows.

$$T = \lim_{n \to +\infty} T_n$$

It is observed that $T = +\infty$ (a.s.) iff $\Pr\{T > t\} = 1$ for all $t < +\infty$. Suppose the opposite is true, i.e. that there exists $\alpha < +\infty$ such that $\Pr\{T \leq \alpha\} > 0$. By the fact that N_t is increasing, we have that $E\{N_\alpha\} = +\infty$, which contradicts our assumption, therefore (28) is proved.

Finally, we address the process R_t. In Section 2.1, it was shown that $Q_t = \int_0^t \overline{u}\lambda_s\, ds$ is the compensator of R_t. Expansion of $R^2_{t\wedge T_n}$, and use of the triangle and Cauchy-Schwarz inequalities yields

$$\begin{aligned} E\{R^2_{t\wedge T_n}\} &\leq E\{M^2_{t\wedge T_n}\} \\ &\quad + 2\left[E\{M^2_{t\wedge T_n}\}E\{Q^2_{t\wedge T_n}\}\right]^{\frac{1}{2}} + E\{Q^2_{t\wedge T_n}\} \end{aligned}$$

The definition of T_n implies:

$$E\{M^2_{t\wedge T_n}\} \leq E\left\{\left(\sum_{i=1}^{n} u_i\right)^2\right\} \leq n^2\overline{u^2}.$$

The remaining term is bounded as follows.

$$E\{Q^2_{t\wedge T_n}\} \leq \overline{u}^2 K^2 T_n^2$$

Therefore, $E\{R^2_{t\wedge T_n}\} < +\infty$ and R_t is locally square integrable.

To prove square integrability of x_t, it is convenient to be able to define the quadratic variance of R_t without having to use the stopping times $\{T_n\}$. The property we desire is for R_t to be an L^2-martingale, which means that for each $t \geq 0$, $E\{R_t^2\} < +\infty$ [17, p.112]. Since $R_{t\wedge T_n}$ is square integrable, its quadratic variation exists [19, p.238], and is defined as the predictable compensator for the quadratic variation $[R,R]_{t\wedge T_n}$. Since $R_{t\wedge T_n}$ has no Wiener component,

$$[R,R]_{t\wedge T_n} = \sum_{0<s\leq(t\wedge T_n)} (\Delta R_s)^2$$

where the summation is over all jumps in R_s up to time $t \wedge T_n$. Since Q_t is continuous, $(\Delta R_{t\wedge T_n})^2 = (\Delta M_{t\wedge T_n})^2$. Therefore $[R,R]_{t\wedge T_n}$ is a jump process with the same counting process as $M_{t\wedge T_n}$, but whose marks are the square of the marks of $M_{t\wedge T_n}$. We may deduce from equation (5) that

$$< R,R >_{t\wedge T_n} = \int_0^{t\wedge T_n} \overline{u^2}\lambda_s\, ds.$$

It is known that $< R,R >_{t\wedge T_n}$ also compensates $R^2_{t\wedge T_n}$, therefore

$$\begin{aligned} E\{R^2_{t\wedge T_n}\} &= E\{< R,R >_{t\wedge T_n}\} \\ &\leq \overline{u^2}K(t\wedge T_n). \end{aligned}$$

Taking the limit of both sides as $n \rightarrow +\infty$ yields

$$E\{R_t^2\} \leq \overline{u^2} K t.$$

References

[1] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, Vol. **12**, pp. 601–611, 1976.

[2] R. L. Moose, "An adaptive state estimation solution to the maneuvering target problem," *IEEE Trans. Automatic Control*, Vol. **AC-20**, No. 3, pp. 359–362, June 1975.

[3] D. L. Snyder, *Random Point Processes*, Wiley, New York, 1975.

[4] H. Kwakernaak, "Filtering for systems excited by Poisson white noise," Eds.: A. Bensoussan, J. L. Lions, *Control Theory, Numerical Methods, and Computer Systems*, Springer Lecture Notes in Economics and Mathematical Systems, Vol. **107**, Berlin, pp. 468–492, 1975.

[5] H. Kwakernaak, "Estimation of pulse heights and arrival times," *Automatica*, Vol. **16**, pp. 367–377, 1980.

[6] S. P. Au and A. H. Haddad, "Suboptimal sequential estimation-detection scheme for Poisson driven linear systems," *Information Sciences*, Vol **16**, pp. 95-113, 1978.

[7] R. M. Rogers, "Optimal and suboptimal filtering for time-invariant systems excited by compound Poisson processes," *Proc. 1984 American Control Conf.*, San Diego, CA, June 6–8, 1984, Vol. **3**, pp. 1486–1488.

[8] J. Goutsias and J. M. Mendel, "Optimal simultaneous detection and estimation of filtered discrete semi-Markov chains," *IEEE Trans. Info. Theory*, Vol.

[9] D. D. Sworder and V. G. Robinson, "Feedback regulators for jump parameter systems with state and control dependent transition rates," *IEEE Trans. Auto. Control*, Vol. **AC-18**, No. 4, pp. 355–360, August 1973.

[10] Alan F. Karr, *Point Processes and Their Statistical Inference*, Marcel Dekker, Inc., New York, 1986.

[11] P. Brémaud, *Point Processes and Queues, Martingale Dynamics*, Springer-Verlag, New York, 1981.

[12] A. Segall, "Stochastic processes in estimation theory," *IEEE Trans. Inf. Theory*, Vol. **IT-22**, pp. 275–286, May 1976.

[13] A. Segall and T. Kailath, "The modeling of randomly modulated jump processes," *IEEE Trans. Inf. Theory*, Vol. **IT-21**, No.2, pp. 135–143, March 1975.

[14] A. Segall, "A martingale approach to modeling, estimation and detection of jump processes," Ph.D. dissertation, Dep. Elec. Eng., Stanford Univ., Calif., 1973.

[15] M. A. Ingram and A. H. Haddad, "Optimal and suboptimal filtering for linear systems driven by self-excited Poisson processes," *Proc. Twenty-Fifth Annual Allerton Conference on Communication, Control, and Computing*, Sept. 30– Oct. 2, 1987, Monticello, Illinios, pp. 426–435.

[16] M. A. Ingram and A. H. Haddad, "A sequential detection approach to state estimation of linear systems driven by self-excited point processes," *Proc. 27th IEEE Conference on Decision and Control*, Houston, TX, December 1988.

[17] Michel Métivier, *Semimartingales, a Course on Stochastic Processes*, Walter de Gruyter & Co., Berlin, 1982.

[18] M.H.A. Davis, "Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models," *Journal of the Royal Statistical Society*, Ser. B, **46**, No. 3, pp. 353–388.

[19] Eugene Wong and Bruce Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.

[20] Roger W. Brockett, *Finite Dimensional Linear Systems*, John Wiley and Sons, Inc., New York, 1970.

[21] T. Kailath, "A note on least squares estimation by the innovations method," *SIAM J. Contr.*, Vol. **10**, No. 3, pp. 477-486, August 1972.

[22] Andrew H. Jazwinsky, *Stochastic Processes and Filtering Theory*, Academic Press, Inc., San Diego, 1970.

[23] T. Kailath and P. Frost, " An innovations approach to least-squares estimation, Part II: Linear smoothing in additive white noise," *IEEE Trans. Automatic Control,* Vol. **AC-13**, No. 6, pp. 655–660, December 1968.

APPENDIX N

E. I. Verriest and S. W. Gray, "Robust Design Problems: A Geometric Approach", in Linear Circuits, Systems and Signal Processing: Theory and Application, Byrnes, Martin and Saeks, eds., Elsevier 1988.

Linear Circuits, Systems and Signal Processing:
Theory and Application
C.I. Byrnes, C.F. Martin and R.E. Saeks (editors)

ROBUST DESIGN PROBLEMS: A GEOMETRIC APPROACH

Erik I. Verriest and W. Steven Gray

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250*

Analogous to the correspondence between observability and identification, a correspondence relating controllability to a "dual" of the identification problem: the "DESIGN"-problem is established. This amounts to the choice of a realization or approximation of a desired system response, e.g., in view of minimizing the effects of component tolerances in analog systems or finite wordlength effects in the discrete case. A geometric approach to the design problem is presented, and its solution given under a useful criterion for optimality. For linear time invariant systems, the minimum sensitivity realizations are linked to the Balanced Realizations.

1. THE PROBLEM DEFINITION AND HISTORY

This paper deals with a new geometric approach to the robustness problem. Classically, the sensitivity properties of a given realization have been investigated, via a "sensitivity system" [12,3], or via the operator form [11]. The questions of robustness with respect to variations of certain structural parameters is closely related to this problem, and treated by Ackermann in [1]. A geometric point of view was recently introduced by Delchamps [2], and applied to compensation and feedback. Our emphasis will be in optimal implementations of systems with quantized or inaccurate parameters.

Consider a linear time invariant system (A,B,C) with m inputs and p outputs. This may be a model for a real system one wants to simulate, the implementation of a digital or analog filter, or an observer-controller implementating an optimal regulator for some given plant. In all these applications, only the relationship between the input and the output of the implemented system is important. Usually the so-called "Canonical Forms" are implemented because they minimize the number of parameters and allow for a pipelined realization. This corresponds to minimal complexity, a quality that may be important if the operation count becomes important. However, a minimal set of parameters has no redundancy, and therefore, high sensitivity.

This paper investigates how the nonuniqueness of the state space realizations can be utilized to determine optimal parameterizations under various measures of "optimality" or robustness.

*This research is supported by the U.S. Air Force under Contract No. F08635-84-C-0273, and AFOSR-87-0308.

Because addition and scalar multiplication of systems have no meaningful natural interpretations, the realization space is simply assumed to have the structure of an affine space of dimension n(n+m+p). The space is given the structure of a Riemannian manifold by introducing an Euclidean metric in the tangent space at each point. For instance, in the analysis and design of the finite wordlength effects with fixed point processing, a uniform metric for all tangent spaces is appropriate, whereas for floating point processing, a metric varying smoothly from point to point is more appropriate.

This space can be resolved (i.e., partitioned into equivalence classes) into disjoint sets, corresponding to different input/output behaviors. For a particular realization, the proximity of neighboring sheets will be an indication for the robustness or sensitivity of this realization. These geometric notions are made precise in Section 3, after giving a more philosophical introduction in Section 2 on the design problem and its relation with other systems problems. This theory is applied to systems design in Section 4. The most interesting result is the one relating the minimum sensitivity (under the fixed point metric) realizations to the balanced realizations.

2. SITUATION OF THE PROBLEM

Consider the phenomenon "linear system" as a mapping σ from a suitable subset of the cartesian product of input functions (U) and realizations (L) to the set of output functions (Y). For continuous linear time-invariant systems, the mapping stands for the convolution operator

$$\sigma : U \times L \rightarrow Y : (u(\cdot), S) \rightarrow y(\cdot)$$

$$y(t) = \int_{-\infty}^{t} Ce^{A(t-\tau)}Bu(\tau)d\tau$$

For discrete systems a similar expression results. We can now look at the marginal maps derived from the linear system operator. In particular, if S = (A,B,C) is fixed, we define the usual linear input/output map as

$$\sigma_s : U \times \{S\} \rightarrow Y : u(\cdot) \rightarrow y(\cdot)$$

On the other hand, for a fixed input u(·), the marginal maps

$$\sigma_u : \{u\} \times L \rightarrow Y : S \rightarrow y(\cdot)$$

associate with each realization S, e.g. the impulse response h(t) if u(t) = δ(t), or the transfer function H(p) characterizing the steady state response to a sinusoid $u(t) = e^{pt}$ of complex frequency p.

The <u>control</u> and <u>decon</u>
in the sense that the fo
the latter to a <u>left-inve</u>
is implicit in the prob
output, invariably "futur
problem one acts on obs
y(·). Similarly, the co
<u>system identification</u> pro
or time series, and hence
finding a <u>right-inverse</u>
desired "future" behavior.

In the identification
ties due to the finite ob
isolation. Similary, un
parameter settings neces
"uniquely" an "optimal" s
distance or norm in the do

3. MAIN RESULTS

A summary of some kn
realizations is first give
on an abstract level.

3.1 <u>The Geometric Stru</u>

Let $L_{m,n,p}$ be the rea
matrices (F,G,H) of dimens
with an affine structure w
there is an attached vect
$R^{n(m+n+p)}$. The group Gl_n
$(A,B,C) \rightarrow (A,B,C)^T = (TAT^{-}$
state space z = Tx. T
Restricted to the complete
systems, the action of
observability) and the quo
a smooth (real) analytic m
set of equivalence classe
submanifolds [5]. This
identification, and is we
of continuous canonical fo

Since the isotropy sub
realizations, its dimensio

systems have no meaningful simply assumed to have the . The space is given the an Euclidean metric in the analysis and design of the g, a uniform metric for all point processing, a metric iate.

into equivalence classes) t/output behaviors. For a sheets will be an indica- lization. These geometric ving a more philosophical nd its relation with other s design in Section 4. The imum sensitivity (under the lizations.

mapping σ from a suitable (U) and realizations (L) to ear time-invariant systems,

$y(\cdot)$

We can now look at the rator. In particular, if ut/output map as

$y(\cdot)$

nal maps

$(\cdot)$

se response $h(t)$ if $u(t) =$ he steady state response to

The control and deconvolution problems are inverse problems for the map σ_s in the sense that the former relates to the derivation of a right-inverse and the latter to a left-inverse of the map. Moreover, a certain causal structure is implicit in the problem. In designing a control to achieve a desired output, invariably "future" actions are understood, while in the deconvolution problem one acts on observed data, and thus relates the "past" of $u(\cdot)$ and $y(\cdot)$. Similarly, the construction of a left-inverse for σ_u pertains to the system identification problem, invariably tied to an observation of functions or time series, and hence relating the "past" of $y(\cdot)$ to the system. Finally, finding a right-inverse of σ_u is the problem of "designing" a system with desired "future" behavior.

In the identification problem, the measured data necessarily has uncertainties due to the finite observation time, finite memory effects, and imperfect isolation. Similary, uncertainties interfere with the design problem: the parameter settings necessarily have finite precision. In order to find "uniquely" an "optimal" solution to these problems, one introduces a suitable distance or norm in the domain and range spaces [14].

3. MAIN RESULTS

A summary of some known results on the geometry of systems and their realizations is first given. The next subsection discusses the robust design on an abstract level.

3.1 The Geometric Structure of the Realization Space

Let $L_{m,n,p}$ be the realization space, i.e. the space of all triples of matrices (F,G,H) of dimension n x n, n x m, and p x n over R. Endow this space with an affine structure with vector space $R^{n(m+n+p)}$. Hence, at each point S, there is an attached vector space T_SL (the tangent space at S), isomorphic to $R^{n(m+n+p)}$. The group $Gl_n(R)$ acts differentiably on the right to $L_{m,n,p}$, via $(A,B,C) \to (A,B,C)^T = (TAT^{-1},TB,CT^{-1})$ corresponding to a change of base in the state space $z = Tx$. The quotient space is non-Hausdorff in general. Restricted to the completely reachable (or dually, the completely observable) systems, the action of $Gl_n(R)$ is free (as a consequence of reachability/observability) and the quotient space (set of orbits) $M^{cr}_{m,n,p} = L^{cr}_{m,n,p}/Gl_n(R)$ is a smooth (real) analytic manifold (hence Hausdorff) of dimension n(m+p). The set of equivalence classes of minimal realizations $M^{co,cr}_{m,n,p}$ are analytic open submanifolds [5]. This space, called parameter space, is crucial in identification, and is well studied (e.g. in relation to the (non)existence of continuous canonical forms [5], and degeneration phenomena [6]).

Since the isotropy subgroup is trivial for all reachable or observable realizations, its dimension is constant on $L^{co,cr}_{m,n,p}$, and hence, the orbits of

$Gl_n(R)$ form a foliation F of $L^{co,cr}_{m,n,p}$ of dimension n(m+p) [9]. The field of tangent spaces to the leaves form an n(m+p)-dimensional subbundle τ(F) of the tangent bundle, called the tangent bundle to F. The quotient bundle ν(F) = TL/τ(F) is called the normal bundle to F.

Our interest is not in the universal parameterization, but in the orbits under the action of $Gl_n(R)$ itself. These orbits are open, and the boundary points of reachable realizations are nonreachable realizations. The explicit form of the closure of the orbits was addressed in [8]. We shall endow the tangent bundle $TL^{co,cr}_{m,n,p}$ with a positive definite metric

$$\langle\cdot,\cdot\rangle_S : T_S L \times T_S L \rightarrow R \quad \text{for all } S \text{ in } L^{co,cr}_{m,n,p} .$$

3.2 <u>The Robust Design Problem: A Geometric Approach</u>

Before proceeding with our system design, we shall prove a general result on sensitivity:

<u>Definition</u>: Let Θ be an N-dimensional open subset of an affine space A^N of design parameters (configurations). By an <u>Observable</u>, we shall mean any smooth function f : Θ → R which has no critical points.

Any two configurations θ_1 and θ_2 in the parameter space are indiscernible by observation of f if $f(\theta_1) = f(\theta_2)$. This allows us to regard two parameterizations yielding the same observable value(s) as being the same (or equivalent) for some purpose. In a systems context, observables are, for instance, a mapping from the realization space to the transfer function (scalar case) evaluated at a particular frequency, or the impulse response evaluated at a specific instant, i.e., the "system functions" [3].

An observable induces a partition of Θ into equivalence classes, known as a foliation. In this case, the submanifolds are the level surfaces of f, and have dimension N-1. There exists a vector field normal (in terms of some arbitrarily chosen Riemannian metric) to the leaves.

The whole issue of the sensitivity problem is now to find the points on the leaves corresponding to a maximal "separation" of the leaves of the foliation.

3.2.1 <u>Riemannian Metrics</u>. If Θ is paracompact, then a Riemannian structure G can be put on Θ (or, more exactly, on its tangent bundle). This means that for each θ ε Θ, a symmetric, positive definite bilinear form G_θ is defined on the vector space $T_\theta\Theta$, such that G defines a metric on TΘ, i.e. is a smooth section of the vector bundle $T^o_2\Theta$. Let $^\#$: $T^*\Theta \rightarrow T\Theta$ be the natural isomorphism of each space $T^*_\theta\Theta$ with $T_\theta\Theta$. If f is a smooth map, the gradient of f is defined as the element $df^\#$ of TΘ (i.e. the vector field corresponding under the map $^\#$ to the differential form df). In the local coordinates, this is given by

where the summation conver
metric tensor $\{g_{ij}\}$

The squared norm of the gr

$\|\nabla_G$

If θ is foliated by f,
N-1-dimensional subspace o

3.2.2 <u>Extremal Sens</u>
(with respect to an observ
$\frac{1}{2}\|\nabla_G f\|^2$ over the leaf char

A worst case analysis l
$G(\nabla_G f,\nabla_G f)^{1/2}$, or equivale

This scalar field induces
However, note that $d\overline{h}^\#$ = d
Its projection on the tang
$dG(df^\#,df^\#)^\# - \lambda df^\#$ to the
<u>Theorem 1</u>: If f is an
points of extremal sensiti
the equation

<u>Proof</u>: The stated conditio
optimization problem.

The gradients of $\overline{h}$ and f
particular, for the uniform

while for the relative me
with the floating point ari

$[\mathrm{diag}(\theta)\mathrm{diag}(f_\theta$

n(m+p) [9]. The field of
.mensional subbundle $\tau(F)$ of
The quotient bundle $\nu(F)$ =

rization, but in the orbits
are open, and the boundary
realizations. The explicit
in [8]. We shall endow the
ric

S in $L^{co,cr}_{m,n,p}$.

roach
ll prove a general result on

et of an affine space A^N of
le, we shall mean any smooth

r space are indiscernible by
s to regard two parameteri-
ing the same (or equivalent)
ables are, for instance, a
sfer function (scalar case)
lse response evaluated at a

ivalence classes, known as a
he level surfaces of f, and
d normal (in terms of some
).
ow to find the points on the
the leaves of the foliation.
, then a Riemannian structure
nt bundle). This means that
linear form G_θ is defined on
ric on $T\Theta$, i.e. is a smooth
$\to$ $T\Theta$ be the natural isomor-
th map, the gradient of f is
field corresponding under the
coordinates, this is given by

$$\nabla_G f = g^{ij} \frac{\partial f}{\partial \theta^i} \frac{\partial}{\partial \theta^j}$$

where the summation convention is used. The matrix $\{g^{ij}\}$ is the inverse of the metric tensor $\{g_{ij}\}$

$$g_{ij} = G(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \theta^j})$$

The squared norm of the gradient is

$$\|\nabla_G f\|^2 = G(\nabla_G f, \nabla_G f) = g^{ij} \frac{\partial f}{\partial \theta^i} \frac{\partial f}{\partial \theta^j}$$

If Θ is foliated by f, then the tangent space Δ_θ to the leaf through θ is an N-1-dimensional subspace of $T_\theta\Theta$. ■

3.2.2 Extremal Sensitivity Theorem. Points of extremal sensitivity (with respect to an observable $f(\theta)$), are determined by minimization of $L(\theta) = \frac{1}{2}\|\nabla_G f\|^2$ over the leaf characterized by a particular value of the observable f.

A worst case analysis leads to the minimization of the gradient norm $\|\nabla_G f\| = G(\nabla_G f, \nabla_G f)^{1/2}$, or equivalently,

$$\overline{h} = \frac{1}{2}\|\nabla_G f\|^2$$

This scalar field induces a vector field in the tangent space Δ_θ of the leaf. However, note that $d\overline{h}^{\#} = dG(df^{\#}, df^{\#})^{\#}$ is, in general, not tangent to the leaf. Its projection on the tangent space to the leaf at θ yields the tangent vector $dG(df^{\#}, df^{\#})^{\#} - \lambda df^{\#}$ to the leaf through θ, for some $\lambda \in R$.

Theorem 1: If f is an observable for the parameter space (Θ, G), then the points of extremal sensitivity with respect to f are implicitly determined by the equation

$$dG(df^{\#}, df^{\#})^{\#} - \lambda df^{\#} = 0$$

Proof: The stated condition is the Euler-Lagrange equation for the constrained optimization problem.

The gradients of $\overline{h}$ and f are aligned at the extremal sensitivity points. In particular, for the uniform metric, $g_{ij} = \delta_{ij}$, the condition specializes to

$$(f_{\theta\theta}(\cdot) - \lambda I)f_\theta(\cdot) = 0$$

while for the relative metric $g_{ij} = \delta_{ij}/\theta^i\theta^j$, which is useful in connection with the floating point arithmetic, the condition is

$$[\mathrm{diag}(\theta)\mathrm{diag}(f_\theta) + \mathrm{diag}(\theta^2)f_{\theta\theta}]\mathrm{diag}(\theta^2)f_\theta = \lambda \mathrm{diag}(\theta^2)f_\theta$$

In the latter case, a simpler form is obtained by using the "generalized" gradient $\hat{\nabla}f$ with components $\theta_i\partial f/\partial\theta_i$ instead; corresponding to the generalized Hessian $\hat{H} = \mathrm{diag}(\hat{\nabla}f) + \mathrm{diag}(\theta)f_{\theta\theta}\mathrm{diag}(\theta)$. We state what was just shown as an important

Corollary: The extremal sensitivity points of (Θ,G), where G is the uniform or relative metric, are the points where the gradient $df^{\#}$ is in the eigenspace of the generalized Hessian operator $\hat{H} : T_\theta\Theta \rightarrow T_\theta\Theta$, i.e.

$$(\hat{H}(f) - \lambda I)df^{\#} = 0$$

4. APPLICATION TO ROBUST REALIZATIONS

Express the parameterization in terms of the components of a factorization of the system Hankel matrix H = OR, where O and R are, respectively, the observability and reachability matrices of the realization. The Hankel matrix defined as a map with domain $L_{m,n,p}$ plays the role of a multidimensional observable. The continuous time systems design under the uniform metric is discussed, for square (p = m) systems only.

Definitions: Let $L_2^m[0,\infty)$ be the Hilbert space of m-vector functions with inner product $\langle x(\cdot),y(\cdot)\rangle = \int_0^\infty x(t)'y(t)dt$. The reachability operator $\underline{R} : L_2^m[0,\infty) \rightarrow R^n$ for a realization (A,B,C) is defined by $\underline{R}u(t) = \int_0^\infty e^{At}Bu(t)dt$. Its adjoint $\underline{R}^*$ is the operator $\underline{R}^* : R^n \rightarrow L_2^m[0,\infty) : \underline{R}^*x = B'e^{A't}x$. The observability operator is $\underline{O} : R^n \rightarrow L_2^p[0,\infty) : \underline{O}x = Ce^{At}x$. Since $\underline{R}$ and $\underline{O}$ have a finite dimensional range and domain, respectively, they are compact, and their composition $\underline{OR}$ is also compact [7]. Finally, we introduce the Hankel operator $\underline{H} : L_2^m[0,\infty) \rightarrow L_2^p[0,\infty) : \underline{H}u(t) = \int_0^\infty h(t+\tau)u(\tau)d\tau$, where $h(t) = Ce^{At}B$. It is readily verified that indeed $\underline{H} = \underline{OR}$. An operator $\Lambda : L_2^m[0,\infty) \rightarrow L_2^m[0,\infty)$ satisfying $\Lambda\Lambda^* = \Lambda^*\Lambda = \mathrm{Id}$ (the Identity operator) is called isometric. We shall also assume that the set $\{e_i\}_{i=1}^n$ is the standard basis for R^n and that the functions $\{\psi_i\}_{i=1}^\infty$ form a complete orthonormal basis in $L_2^m[0,\infty)$.

$$\underline{H} = \sum_{uv} h_{uv}|\psi_u\rangle\langle\psi_v| \qquad \underline{R} = \sum_{ij} r_{ij}|e_i\rangle\langle\psi_j| \qquad \underline{O} = \sum_{kl} o_{kl}|\psi_k\rangle\langle e_l|$$

The matrix representations $[h_{ij}]$, $[r_{ij}]$, and $[o_{ij}]$ will be, respectively, denoted by $\mathrm{Mat}(\underline{H})$, $\mathrm{Mat}(\underline{R})$, and $\mathrm{Mat}(\underline{O})$. By Vec(M), we mean the vector formed by stacking the elements of the matrix M columnwise.

It is now possible to state our first auxiliary result:

Lemma: Let $E : L_2^m[0,\infty) \rightarrow L_2^m[0,\infty)$ be such that $\mathrm{Tr}\Lambda E = 0$ for all isometric operators Λ, then E = 0.

Proof: Suppose E has the singular value decomposition [7, p. 261]

where $\{u_i\}$ and $\{v_i\}$ ar

$\Sigma|v_j\rangle\langle u_j|$ yields $\Sigma\theta_i = 0.$

must have all $\theta_i = 0$ and,

In order to apply the

the affine space formed

eter vector is $\theta' = [\mathrm{Vec}($

case [4], we shall consi

$m_o(f_\Lambda)$ the leaf on which

Theorem 2: The extremal

$\underline{RR}^* = \underline{O}^*\underline{O}.$

Proof: Substitute the bra

$f(\theta)$, and use the orthon

time problem to the matrix

the Corollary to Theorem 1

M

Expressing $\mathrm{Mat}(\underline{R})\mathrm{Mat}(\underline{R})'$

condition in terms of the

Corollary: The minimal

minimal realization of $\underline{H}$ a

orthogonal transformation)

Proof: Observe first that

choice of Λ, and therefore,

All extremal sensitivity

$\bigcap_\Lambda M_o(f_\Lambda)$. By the lemma,

submanifold characterized

operator $\underline{H}$ under the actio

$\underline{O}^*\underline{O}$ so that

$\langle x,\underline{R}$

which leads to the equality

Realizations having this p

similarity transformation

gramians) [10,13]. The s

solutions obtained correspo

infinitesimal variations in

by using the "generalized"
esponding to the generalized
e what was just shown as an

G), where G is the uniform or
: $df^{\#}$ is in the eigenspace of
i.e.

components of a factorization
nd R are, respectively, the
alization. The Hankel matrix
role of a multidimensional
under the uniform metric is

m-vector functions with inner
ility operator $\underline{R} : L_2^m[0,\infty) \to$
$= \int_0^\infty e^{At}Bu(t)dt$. Its adjoint
$B'e^{A't}x$. The observability
ince $\underline{R}$ and $\underline{O}$ have a finite
hey are compact, and their
introduce the Hankel operator
, where $h(t) = Ce^{At}B$. It is
erator $\Lambda : L_2^m[0,\infty) \to L_2^m[0,\infty)$
r) is called isometric. We
tandard basis for R^n and that
basis in $L_2^m[0,\infty)$.

$$\underline{O} = \sum_{kl} o_{kl}|\psi_k\rangle\langle e_l|$$

$[o_{ij}]$ will be, respectively,
, we mean the vector formed by

y result:
at $Tr\Lambda E = 0$ for all isometric

ition [7, p. 261]

$$E = \Sigma\theta_i|u_i\rangle\langle v_i|$$

where $\{u_i\}$ and $\{v_i\}$ are orthonormal sets in $L_2^m[0,\infty)$, then choosing Λ as $\Sigma|v_j\rangle\langle u_j|$ yields $\Sigma\theta_i = 0$. Since the singular values θ_i are nonnegative, we must have all $\theta_i = 0$ and, hence, $E = 0$. ∎

In order to apply the theory developed in the previous section, we consider the affine space formed by the matrix elements of $\underline{O}$ and $\underline{R}$, so that the parameter vector is $\theta' = [Vec(Mat(\underline{R})')', Vec(Mat(\underline{O}))']$. Analogous to the discrete case [4], we shall consider the observables: $f_\Lambda(\theta) = Tr\Lambda(\underline{H}-\underline{OR})$. Denote by $m_o(f_\Lambda)$ the leaf on which f_Λ is constant, zero say, then we have the:

<u>Theorem 2</u>: The extremal sensitivity points of $M_o(f_\Lambda)$ have the property that $\underline{RR}^* = \underline{O}^*\underline{O}$. ∎

<u>Proof</u>: Substitute the bra-ket expansions in the expression for the observable $f(\theta)$, and use the orthonormality of the bases. This reduces the continuous time problem to the matrix problem, solved in [4], where it was shown, based on the Corollary to Theorem 1, that the extremal sensitivity points satisfy

$$Mat(\underline{R})Mat(\underline{R})' = Mat(\underline{O})'Mat(\underline{O})$$

Expressing $Mat(\underline{R})Mat(\underline{R})'$ and $Mat(\underline{O})'Mat(\underline{O})$ in the basis $\{e_i\}_{i=1}^n$ gives then the condition in terms of the original operators: $\underline{RR}^* = \underline{O}^*\underline{O}$.

<u>Corollary</u>: The minimal sensitivity realizations on the $Gl_n(R)$ orbit of a minimal realization of $\underline{H}$ are the essentially balanced (i.e. balanced modulo an orthogonal transformation) realizations.

<u>Proof</u>: Observe first that the condition for an extremum did not depend on the choice of Λ, and therefore, must be true for all isometries, or observables f_Λ. All extremal sensitivity points of f_Λ belong, therefore, to the intersection $\bigcap_\Lambda M_o(f_\Lambda)$. By the lemma, the intersection of the manifolds $M_o(f_\Lambda)$ is the submanifold characterized by $\underline{H} = \underline{OR}$, i.e. the orbit of the system with Hankel operator $\underline{H}$ under the action of $Gl_n(R)$. Then, by the previous theorem, $\underline{RR}^* = \underline{O}^*\underline{O}$ so that

$$\langle x,\underline{RR}^*y\rangle = \langle x,\underline{O}^*\underline{O}y\rangle \qquad \forall x,y \in R^n$$

which leads to the equality of the Reachability and the Observability Gramian. Realizations having this property are essentially balanced, as an orthogonal similarity transformation will make them truly balanced (equal and <u>diagonal</u> gramians) [10,13]. The second variation property shows that the extremal solutions obtained correspond to minimum sensitivity solutions. Finally, all infinitesimal variations in the parameters of the factorizations of the Hankel matrix lead to second order variations in H. But small (first order) variations in the reachability and observability matrices are themselves linked to first order variations in the realization parameters. ∎

As shown by this corollary, it suffices to find an essentially balanced realization. The characterization as a factorization of the Hankel matrix is, therefore, independent of the size of the Hankel matrix considered, as long as it specifies the given input/output relation.

REFERENCES

[1] J. Ackerman, "Parameter Space Design of Robust Control Systems," IEEE Trans. Auto. Control, Vol. AC-25, No. 6, 1980.

[2] D.F. Delchamps, "New Geometric Approaches to Parameter Sensitivity in Feedback Systems," in Modelling, Identification and Robust Control, C.I. Byrnes and A. Lindquist, eds., North-Holland, 1986.

[3] P.M. Frank, "Introduction to System Sensitivity Theory," Academic Press, 1978.

[4] W.S. Gray and E.I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity," 1987 IEEE Conf. Dec. Control, Los Angeles, CA, December 1987.

[5] M. Hazewinkel, "(Fine) Moduli (Spaces) for Linear Systems: What Are They Good For?", in Geometric Methods in the Theory of Linear Systems, C.I. Byrnes and C. Martin, eds., North-Holland, 1979, pp. 125-193.

[6] M. Hazewinkel, "On Families of Linear Systems: Degeneration Phenomena," in Algebraic and Geometric Methods in Linear Systems Theory, C.I. Byrnes and C.F. Martin, eds., Lecture Notes in Mathematics, Vol. 18, 1980, pp. 157-189.

[7] T. Kato, "Perturbation Theory for Linear Operators," Springer-Verlag, 1976.

[8] A.S. Khadr and C. Martin, "On the Gln(R) Action on Linear Systems: The Orbit Closure Problem," in Algebraic and Geometric Methods in Linear System Theory," C.I. Byrnes and C.F. Martin, eds., Lectures in Applied Mathematics, Vol. 18, 1980.

[9] H.B. Lawson, Jr., "The Quantitative Theory of Foliations," Conf. Board of the Mathematical Sciences, 1977.

[10] B.C. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," IEEE Trans. Auto. Control, Vol. AC-26, No. 1, February 1981.

[11] J.G. Reid, P.S. Maybeck, R.B. Asher, and J.D. Dillow, "An Algebraic Representation of Parameter Sensitivity in Linear Time-Invariant Systems," J. Franklin Inst. 1, Vol. 301, Nos. 1 and 2, January-February 1976.

[12] R. Tomovic and M. Vukobratovic, "General Sensitivity Theory," American Elsevier, 1972.

[13] E.I. Verriest, "The Structure of Multivariable Balanced Realizations," Proc. 1983 Int'l. Symp. Circuits and Systems, Newport Beach, CA.

[14] J.L. Willems, "Models for Dynamic Systems," submitted to Dynamics Reported, 1986.

APPENDIX O

E. I. Verriest and A. H. Haddad, "Filtering and Implementation for Air-to-Air Target Tracking", Proc. 1988 American Control Conference, Atlanta, pp.143-148, June 1988.

FILTERING AND IMPLEMENTATION FOR AIR-TO-AIR TARGET TRACKING[1]

Erik I. Verriest and Abraham H. Haddad

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

This paper discusses some aspects of the design problem involved in the choice of a realization or approximation of a desired system behavior (as for instance dictated by the analytical solutions to a filtering problem) by parameters that can only be approximately adjusted, e.g., due to quantization, component tolerances (analog case) and finite wordlength (discrete case). The paper first addresses the mathematical characterization of this robustness problem, and its solutions under various criteria of optimality. Earlier results are here extended to multi-mode systems which can arise in nonlinear approximation problems. The feasibility of this approach in multi-mode filtering is shown, and is illustrated by an air-to-air tracking example.

1. INTRODUCTION

The air-to-air target tracking problem is highly nonlinear because of the nonlinear relations between measurements and dynamical states, and the different flight regimes that occur. Differences in Mach number and or geometry of the target induce large changes in the dynamical model. A good knowledge of the dynamical model is primordial to the design of good tracking filters, as the predictive behavior of the filters are determined by the dynamics of the system. Mach number changes with air density, hence altitude, and velocity, and is therefore coupled to the position and momentum of the target. These are of course state components of direct interest. As the target is maneuvring, perhaps beyond the anticipation of the tracker, its trajectory is modeled as a smooth stochastic process, the statistics of which are clearly dependent again on its position and momentum, as well as the geometry of the vehicle.

This paper investigates the tracking problem under the assumptions that the data sampling rate is sufficiently high. This implies small increments (as compared to the sizes of the domains in the different flight regimes) in the state variables from one sample to another, so that the same flight regime can be assumed over a large number of samples. Under this condition the system is reasonably well approximated by a piecewise affine stochastic system [1]. The transitions from one flight regime to another is determined by the state vector itself. Given enough (good) samples, the state estimator will have a good performance in each domain. At the transients from one domain to another, unmodeled uncertainty would be introduced because of the mismatches of the updates near the boundaries. A good filtering scheme needs likelihood type methods as developed in [2-3] to deal with this additional uncertainty. This typically leads to a filter composed of a parallel bank of Kalman filters, together with a likelihood updating scheme. In this paper, the assumptions ensure that the time spent in these transition regions is relatively small statistically speaking. We simply propose to artificially reset the state covariance whenever a state domain transition occurs. This simplifies the filter from the parallel bank plus likelihood estimator to a simple sequentially switched estimator.

The new feature of this paper is the optimal implementation of such a sequentially switched filter from the point of view of parameter sensitivity. Section 2 describes some typical problems in the implementation of systems, i.e. the minimization of the effects of component tolerances for analog systems, and the finite wordlength effects in digital systems. In section 3 the results are extended to a more general type of systems: the switched systems and piecewise linear systems. Finally in section 4, the air-to-air tracking example is discussed.

2. OPTIMAL IMPLEMENTATIONS FOR LINEAR MODELS

This work builds on the earlier work on robust design problems [4-5]. Consider a linear time invariant system (A,B,C) with m inputs, p outputs and McMillan degree n. This may be a model of a system to be faithfully simulated, the implementation of an analog or discrete filter, or an observer-controller implementing an optimal regulator for a given plant. As in all these applications, not the actual state coordinates, but the input-output relation is important, they are usually implemented by a so- called canonical form. The reason for this is that these implementations minimize the number of parameters, and allow a pipelined realization of the devices, e.g. the "Direct Form" realizations in digital signal processing. A minimal number of parameters corresponds to minimal complexity, an important quality if operation count becomes important. However, a minimal set of parameters has no redundancy, and therefore one may expect high sensitivity with respect to these parameters. It is clear that the freedom of coordinate basis of the implementation should be

[1] This work was supported by the U. S. Air Force under Grant AFOSR-87-0308

utilized to determine optimal realizations under various criteria for optimality. In particular, two issues seem to be important: sensitivity and clustering. The sensitivity requirement guarantees robustness of the actual implementation, while clustering deals with the parameter ranges. It relates to the problem of approximately implementing a certain system with parameters chosen from a finite set with fixed values.

The approach taken in [4] is geometric. The realization space $L_{m,n,p}$ is modeled as an $n(n+m+p)$ dimensional affine space with an Euclidean metric metric defined in the tangent space at each point. The Extremal Sensitivity Theorem asserts that the minimum sensitivity points of an observable are the points where the (generalized) gradient is in the eigenspace of the (generalized) Hessian. In the case of fixed point implementations, the uniform Euclidean metric is appropriate, and the gradient and Hessian correspond with the the usual notions in calculus. All results are therefore also "infinitesimal". One can reasonably so argue that in finite wordlength arithmetic, the notions of infinitesimal perturbations do not apply, as all perturbations are due to the truncation of the coefficients. This may indeed invalidate the above mentioned method. For this reason, we shall develop the analysis for non-infinitesimal perturbations in this paper. It will be shown that this approach makes a connection with the notion of clustering. In this section, by finite we shall mean non-infinitesimal. By L we denote a particular realization in $L_{m,n,p}$, and by M the equivalence class of all similar realizations having a particular input-output behavior, i.e. the orbit $\pi^{-1}(L)$ under $Gl_n(R)$. The equivalence class will be referred to as "system", the orbit space is denoted by $M_{m,n,p}$, and the projection map from $L_{m,n,p}$ to $M_{m,n,p}$ by π. Two problems related to the sensitivity and robustness are studied.

<u>Problem A: Discrimination:</u> How do we choose a realization of a given system M_o such that it is maximally distant from the orbit $\pi^{-1}(M)$, where M is another given system.

<u>Problem B: Worst Case Defects:</u> If the system parameters are perturbed over a fixed non-infinitesimal amount, and if f is a scalar system function (i.e. invariant under similarity) then find the realizations L of M_o, for which the f-perturbation

$$\max_{C_\Delta} \{ f(L+\Delta) - f(L) \} \quad \text{where } C_\Delta = \{ \Delta \mid \|\Delta\| = 1 \}$$

is minimal.

In both problems, the norm function in the realization space will be fixed to be the Eising norm (compatible with the uniform metric).

$$d_E^2 = \mathrm{Tr} \{ AA' + BB' + C'C \}$$

It follows then that if (A,B,C) is a solution to each of the above problems, then also any realization obtained from (A,B,C) by an orthogonal similarity transformation is also a solution. A third problem, related to problem A, is now introduced:

<u>Problem C: Clustering:</u> Find the realizations L of M with minimal system (Eising) norm.

This has the physical significance that cooperatively the components of the realization are as small as possible, hence clustered near zero. It is a special case of the more general (and more significant) clustering problem. Let $\Gamma = \{\gamma_1,\ldots,\gamma_M\}$ be a finite subset of R, then we formulate

<u>Problem C': Γ-Clustering:</u> Find the realizations L of M with component values closest (in the Eising-norm induced metric) to the set $\Gamma = \{\gamma_1,\ldots,\gamma_M\}$.

Problem C corresponds then to $\Gamma = \{0\}$. It is also helpful to define a bilinear map on $L_{m,n,p}$:

$$[[.,.]] = L_{m,n,p} \times L_{m,n,p} \dashrightarrow R^{n\times n}$$

$$[[(A,B,C),(F,G,H)]] = [A',F] - G\,B' + C'\,H$$

where [.,.] is the usual Lie product:

$$[A,F] = AF - FA$$

<u>Theorem 1:</u> The class of optimally clustered realizations coincides with the class for which [[L,L]] vanishes.

<u>Proof:</u> If (A,B,C) is constrained to realize M, then if (A_0,B_0,C_0) is a representant for M, we get a constrained optimization problem, for which the Hamiltonian is

$$H = \mathrm{Tr} \{ AA' + BB' + C'C + \Lambda_A\{TA_0-AT\}' + \Lambda_B\{TB_0-B\}' + \Lambda_C\{C_0-CT\} \}$$

The optimality conditions lead then directly to the stated result.

Not every system allows an optimally clustered realization. A counter-example can be found which is based on the phenomenon that the orbits under similarity are not closed [6].

With this solution, we can show the following:

<u>Theorem 2:</u> The realization of M_o which has maximal Eising distance to the orbit of M_1 is implicitly given by $L_* \in \pi^{-1}(M_o)$, satisfying

$$[[\Delta',L_*]] = 0$$
$$\Delta \text{ optimally clustered}$$

where Δ is the perturbation, $d_E(L^*,L_*) = L^*-L_*$, with $L^* \in \pi^{-1}(M_1)$. The maximal distance is then the Eising norm of Δ.

<u>Proof:</u> This follows easily by solving the minimax problem: First determine for a given realization L, the point $L^\#$ on the orbit of M_1 for which theEising distance $d_E(L,L^\#)$ is minimal. This problem has always a solution, but the realization $L^\#$ may not be unique. The proof is similar as in the clustering theorem. The condition is $[[L,L-L^\#]] = 0$. Next we slide L on its orbit, the associated realization $L^\#$ will

clearly vary with L, so that we may define a map $\#: \pi^{-1}(M_o) \dashrightarrow \pi^{-1}(M_1) : L \dashrightarrow L^{\#}$. Now find the similar realization L_*, for which $d_E(L_*,L^*)$ is maximal. The realization L^* is the corresponding point $(L_*)^{\#}$. This constrained optimization problem yields the additional condition $[[L-L^{\#},L-L^{\#}]] = 0$.

There remain some open problems. It is not clear whether or not the orbits can diverge, in the sense that the optimum may be on the closure of the orbits, and therefore not attainable. Also, if a solution exists, it may not be unique (modulo O(n)). The example of $\pi^{-1}(0,1,1)$ and $\pi^{-1}(1,1,1)$ illustrates that to every L in the first, a corresponding $L^{\#}$ on the other orbit exists, for which the distances are constant and equal to 1. It is also natural to look at the extension of problem A:

Problem A': Multimode Discrimination: Given the orbits $\pi^{-1}(L_i)$, find the realizations L_i^* on each orbit such that the set $\{L_i^*\}$ is maximally separated.

This is of interest for realizing multi mode systems. The practical significance of all this is that the realizations with the largest intraset distance are the most robust with respect to parameter inaccuracy, as for instance due to coefficient truncation. The problem will be discussed in section 3.4.

As to problem B, we shall just state the following results for the scalar observable $f: L_{m,n,p} \dashrightarrow R$, which in fact only gives an implicit solution to problem B:

Theorem 3: i) Let the realization L be given. The deviation $f(L+\Delta) - f(L)$, is extremal if the perturbation Δ is in the direction of the gradient of the observable f, evaluated at $L+\Delta$.

ii) If only the system M is given, i.e. the orbit $\pi^{-1}(L)$, then the deviation $f(L+\Delta) - f(L)$ is extremal at L_* if the perturbation Δ_*, grad $f(L_*+\Delta_*)$ and grad (L_*) are all aligned.

Proof: Again this follows simply from adjoining the constraints $\|\Delta\|^2 = 1$ and for part ii) also $f(L) = c$, with Lagrange multipliers λ (and μ for ii)) to the performance function $f(L+\Delta)-f(L)$. The optimality conditions are

$$\text{grad } f(\theta+\Delta) + \lambda\Delta = 0$$

$$\text{grad } f(\theta+\Delta) + (\mu-\lambda) \text{ grad } f(\theta) = 0.$$

Remark: The infinitesimal result of [4] is recovered for the uniform metric if the perturbation becomes infinitesimally small.

3. MULTI-MODE SYSTEMS

Two types of models with many similarities are discussed: Switched parameter linear systems and piecewise linear systems. Each "mode" of a system Σ_i will be denoted by a triple of systems functions (A_i,B_i,C_i). We shall also assume that the number of different modes is finite, N.

3.1 Switched Systems

The system is assumed to be modeled by

$$x_{k+1} = A_{[k]}x_k + B_{[k]}u_k$$

$$y_k = C_{[k]}x_k$$

where $[k] \in \{1,\dots,N\}$ is the function determining the mode switched on at time k. We assume that this switching is state-independent, but otherwise a purely deterministic sequence, or a random time series. A theory for deterministic discrete time periodic systems was developed in [7]. Here we allow general time variation, thus not necessarily periodically switching sequences. In the randomly switched case, we assume that the statistics are stationary and known. The domain of validity of each mode is the entire state space. This is in contrast with the next class.

3.2 Piecewise Linear Systems

This is a multi-mode system as described above, but the domains for the validity of each mode partitions the state space, i.e. they form a "patchwork" which pieces together a single "global" system. Clearly, one can think of such a system as a switched system with the switching completely determined by the state x_k of the system. Such a model results for instance in the approximation of a nonlinear system by a piecewise linear one [1]. Despite its local linearity, the dynamical behavior of such a system can be very complex and sustain chaotic motion [1].

3.3 Robust Design Problems for Multi-Mode Case

As usual, the problem is to design an optimal implementation of the multi-mode system. Characteristic for the multi-mode systems is the fact that the state is communicated from one mode to the other. This implies that despite the fact that each mode separately can be realized in many different ways, the total system is only left invariant under the action of $Gl_n(R)$, and not $Gl_n(R)^N$, where N is the number of modes. Hence a straightforward optimization by realizing each mode in the optimal way (i.e. essentially balanced) will only be valid if a state transformation is performed at each mode transition. We formalize this as

Theorem 4: Defining the observable for the multimode system as a weighted average of the observables in the single modes, the optimal unconstrained realization is obtained by realizing each mode individually in essentially balanced form.

Proof: The observable is

$$f = \text{Tr} \sum_i \pi_i \Lambda_i (H_i - O_i R_i)$$

which is a weighted sum of the observables defined in [4-5] for single mode systems. The parameterization is with respect to the components of the observability matrices O_i, and the reachability matrices R_i. The gradients are linear in these parameters, and the Hessian eigenproblem therefore decouples into the individual components, giving the simple condition $O_i'O_i = R_iR_i'$, expressing essential balancedness of all mode realizations.

If all modes communicate, this means that N(N-1) transformation matrices need to be stored. This additional computation induces also inaccuracies, and may therefore upset the optimality for that scheme. We present here the more direct approach by choosing the optimality criterion as a weighted version of the objective in each mode. The weights $\{\pi_i\}$ are most reasonably set equal to the relative time spent in each mode. From our assumptions these relative times are precomputable.

Theorem 5: Let the relative time spent in mode i be π_i, then the constrained minimal sensitivity realizations are given by the essentially balanced multi-mode systems defined by the requirement:

$$\sum_i \pi_i^2\, O_i'O_i = \sum_i \pi_i^2\, R_iR_i'$$

Proof: Follows directly from the EST, using the observable

$$f_i = \sum_i \pi_i \operatorname{Tr} \Lambda_i(H_i - O_iR)$$

with the constraints:

$$O_i = O_i^{o}T^{-1} \quad \text{and} \quad R_i = TR_i^{o}$$

where R_i^{o} and O_i^{o} are respectively the reachability and observability matrices for a nominal realization, and T is the $Gl_n(R)$ element to be determined, i.e. the parameterization for the problem. First the gradients of the observable with respect to R_i and O_i are computed, and the constraints are substituted. Noting that the time π_i spent in mode Σ_i does not depend on the realization of that mode, the gradient components are readily obtained,

$$\partial_{Oi} = \pi_i T R_i^{o} \Lambda_i$$

$$\partial_{Ri} = \pi_i \Lambda_i O_i^{o} T^{-1}$$

Minimization of the norm of the gradient with respect to T yields then the condition

$$TPT' = T^{-1}QT^{-1}$$

where we used the fact $\Lambda_i\Lambda_i' = \Lambda_i'\Lambda_i = I$, and defined the generalized gramians, weighted by the sojourn-times as

$$P = \sum_i \pi_i^2\, R_i^{o}R_i^{o\prime}$$

$$Q = \sum_i \pi_i^2\, O_i^{o\prime}O_i^{o}$$

The optimal T is then simply the balancing transformation for P and Q, which can always be found [8].

3.4 Non-Infinitesimal Perturbations of Multi-Mode Systems

It is also natural to look at the extension of problem A: Given the orbits $\pi^{-1}(L_i)$, find the realizations L_i^* on each orbit such that the set $\{L^*_i\}$ is maximally separated. The practical significance of all this is that the realizations with the largest intraset distance are the most robust with respect to parameter inaccuracy, as for instance due to coefficient truncation. We give a constructive solution of problem A' in the unconstrained case, i.e. when the states do not necessarily have to communicate directly, and transformations are allowable at each mode transition.

1. For each realization L on M_o, determine realizations L_i on the orbits of the other modes for which $d_E(L,L_i)$ is minimal. Let the minimal distance be $\Delta_L(M_i)$.

2. Determine $\Delta(L,\{M\}) = \min\{\, \Delta_L(M_i);\ i=1,\dots,N \,\}$. Note that this disance does no longer vary smoothly as L moves on M_o.

3. Determine $L^{\#}$ on M_o such that

   $\Delta(L^{\#},\{M\}) = \max\{\Delta(L,\{M\});\ L \text{ realizes } M_o\}$.

4. Perform steps 1-3 for each of the modes M_i, to find the optimal realizations $L_i^{\#}$.

Because of the nondifferentiable structure, the maximization in step 3 cannot be performed by simple differentiation. In the constrained problem, we start from the realizations $L_1,\dots,L_2$ in modes $M_1,\dots,M_2$ respectively, and solve for the transformation T such that (in the notation of the unconstrained problem)

$$\min\{\Delta(T(L_i),\{M\});\ i=1,\dots,N\}$$

is maximized over $T \in Gl_n(R)$. Many variants of the problem can be defined. For instance, the π_i-weighted average, rather than the minimum of the distances $\Delta(T(L_i),\{M\})$ may be maximized.

4. APPLICATION TO FILTERING

4.1 The Model

As discussed in the introduction, we shall assume that the nonlinear dynamics are satisfactorilly modeled by a piecewise linear multi model stocahstic system, thus a combination of the systems discussed in section 3. Each flight regime corresponds with one domain, and these domains are smoothly patched together. Moreover, in each flight regime, we also assume a multi--mode model because of the variable geometry. The system is assumed to be controlled, the control being conditioned on the observations, and therefore deterministic. This is superimposed on the stochastic inputs, modeling the noise in the system as well as the unpredictable component of the motion of the craft to be tracked.

The general discrete model is

$$x_{k+1} = A_{[k]}x_k + B_{[k]}d_k + Q_{[k]}^{\frac{1}{2}}u_k$$

$$y_k = C_{[k]}x_k + D_{[k]}e_k + R_{[k]}^{\frac{1}{2}}v_k$$

where u and v are white noise processes, modeling the measurement and dynamical uncertainties (e.g.

the unknown inputs due to the unpredictable motion of the craft to be tracked are typically modelled by colored noise, the noise shaping filter is then included in the dynamical equation). d is the deterministic input, which is a feedback of the filtered signal and some externally applied known component. Offsets (the biases due to an affine approximation of the nonlinearities) can be modeled in these terms as well.

4.2 The Steady State Filter

Under the above assumptions, a steady state Kalman filter approximation is implemented in each of the domains

$$\hat{x}_{k+1} = A_{[k]}\hat{x}_k + B_{[k]}d_k + K_{[k]}(y_k - C_{[k]}x_k - D_{[k]}e_k)$$

where the gains are computed for the steady state. This of course requires some assumptions on the deterministic signals d_k and e_k. Typically such a filter is used in a feedback scheme in order to provide the control command, i.e. we also have an "output" equation

$$d_k = r_k - M_{[k]}\hat{x}_k$$

which generates the control command. The combined equations are then

$$\hat{x}_{k+1} = (A_{[k]} - B_{[k]}M_{[k]} - K_{[k]}C_{[k]})\hat{x}_k + B_{[k]}r_k - K_{[k]}D_{[k]}e_k + K_{[k]}y_k$$

i.e, a multi-mode system

$$\hat{x}_{k+1} = F_{[k]}\hat{x}_k + G_{[k]}w_k$$

$$d_k = H_{[k]}\hat{x}_k + r_k$$

with the modes defined by

$$F_i = A_i - B_iM_i - K_iC_i$$

$$G_i = [\ B_i\ ;\ -K_iD_i\ ;\ K_i\ ]$$

$$H_i = -M_i$$

and the input w_k is $[r_k', e_k', y_k']'$.

4.3 The Optimal Implementation

The equations for the filter modes obtained in the previous subsection are of the form of the multi-mode systems in section 3. The results obtained there apply therefore directly. In particular, the sojourn-times $\{\pi_i\}$ can be estimated, either via simulation on the exact dynamical (nonlinear) system, or in the simpler cases, by direct analysis. The Gramians $Q_i = O_i'O_i$ and $P_i = R_iR_i'$ can be computed by solving the Lyapunov equations

$$F_iP_iF'_i + G_iG'_i = P_i$$

$$F'_iQ_iF_i + H'_iH_i = Q_i$$

The transformation to the minimal sensitivity coordinate basis is then obtained by balancing [7] the matrices

$$P = \Sigma\ \pi_i^2\ P_i \qquad \text{and} \qquad Q = \Sigma\ \pi_i^2\ Q_i$$

Finally, each of the modes of the filter is then transformed to the optimal form. We have worked out the ideas for discrete time filters. The concept works just as well in continuous time [5]. The conditions are the same (i.e. essential balancedness of the averaged system).

4.4 Suboptimal Implementations

While the optimal implementation is based on the steady-state filters described in section 4.2, the applicability of these filters is not appropriate for the stochastic transition case when the transitions are not known to the observer. The reasons for this is that the filters were derived for the steady-state case, which assumes long sojourn times for such a steady-state to be achieved. The results also assume that the mode of the system may be known so that the appropriate filter can be selected. Such an assumption can be justified for long sojourn times that allow mode identification. Finally, when fast transitions among the modes can occur, a steady-state will also never be achieved for any of these states.

In this case a suboptimal choice of the filters is considered based on the size of π_i. A small parameter δ is selected, and the modes are classified into two types: those with $\pi_i > \delta$ (slow modes) and those with $\pi_i < \delta$ (fast modes). A set of filters as shown in section 4.2 is designed to run in parallel for all slow modes, and the correct one is chosen based on a likelihood function that is based on these steady-state filters (the transitions are ignored to avoid the exponential rise in complexity). Two alternatives are considered for the fast mode filters. The first is to define an average model just based on these modes and their sojourn times π_i, and use this to obtain a filter for these modes with corresponding aggregated likelihood function. The second is to use a separate filter for each mode and use a weighted average using the lilelihood functions of each mode.

In these cases the filters will be as given in section 4.2 and the likelihood function is obtained by standard expressions, which may be modified as in [2] when the multi-models result from nonlinear approximations. In that case the regions in the state space can be approximately found from the estimates and combined with the statistical expressions.

4.5 Example: Two-Dimensional Intercept Problem

In this two-domensional example the states represent the relative positions of the missile and the target, which may be given as

$$\dot{x} = V_x$$

$$\dot{y} = V_y$$

$$\dot{V}_x = d_1 - a_1$$

$$\dot{V}_y = d_2 - a_2$$

where x and y are the relative positions in the (x,y) plane, d_i are the control forces of the missile which will be based on the estimates of the states assuming separation holds, and a_i are the maneuvering acceleration of the target which may be modeled by a first-order Markov process

$$\dot{a}_i = -\mu_i a_i + w_i, \qquad i = 1,2.$$

The objective may be formulated as a quadratic control problem except that the observation are nonlinear in the states, namely

$$z_1 = (x^2 + y^2)^{\frac{1}{2}} + v_1 = r + v_1$$

$$z_2 = \tan^{-1} (y/x) + v_2 = \theta + v_2$$

where v is modelled as a white noise with covariance dependent on the relative distance.

The model may be approximated in two possible ways. The first is to define new states involving the angle and the range and these will lead to a nonlinear model for the state equations. This model is then approximated by a piecewise affine multi-mode set of equations. The second is to approximate the observation functions by piecewise affine multi-mode system with linear state model. If we define the set points for the approximations as r_i and θ_i then the observations will be given approximately by

$$z_1 = r_i + \cos \theta_i (x-x_i) + \sin \theta_i (y-y_i) + v_1$$

$$z_2 = \theta_i + (1/r_i) (\cos \theta_i (y-y_i) - \sin \theta_i (x-x_i)) + v_2.$$

The multi-mode filter is then derived using the expressions of section 4.2. The control command may also be incorporated in the design using either given control strategy such as proportional navigation or suboptimal implementation of an optimal quadratic cost state feedback control law. The result can be evaluated using simulation of the system under several engagement scenarios.

5. CONCLUSIONS

The optimal sensitivity properties for the multi-mode realizations have been derived. They extend nicely the notions of Essentially Balanced Realizations derived in [4-5]. These optimal realizations have been applied to obtain an optimal implementation of a simple multi-mode filter, which allows the tracking of a target with low-complexity, small wordlength hardware. This simple multi-mode model can be justified if the sampling rate is sufficiently high. More quantitative results are presently under investigation based on a simple two-dimensional tracking example.

We have restricted our discussion to square systems (m=p) and minimal realizations. Extensions of the theory are in progress. It seems intuitively clear that one could further exploit the redundancy of a realization by deliberately using nonminimal realizations. A heuristic argument for this possibility is as follows: Let N = kn, and let (A,B,C) be a minimal sensitivity minimal realization of a system H. Construct now k different realizations

$$\{A_i,B_i,C_i\} = \{T_iAT_i^{-1},T_iB,CT_i^{-1};\ i=1,\dots,k\}.$$

With these realizations construct the nonminimal diagonal realization of order N

$$\hat{A} = \text{diag}\ (A_i)\ ;\ \hat{B} = \text{vec}(B_i)\ ;\ \hat{C} = \text{vec}(C_i)/k$$

If the T_i's are chosen in a neighborhood of the identity, such that the rounding errors in each component system are independent, then as $k \dashrightarrow \infty$

$$\|\Delta y_k\|^2 \le \|\Delta C_i + C_i \Delta A_i\|^2_{max}/k\ \|x^{(k)}\|^2 \dashrightarrow 0$$

It is possible to overparameteize the system in order to obtain minimal sensitivity realizations. Finally, the idea in the proof of the main sensitivity theorem leads to gradient type algorithms for the optimal sensitivity realizations. This of course is to be performed off line, during the design stage, and poses therefore no restrictions on the hardware. Preliminary remarks regarding these appear in [5].

REFERENCES

[1] E. I. Verriest and A. H. Haddad, "Linear Markov Approximations of Piecewise Linear Stochastic Systems", Stochastic Analysis and Applications, vol. 5, pp. 213-244, 1987.

[2] A. H. Haddad, E. I. Verriest, and P. D. West, "Approximate Nonlinear Filtering for Piecewise Linear Systems," NATO/AGARD Guidance and Control Panel's 44th Symposium, Athens, Greece, 5-8 May 1987.

[3] E. I. Verriest and A. H. Haddad, "Approximate Nonlinear Filters for Piecewise Linear Models", Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 526-529, March 1986.

[4] E. I. Verriest and W. S. Gray, "Robust Design Problems: A Geometric Approach", in Mathematical Theory in Networks and Systems, Martin and Byrnes, eds., North-Holland 1988.

[5] W. S. Gray and E. I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, pp. 124-128, December 1987.

[6] A. S. Khadr and C. Martin, "On the GLn(R) Action on Linear Systems: The Orbit Closure Problem", in Algebraic and Geometric Methods in Linear System Theory, Byrnes and Martin, eds., Lectures in Applied Mathematics, Vol. 18, Springer-Verlag 1980.

[7] E. I. Verriest, "Alternating Discrete Time Systems: Invariants, Parametrization and Realization", Proc. Annual Conference on Information Systems and Sciences, Princeton University, March 1988.

[8] E. I. Verriest and T. Kailath, "On Generalized Balanced Realizations", IEEE Trans. Automatic Control, Vol. AC-28, pp. 833-844, August 1983

APPENDIX P

E. I. Verriest, "Minimum Sensitivity Implementations for Multi-Mode Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 2165-2170, Dec. 1988.

MINIMUM SENSITIVITY IMPLEMENTATIONS FOR MULTI-MODE SYSTEMS

Erik I. Verriest

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

ABSTRACT

This paper addresses some aspects of the design and implementation of multi-mode systems, under finite precision restrictions. This occurs, for instance, when quantization and finite wordlength effects need to be incorporated, or when high component tolerances in analog designs need to be considered. The freedom in the design is exploited in order to obtain the realizations closest to the normal or desired behavior, despite the interference of quantization, component tolerances (analog case) and finite wordlength (discrete case). Our interest is in the mathematical characterization of this new type of robustness problem, and its solutions under various criteria of optimality. Earlier results, linking these optimal realizations for linear time-invariant systems to the Balanced Realizations are here extended to the multi-mode and general time-varying systems. The feasibility of this approach in multi-mode stochastic problems is shown.

1. INTRODUCTION

This work builds on our previous work on robust design problems for single-mode time-invariant systems [4,5]. Consider a linear time invariant system (A,B,C) with m inputs, p outputs, and McMillan degree n. This may be a model of a system to a faithfully simulated, the implementation of an analog or discrete filter, or an observer/controller implementing an optimal regulator for a given plant. As in all these applications, not the actual state coordinates, but the input-output relation is important, they are usually implemented by a so-called canonical form. Reason for this is that these implementations minimize the number of parameters, and allow a pipelined realization of the devices, e.g. the "Direct Form" realizations in digital signal processing. A minimal number of parameters corresponds to minimal complexity, an important quality if operation count becomes important. However, a minimal set of parameters has no redundancy, and therefore, one may expect high sensitivity with respect to these parameters. It is clear that the freedom of coordinate basis of the implementation should be utilized to determine optimal realizations under various criteria for optimality. In particular, two issues seem to be important: sensitivity and clustering. The sensitivity requirement guarantees robustness of the actual implementation, while clustering deals with the parameter ranges. It relates to the problem of approximately implementing a certain system with parameter values chosen from a finite set.

These results will be extended here to multi-mode systems. These systems have recently become of interest as models for multi-rate systems [10], nonuniformly sampled continuous systems and as approximations to nonlinear systems [2]. Several interesting aspects (e.g. reachability) have been studied for these systems, and connections with other fields of study (Yang-Mills Theory) have recently been made [9]. The next section poses the main design problems in the geometric framework. Emphasis is here given to the noninfinitesimal perturbations, the results on infinitesimal perturbations being presented earlier [4,5]. The interest is, of course, to problems involving quantization and finite wordlength effects in digital data processing, where a random variable approach to the problem may lead to unsuccessful modeling of its behavior, as for instance illustrated in [11]. Section 3 then goes on with the application of the geometric theory to multi-mode systems. Two types are discussed in detail: the switched systems and the piecewise linear systems. The general time-varying case is presented in Section 4. In Section 5, implementations of approximate filters are discussed, based on a piecewise linear approximation of the nonlinearity, and their optimal implementation as discussed in Section 3.

2. A GEOMETRIC APPROACH TO OPTIMAL IMPLEMENTATIONS: NONINFINITESIMAL THEORY

The approach taken in [4] is geometric. The realization space $L_{m,n,p}$ is modeled as an n(n+m+p) dimensional affine space with an Euclidean metric defined in the tangent space at each point. The Extremal Sensitivity Theorem [4] asserts that the minimum sensitivity points of an observable (a smooth map from $L_{m,n,p}$ to $\mathbb{R}$) are the points where the (generalized) gradient is in the eigenspace of the (generalized) Hessian. In the case of fixed point implementations, the uniform Euclidean metric is appropriate, and the gradient and Hessian correspond with the usual notions in calculus. All results are, therefore, also "infinitesimal." One can reasonably argue that in finite wordlength arithmetic, the notions of infinitesimal perturbations of the coefficients are meaningless. This may indeed invalidate the application to finite wordlength effects. For this reason, we shall develop the analysis for noninfinitesimal perturbations in this paper. It will be shown that this approach makes a connection with the notion of clustering. In this paper, by finite we shall mean noninfinitesimal. By L we denote a particular realization in $L_{m,n,p}$, and by M the equivalence class of all similar realizations having a particular input-output behavior, i.e. the orbit $\pi^{-1}(M)$ under $GL_n(\mathbb{R})$. The equivalence class will be referred to as "system," the orbit space is denoted by $M_{m,n,p}$, and the projection map from $L_{m,n,p}$ to $M_{m,n,p}$ by π. We study two problems, which are related to the sensitivity and robustness.

Problem A: Discrimination. How do we choose a realization of a given system M_o such that it is maximally distant from the orbit $\pi^{-1}(M)$, where M is another given system?

Problem B: Worst Case Defects. If the system parameters are perturbed over a fixed noninfinitesimal amount, and if f is a scalar system function (i.e.

invariant under similarity), then find the realizations L of M_o for which the f-perturbation

$$\max_{C_\Delta} \{f(L+\Delta) - f(L)\} \qquad \text{where } C_\Delta = \{\Delta | \, \|\Delta\| = 1\}$$

is minimal.

In both problems, the norm function in the realization space will be fixed to be the Eising norm (compatible with the uniform metric).

$$\|L\|^2 = \|(A,B,C)\|^2 = \mathrm{Tr}\{AA' + BB' + C'C\}$$

It follows then that if (A,B,C) is a solution to each of the above problems, then also any realization obtained from (A,B,C) by an orthogonal similarity transformation is also a solution. Before outlining the solution to the above problems, we shall introduce a third one, to which problem A is related:

Problem C: Clustering. Find the realizations L of M with minimal system (Eising) norm.

This has the physical significance that cooperatively the components of the realization are as small as possible, hence clustered near zero. It is a special case of the following more general (and more significant) clustering problem. Let $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$ be a finite subset of R, then we formulate:

Problem C': Γ-Clustering. Find the realizations L of M with component values closest (in the Eising norm induced metric) to the set $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$.

Problem C corresponds then to $\Gamma = \{0\}$. It is also helpful to define a bilinear map on $L_{m,n,p}$:

$$[[.,.]] : L_{m,n,p} \times L_{m,n,p} \longrightarrow R^{n \times n}$$

$$[[(A,B,C),(F,G,H)]] = [A',F] - GB' + C'H$$

where [.,.] is the usual Lie product: $[A,F] = AF - FA$.

Theorem 1. The class of optimality clustered realizations coincides with the class for which [[L,L]] vanishes.

Proof. If (A,B,C) is constrained to realize M, then if (A_0,B_0,C_0) is a representative for M, we get a constrained optimization problem, for which the Hamiltonian is

$$H = \mathrm{Tr}\{AA' + BB' + C'C + \Lambda_A\{TA_0 - AT\}' + \Lambda_B\{TB_0 - B\}' + \Lambda_C\{C_0 - CT\}\}$$

The optimality conditions lead then directly to the stated result. ●

Not every system allows an optimally clustered realization. A counter-example is the system

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad C = [1 \;\; 1]$$

Converging sequences of transformations can be found which yield equivalent systems with decreasing Eising norm, but the limit realization is not a point on the orbit of the given realization. It is a known phenomenon that the orbits under similarity are not closed [6].

Remark. A realization is optimally clustered iff there exists an orthogonal transformation $S \in O(n+m)$, such that $[A,B]S = [A',C']$.

With this solution, we can show the following:

Theorem 2. The realization of M_o which has maximal Eising distance to the orbit of M_1 is implicitly given by $L_* \in \pi^{-1}(M_0)$, satisfying

$$[[\Delta', L_*]] = 0$$

$$\Delta \text{ optimally clustered}$$

where Δ_1 is the perturbation, $d_E(L^*,L_*) = \|L^*-L_*\|$, with $L^* \in \pi^{-1}(M_1)$. The maximal distance is then the Eising norm of Δ.

Proof. This follows easily by solving the minimax problem: First determine for a given realization L of M_o, the point $L^\#$ on the orbit $\pi^{-1}(M_1)$ for which the Eising distance $d_E(L,L^\#)$ is minimal. This problem has always a solution, but the realization $L^\#$ may not be unique. The proof is similar as in the clustering theorem. The condition is $[[L,L-L^\#]] = 0$. Next we slide L on its orbit, the associated realization $L^\#$ will clearly vary with L, so that we may define a map $\#: \pi^{-1}(M_0) \longrightarrow \pi^{-1}(M_1) : L \longrightarrow L^\#$. Now find the similar realization L_*, for which $d_E(L_*,L^*)$ is maximal. The realization L^* is the corresponding point $(L_*)^\#$. This constrained optimization problem yields the additional condition $[[L-L^\#,L-L^\#]] = 0$. ●

There remain some open problems. It is not clear whether or not the orbits can diverge, in the sense that the optimum may be on the closure of the orbits, and therefore not attainable. Also, if a solution exists, it may not be unique (modulo O(n)). The example of $\pi^{-1}(0,1,1)$ and $\pi^{-1}(1,1,1)$ illustrates that to every L in the first, a corresponding $L^\#$ on the other orbit exists for which the distances are constant and equal to 1. It is also natural to look at the extension of problem A.

Problem A': Multi-Mode Discrimination. Given the orbits $\pi^{-1}(L_i)$, find the realizations L_i^* on each orbit such that the set $\{L_i^*\}$ is maximally separated.

This is of interest for realizing multi-mode systems. The practical significance of this all is that the realizations with the largest intraset distance are the most robust with respect to parameter inaccuracy, as for instance due to coefficient truncation. The problem will be discussed in Section 3.4.

As to Problem B, we shall just state the following results for the scalar observable $f: L_{m,n,p} \longrightarrow \mathbb{R}$, which in fact only gives an implicit solution to Problem B:

Theorem 3.

(1) Let the realization L be given. The deviation $f(L+\Delta) - f(L)$, is extremal if the perturbation Δ is in the direction of the gradient of the observable f, evaluated at L+Δ.

(2) If only the system M is given, i.e. the orbit $\pi^{-1}(M)$, then the deviation $f(L+\Delta) - f(L)$ is extremal at L_* if the perturbation Δ_*, grad $f(L_*+\Delta_*)$ and grad (L_*) are all aligned.

Proof. Again this follows simply from adjoining the constraints $\|\Delta\|^2 = 1$ and for part (2) also $f(L) = \text{cst}$, with Lagrange multipliers λ (and μ for (2)) to the performance function $f(L+\Delta) - f(L)$. The optimality conditions obtained by nulling the partials, with respect to Δ and L, are

$$\operatorname{grad} f(\theta+\Delta) + \lambda\Delta = 0$$

$$\operatorname{grad} f(\theta+\Delta) + (\mu-1) \operatorname{grad} f(\theta) = 0 \qquad \bullet$$

Remark. The infinitesimal result of [4] is recovered for the uniform metric if the perturbation becomes infinitesimally small.

3. MULTI-MODE SYSTEMS

A multi-mode system is in effect a time-variant system. However, the term will be used to designate the particular case where the time spent by the system in each mode is significantly longer than the dynamical characteristic times (e.g. time constants and oscillation periods), in each mode. Heuristically speaking, a multi-mode system behaves locally (in a temporal sense) like a time-invariant system. Each "mode" of a system Σ_i will be denoted by a triple (A_i, B_i, C_i) of nxn, nxm, and pxn matrices, respectively. We shall also assume that the number of different modes is finite, N say. (Although the theoretical development remains valid with a countable set of modes, its justification in finite time data processing is elusive.)

This is in contrast to fast switching (Section 4). In this case it is well known that, for instance, the stability properties are not directly determined by the individual dynamical interests A_i. Two types of models, with many similarities are discussed: switched systems and piecewise linear systems.

3.1 Switched Systems

The system is assumed to be modeled by

$$x_{k+1} = A_{[k]}x_k + B_{[k]}u_k$$

$$y_k \quad = C_{[k]}x_k$$

where $[k] \in \{1,\ldots,N\}$ is the function determining the mode switched on at time k. We assume that this switching is state-independent, but otherwise, can be a purely deterministic sequence, or a random time series. A theory for deterministic discrete time periodic systems was developed in [7]. Here we allow general time variation, thus not necessarily periodically switching sequences. In the randomly switched case, we assume that the statistics are stationary and known. The domain of validity of each mode is all of R^n, i.e. the whole state space. This is in contrast with the next class.

3.2 Piecewise Linear Systems

This is a multi-mode system as described above, but the domains for the validity of each mode partition the state space, i.e. they form a "patchwork" which pieces together a single "global" system. Clearly, one can think of such a system as a switched system with the switching completely determined by the state x_k of the system. Such a model results, for instance, in the approximation of a nonlinear system by a piecewise linear one [1]. Despite its local linearity, the dynamical behavior of such a system can be very complex and sustain chaotic motion Details are presented in [3].

3.3 Robust Implementation

The problem is to design an optimal implementation of the multi-mode system. As a first approximation to the optimal realization, each mode can be realized in minimal sensitivity form as presented in [4]. This is justified by the following argument. Let H_i be the pxp block Hankel matrix of the Markov parameters in mode i, and consider the Hankel matrix formed by the 2k-1 consecutive samples of the pulse response of the multi-mode system, given that the input pulse occurred while the system was in mode i. If this initial pulse time is uniformly distributed in the interval $[0,T_i]$, where T_i is the duty time of mode i, then the two Hankel matrices will be equal with "probability" T_i-2k+1/T_i as long as 2k is less than T_i. Note that the above probability converges to 1 if k/T_i decreases. In this case, a weighted average of the observables corresponding to each mode is justified, with the weights proportional to the duty cycles, or expected sejourn times of the system in the respective modes. This yields then at once the extension of the sensitivity theorem in [4]:

Theorem 4. Defining the observable for the multi-mode system as a weighted average of the observables in the single modes, the optimal unconstrained realization is obtained by realizing each mode individually in essentially balanced form.

Proof. The observable is

$$f = \operatorname{Tr} \sum_i \pi_i \Lambda_i (H_i - O_i R_i)$$

which is a weighted sum of the observables defined in [4,5] for single-mode systems. The parameterization is with respect to the components of the observability matrices O_i, and the reachability matrices R_i. The gradients are linear in these parameters, and the Hessian eigenproblem, therefore, decouples into the individual components, giving the simple condition

$$O_i'O_i = R_i R_i' ,$$

expressing essential balancedness of all mode realizations. $\bullet$

There is one problem with the above approach. By individually optimizing each mode, there will be no common base for the state spaces in each mode. This means that if at time T a mode switching occurs from i to j, the state existing at time T in mode i, x_T, cannot be directly used as "initial condition" (at time T) for the system in mode j, but needs to be transformed first to the proper coordinates. If all modes communicate, (i.e. if all mode transitions are present or possible), it means that N(N-1) transformation matrices need to be stored. This set consists of pairs of mutual inverses. If only cyclic shifts occur, N transformations suffice. Besides this required overhead in memory, the additional computations induce also inaccuracies, and may therefore upset the optimality for that scheme.

A direct approach exists by solving the problem with the same optimality criterion as in Theorem 4 (i.e. a weighted version of the objective in each mode), but with the additional constraint that the state is communicated from one mode to the other. This implies that despite the fact that each mode separately can be realized in many different ways, the total system is only left invariant under the action of $Gl_n(R)$, and not $Gl_n(R)^N$, where N is the number of modes. The weights $\{\pi_i\}$ are again set equal to our assumptions these relative times are precomputable.

Theorem 5. Let the relative time spent in mode i be π_i, then the constrained minimal sensitivity realizations are given by the essentially balanced multi-mode systems defined by the requirement:

$$\sum_i \pi_i^2 O_i'O_i = \sum_i \pi_i^2 R_i R_i' .$$

Proof. Follows directly from the EST, using the observable

$$f = \sum_i \pi_i \operatorname{Tr}\Lambda_i (H_i - O_i R_i)$$

with the constraints:

$$O_i = O_i^o T^{-1}$$

$$R_i = TR_i^o$$

where R_i^o and O_i^o are, respectively, the reachability and observability matrices for a nominal realization, and T is the $GL_n(R)$ element to be determined, i.e. the parameterization for the problem. First, the gradients of the observable with respect to R_i and O_i are computed, and the constraints are substituted. Noting that the time π_i spent in mode Σ_i does not depend on the realization of the mode, the gradient components are readily obtained,

$$\partial_{Oi} = \pi_i TR_i^o \Lambda_i$$

$$\partial_{Ri} = \pi_i \Lambda_i O_i^o T^{-1}$$

Minimization of the gradient norm with respect to T yields then the condition

$$TPT' = T^{-T}QT^{-1}$$

where we used the fact $\Lambda_i \Lambda_i' = \Lambda_i' \Lambda_i = I$, and defined the generalized gramians, weighted by the sejourn-times as

$$P = \sum_i \pi_i^2 R_i^o R_i^o ,$$

$$Q = \sum_i \pi_i^2 O_i^o{}' O_i^o$$

The optimal T is then simply the balancing transformation for P and Q, which can always be found [8]. ●

3.4 Noninfinitesimal Perturbations of Multi-Mode Systems

It is also natural to look at the extension of Problem A: Given the orbits $\pi^{-1}(M_i)$, find the realizations L_i^* on each orbit such that the set (L_i^*) is maximally separated. The practical significance of this all is that the realizations with the largest intraset distance are the most robust with respect to parameter inaccuracy, as for instance due to coefficient truncation. We give a constructive solution of Problem A' in the unconstrained case, i.e. when the states do not necessarily have to communicate directly, and transformations are allowable at each mode transition:

(1) For each realization L on M_o, determine realizations L_i on the orbits of the other modes for which $d_E(L,L_i)$ is minimal. Let the minimal distance be $\Delta_L(M_i)$.

(2) Determine $\Delta(L,\{M\}) = \min\{\Delta_L(M_i) ; i = 1,\ldots,N\}$. Note that this distance does no longer vary smoothly as L moves on M_o (it is not differentiable at the crossovers).

(3) Determine $L^{\#}$ on M_0 such that

$$\Delta(L^{\#},\{M\}) = \max\{\Delta(L,\{M\}) ; L \text{ realizes } M_o\} .$$

(4) Perform Steps 1 to 3 for each of the modes M_i, to find the optimal realizations $L_i^{\#}$.

Because of the nondifferentiable structure, the maximization in Step 3 cannot be performed by simple differentiation. In the constrained problem, we start from the realizations $L_1,\ldots,L_2$ in modes $M_1,\ldots,M_2$, respectively, and solve for the transformation T such that (in the notation of the unconstrained problem)

$$\min\{\Delta(T(L_i),\{M\}) ; i = 1,\ldots,N\}$$

is maximized over $T \in Gl_n(R)$. Many variants of the problem can be defined. For instance, the π_i-weighted average, rather than the minimum of the distances $\Delta(T(L_i),\{M\})$ may be maximized.

4. GENERAL TIME-VARYING SYSTEMS

Consider now the general time-varying case, with realization $\{(A_i,B_i,C_i) ; i \varepsilon Z\}$. The dynamics are completely specified by the response matrices H(k), whose ij-elements are the response at time k+i to an impulse at time k+j-1

$$H_{ij}(k) = C_{k+i}A_{k+i-1}A_{k+i-2}\cdots A_{k+j+1}A_{k+j}B_{k+j-1}$$

Just as in [4], the optimal implementation will be determined by the optimal factorization of H(k) into a local (at time k) observability and reachability matrix, O(k) and R(k), for the time-varying system; i.e.

$$H(k) = O(k)R(k) = \begin{bmatrix} C_{k+1} \\ C_{k+2}A_{k+1} \\ \cdot\cdot \\ \cdot\cdot \end{bmatrix} [B_k, A_k B_{k-1}, \ldots]$$

Defining the observable f_k as $Tr\Lambda_k(O(k)R(k)-H(k))$, the only difference with the development in [4] lies in the interpretation (i.e. it is now the local observable for a time-varying system). The mathematics carry through in a straightforward manner. The extremal sensitivity realization is then again determined from the criterion

$$R(k)R'(k) = O'(k)O(k)$$

which means that the realization must be locally (at k) essentially balanced. This result leads directly to:

Theorem 6. A time-varying system has a realization of minimal sensitivity which is essentially balanced in the time-varying sense [8].

Proof. Define as the system observable a uniformly weighted sum of the local observables defined above, i.e.

$$f = \lim_{N\to\infty} \frac{1}{2N-1} \sum_{i=-N}^{N} Tr\Lambda_i(H(i)-O(i)R(i))$$

The parameterization of the realizations is with respect to the local observability and reachability matrices. By Theorem 4, the optimality conditions are

$$O(i)'O(i) = R(i)R'(i)$$

i.e. equality of the local observability and reachability gramians. These local gramians can be simultaneously diagonalized by an orthogonal transformation. The corresponding realization is the time variant analog of the balanced realization, and its existence is proven in [8]. As the condition only expresses equality and not diagonality of the gramians, the extended notion of "essential balancedness," i.e. the balancedness up to an orthogonal (now time-varying) transformation is again sufficient. ●

5. APPLICATION TO NONLINEAR STOCHASTIC CONTROL

5.1 The Model

As discussed in the introduction, we shall assume that the nonlinear dynamics are satisfactorily modeled by a piecewise linear multi-model stochastic system [2], thus a combination of the systems discussed in Section 3. The system is assumed to be controlled, the control being conditioned on the observations, and therefore deterministic. This is superimposed on the stochastic inputs, modeling the noise in the system as well as the unpredictable components of the inputs, due to coupling with unmodeled dynamics.

The general discrete model is

$$x_{k+1} = A_{[k]}x_k + B_{[k]}d_k + Q^{1/2}_{[k]}u_k$$

$$y_k = C_{[k]}x_k + D_{[k]}e_k + R^{1/2}_{[k]}v_k$$

where u and v are white noise processes, modeling the measurement and dynamical uncertainties (e.g. the unknown inputs due to unmodeled dynamics are typically modeled by colored noise, the noise shaping filter is then included in the dynamical equation). d is the deterministic input, which is a feedback of the filtered signal and some externally applied known component. Offsets (the biases due to an affine approximation of the nonlinearities) can be modeled in these terms, d and e, as well.

5.2 The Steady State Filter

Under the above assumptions, a steady state Kalman filter approximation is implemented in each of the domains

$$\hat{x}_{k+1} = A_{[k]}\hat{x}_k + B_{[k]}d_k + K\left(y_k - C_{[k]}\hat{x}_k - D_{[k]}e_k\right)$$

where the gains are computed for the steady state. This of course requires some assumptions on the deterministic signals d_k and e_k. Typically such a filter is used in a feedback scheme in order to provide the control command, i.e. we also have an "output" equation

$$d_k = r_k - M_{[k]}\hat{x}_k$$

which generates the command control. The combined equatins are then

$$\hat{x}_{k+1} = (A_{[k]} - B_{[k]}M_{[k]} - K_{[k]}C_{[k]})\hat{x}_k$$
$$+ B_{[k]}r_k - K_{[k]}D_{[k]}e_k + K_{[k]}y_k$$

i.e. a multi-mode system

$$\hat{x}_{k+1} = F_{[k]}\hat{x}_k + G_{[k]}w_k$$
$$d_k = H_{[k]}\hat{x}_k + r_k$$

with the modes defined by

$$F_i = A_i - B_iM_i - K_iC_i$$
$$G_i = [B_i \; ; \; -K_iD_i \; ; \; K_i]$$
$$H_i = -M_i$$

and the input w_k is $[r'_k, e'_k, y'_k]'$.

5.3 The Optimal Implementation

The equations for the filter modes obtained in the previous subsection are of the form of the multi-mode systems in Section 3. The results obtained there apply, therefore, directly. In particular, the sejourn-times $\{\pi_i\}$ can be estimated, either via simulation on the exact dynamical (nonlinear) system, or in the simple cases, by direct analysis. The gramians $Q_i = O'_iO_i$ and $P_i = R_iR'_i$ can be computed by solving the Lyapunov equations

$$F_iP_iF'_i = G_iG'_i = P_i$$
$$F'_iQ_iF_i + H'_iH_i = Q_i$$

The transformation to the minimal sensitivity coordinate basis is then obtained by balancing [7] the matrices

$$P = \Sigma\, \pi_i^2 P_i \qquad \text{and} \qquad Q = \Sigma\, \pi_i^2 Q_i$$

Finally, each of the modes of the filter is then transformed to the optimal form. We have worked out the ideas for discrete time filters. The concept works just as well in continuous time [5]. The conditions are the same (i.e. essential balancedness of the averaged system).

6. CONCLUSIONS

The optimal sensitivity properties for the multi-mode realizations have been derived. They extend nicely the notions of essentially balanced realizations derived in [4,5]. These optimal realizations have been applied to obtain an optimal implementation of a simple multi-mode filter, which allows the tracking of a target with low complexity, small wordlength hardware. This simple multi-mode model can be justified if the sampling rate is sufficiently high. More quantitative results are presently under investigation.

We have restricted our discussion to square systems (m = p) and minimal realizations. Extensions are straightforward. It seems intuitively clear that one could further exploit the redundancy of a realization by deliberately using nonminimal realizations. A heuristic argument for this possibility is as follows: Let N = kn, and let (A,B,C) be a minimal sensitivity minimal realization of a system H. Construct now k different realizations $\{A_i, B_i, C_i\} = \{T_iAT_i^{-1}, T_iB, CT_i^{-1} ;\ i = 1,\ldots,k\}$. With these realizations construct the nonminimal diagonal realization of order N

$$\hat{A} = \text{diag }(A_i) \; ; \; \hat{B} = \text{vec }(B_i) \; ; \; \hat{C} = \text{vec }(C_i)/k$$

If the T_i's are chosen in a neighborhood of the identity, such that the rounding errors in each component system are independent, then as k ⟶ ∞

$$\|\Delta y_k\|^2 < \|\Delta C_i + C_i\Delta A_i\|^2_{\max}/k\|x^{(k)}\|^2 \longrightarrow 0$$

It is possible to overparameterize the system in order to obtain minimal sensitivity realizations. Finally, the idea in the proof of the main sensitivity theorem leads to gradient type algorithms for the optimal sensitivity realizations. This, of course, is to be performed off line, during the design state, and poses, therefore, no restrictions on the hardware. Some preliminary remarks regarding these appear in [5].

ACKNOWLEDGEMENT

This research is supported by the U.S. Air Force under Contract No. AFOSR-87-0308.

REFERENCES

[1] E.I. Verriest and A.H. Haddad, "Approximate Nonlinear Filters for Piecewise Linear Models," Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 526-529, March 1986.

[2] L.O. Chua and A.C. Deng, "Canonical Piecewise-Linear Modeling," IEEE Trans. Circuits and Systems, vol. CAS-33, no. 5, pp. 511-525, May 1986.

[3] E.I. Verriest and A.H. Haddad, "Linear Markov Approximations of Piecewise Linear Stochastic Systems," Stochastic Analysis and Applications, vol. 5, no. 2, pp. 213-244, 1987.

[4] E.I. Verriest and W.S. Gray, "Robust Design Problems: A Geometric Approach," to appear in Mathematical Theory in Networks and Systems, Martin and Byrnes, eds., North-Holland, 1988.

[5] W.S. Gray and E.I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity," Proc. 26th Conference on Decision and Control, Los Angeles, CA, December 1987.

[6] A.S. Khadr and C. Martin, "On the GLn(R) Action on Linear Systems: The Orbit Closure Problem," in Algebraic and Geometric Methods in Linear System Theory, Byrnes and Martin, eds., Lectures in Applied Mathematics, vol. 18, Springer-Verlag, 1980.

[7] E.I. Verriest, "Alternating Discrete Time Systems: Invariants, Parameterization and Realization," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

[8] E.I. Verriest and T. Kailath, "On Generalized Balanced Realizations," IEEE Transactions on Automatic Control, vol. AC-28, no. 8, pp. 833-844, August 1983.

[9] U. Helmke, "Parameterizations for Multi-Mode Systems and Yang-Mills Instantons," Proc. 25th Conference on Decision and Control, Athens, Greece, December 1986.

[10] D.P. Stanford and L.T. Conner, "Controllability and Stabilizability in Multi-Pair Systems," SIAM J. Control and Optimization, vol. 18, no. 5, pp. 488-497, September 1980.

[11] D.F. Delchamps, "New Techniques for Analyzing the Effects of Output Quantization in Feedback Systems," Proc. Annual Conference on Information Sciences and Systems, Princeton University, March 1988.

APPENDIX Q

E. I. Verriest, "On three-dimensional Rotations, Coordinate Frames, and Canonical Forms for It All", Proceedings of the IEEE, vol. 75, pp. 1376-1378, October 1988.

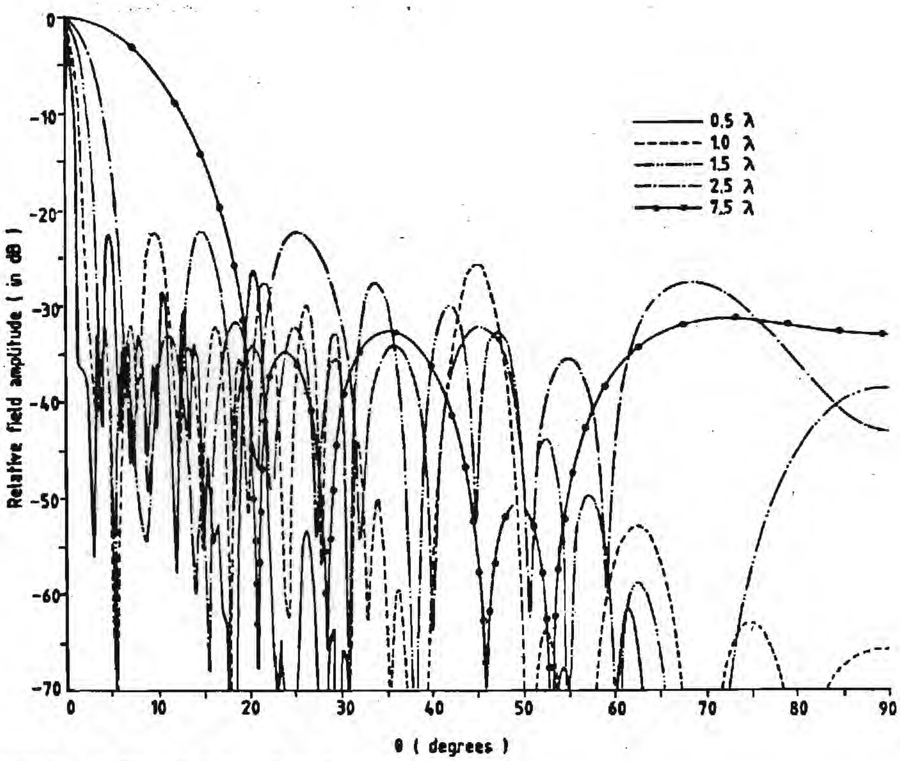

Fig. 4. Vertical beam patterns for a multiplicative coaxial circular array.

[2] ——, "Sidelobe suppressed beampatterns of a coaxial circular array at operating wavelength useful to underwater transducer applications," *Acoust. Lett.*, vol. 11, no. 3, pp. 34–38, 1987.

[3] D. R. Hill, "Reduction of sidelobes in uniformly excited arrays with element pattern control," *Electron. Lett.*, vol. 16, no. 4, pp. 134–135, Feb. 1980.

[4] M. I. Skolnik, *Introduction to Radar Systems.* Singapore: McGraw-Hill, 1985, ch. 7, pp. 233-234.

[5] R. J. Urick, *Principles of Underwater Sound.* New York, NY: McGraw-Hill, 1975, ch. 2, pp. 60–68.

On Three-Dimensional Rotations, Coordinate Frames, and Canonical Forms for It All

ERIK I. VERRIEST

Some properties of the eigenproblem for a three-dimensional rotation matrix are shown, and related to the geometrical rotation parameters. The problem of assigning a unique canonical coordinate frame to a set of three mutually orthogonal axes is considered. The assignment is such that it corresponds to a minimal overall rotation with respect to the reference system. This problem is of interest for the unique and consistent labeling of the principal axes of various tensors related to physical properties of materials, and symmetric matrices that appear in various disciplines of engineering.

Introduction

In the areas of celestial and applied mechanics, robotics, the theory of elasticity, radar and sonar, and in nuclear, molecular and solid-state physics, one frequently needs to express preferential spatial orientations (attached to a "rigid" body) in terms of some fixed reference system (the "laboratory" system). Coordinate transformations are also of interest in expressing material parameters such as dielectric tensors, electrooptic tensors, stress tensors, and so on. Any three fixed mutually orthogonal lines intersecting in 0 (e.g., obtained by solving the eigenproblem for a real symmetric matrix), define 48 possible coordinate frames (of which 24 are right-handed). As the labeling of the preferential axes is usually arbitrary, this paper addresses the problem of providing a "nice" way to uniquely describe or represent a preferential coordinate frame.

As our solution relies on some elementary properties of 3-D rotations, some basic properties of such transformations are first recalled.

Background and Notation

A half-line originating in 0 (the origin) will be called an axis. If u is an axis, then the axis parallel to u but extending in the opposite direction will be denoted by $-u$. By a (right-handed) coordinate frame F, we understand an ordered triple of mutually orthogonal axes, following the right-hand rule. A frame consisting of the axes u, v, and w in that particular order will be denoted by (u, v, w). $\boldsymbol{F}$ is the set of all possible right-handed coordinate frames.

There are many ways to specify the orientation of a coordinate frame relative to another orthogonal coordinate frame with the same origin. Denoting the axes of the fixed reference frame by (x, y, z), and of the preferential coordinate frame by (x', y', z'), it is standard to represent the rotation by the direction cosines of the primed axes relative to the unprimed ones. One can think of the new (primed) coordinate system as the one resulting by operating on the original (unprimed) system by some transformation, and it is well known that the set of matrices Θ representing these transformations form the rotation group SO(3).

Since any rotation can be represented as a global rotation over $\theta \in [0, \pi]$, measured counterclockwise about some axis u, a representation of the set of three-dimensional rotations can be given in spherical coordinates: Longitude ϕ and latitude ψ suffice to identify the global rotation axis u, and the radius $r = \theta$ describes the angle of rotation. However, SO(3) is not topologically equivalent to the open (or closed) ball, since antipodal points on the surface of the sphere represent the same rotation. A standard homotopy argument shows that the fundamental group contains two ele-

Manuscript received November 19, 1987. The research was supported by the U.S. Air Force under Contract AFOSR-87-0308.

The author is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30306, USA.

IEEE Log Number 8821052.

0018-9219/88/1000-1376$01.00 © 1988 IEEE

ments [3]. The covering of SO(3) by SU(2) leads to the Cayley-Klein parameterization [1].

The eigenproblem for rotation matrices is summarized in the following:

Lemma: i) A rotation matrix Θ has all its eigenvalues on the unit circle. If the rotation is nontrivial, only one eigenvalue equals +1. The global geometric rotation angle θ satisfies $\cos\theta = [\mathrm{tr}(\Theta) - 1]/2$, and the rotation axis corresponds to the global rotation vector (i.e., the eigenvector u, corresponding to the eigenvalue 1). ii) The real and imaginary parts of any complex eigenvector corresponding to a nonunity eigenvalue have the same norm, and together with the eigenvector corresponding to the eigenvalue +1 they form a mutually orthogonal set.

Proof: Part i) is shown in [1]. As for ii), let v be the complex eigenvector of Θ, corresponding to the eigenvalue λ. Expressing $u'\Theta v$ in different ways results in $u' \,\mathrm{Re}\,(v) = u' \,\mathrm{Im}\,(v) = 0$, unless the rotation is trivial. Similarly, the simplification of $v'\Theta v$ leads to $v'v = 0$, which in turn implies the orthogonality of Re (v) and Im (v), and equality of their norms.

It follows at once that $\{\mathrm{Re}\,(v), \mathrm{Im}\,(v)\}$ is an orthogonal basis in the rotation plane. A canonical parameterization can be shown to result.

Theorem: Any rotation matrix has a (nonunique) eigenvalue decomposition

$$\Theta = [u, v, \bar{v}] \,\mathrm{diag}\,(1, e^{i\theta}, e^{-i\theta})\, [u, v, \bar{v}]'$$

where θ lies in the interval $[0, \pi]$, measured counterclockwise with respect to u, and such that $[u, \mathrm{Re}\,(v), \mathrm{Im}\,(v)]$ belongs to F.

Proof: See [4].

Labeling of Frames Associated with the Real Symmetric Eigenproblem

In many problems the rotation matrix Θ is only of intermediate interest. In particular, consider the real symmetric eigenvector decomposition $A = UKU'$, also known as the *Principal Axis (Component) Decomposition*. K is a diagonal matrix and U is orthogonal. If $\det U = 1$, then $[u_1, u_2, u_3]$ is the matrix of direction cosines of a new right-handed coordinate system whose coordinate axes are aligned with the vectors u_1, u_2, and u_3. Clearly, this decomposition is not unique: eigenvalues can be permuted, and to each ordering, several different choices for the corresponding eigenvectors may exist, leading to different orthonormal coordinate frames. For reference purposes, a canonical decomposition is desirable. To facilitate the search for canonical forms, the problem is characterized in terms of its invariants, introducing the following definitions:

Definition 1: A Symmetric Eigenvalue Decomposition (SED) is an ordered pair (U, K) where U belongs to the set F, and K is an ordered triple of real numbers (i.e., an element of $\boldsymbol{R}^3$).

Definition 2: Two SEDs (U_1, K_1) and (U_2, K_2) are called

i) *(Weakly) Equivalent* $(\sim)$ iff $U_1 \,\mathrm{diag}\,(K_1) U_1' = U_2 \,\mathrm{diag}\,(K_2) U_2'$ and K_1 is an ordered triple of the permuted elements of K_2.

ii) *Strongly Equivalent* $(\underset{\sim}{s})$ if they are equivalent and $K_1 = K_2$.

The equivalence classes induced by the above equivalences are nontrivial. In the nondegenerate case, the equivalence class of all frames equivalent to a given frame U is generated by operating on U by transformations of the group G_e, consisting of the elementary operations:

1) cyclically relabeling of the coordinate axes x', y', z',
2) changing the directions of any two axes,
3) changing the direction of one axis, and switching the remaining two.

It follows that every orbit of this group (equivalence class $F/\sim$) consists of exactly 24 elements. In the restricted case of Strong Equivalence, the frames can only be related by inversion of any two axes, thus leaving only 4 elements in each class of $F/\underset{\sim}{s}$. The matrix representations of the generators of G_e are:

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad S_a = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$S_b = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad Q = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The generators of the subgroup G_{se} associated with the Strong Equivalence are S_a and S_b. The action of the group element $G \in G_e(G_{se})$ on F is defined by $G(F) = FG \in F$: The frame derived from F by group operation G is the frame with associated matrix FG. The degenerate case is discussed in [4].

Canonical Forms for Real Symmetric Matrix Decompositions

A selection of a canonical form for the decomposition means that to each element $[F]$ of the equivalence classes $F/\sim$ (or $F/\underset{\sim}{s}$) a unique representant is assigned [2]. An obvious choice is the frame obtainable by a rotation of the reference frame over the smallest possible angle. The ideas are made rigorous by introducing a "correlation" metric $\langle \cdot, \cdot \rangle : F \times F \rightarrow \boldsymbol{R}$ defined by $\langle F_1, F_2 \rangle = \mathrm{tr}\, F_1 F_2'$. The function $\langle \cdot, \cdot \rangle$ is not an inner product on the set of frames F, since the latter has not been endowed with a linear structure (i.e., addition of two frames or scalar multiplication of a frame are not defined). An inner product interpretation is possible by embedding F in a 9-dimensional vector space [4]. As the reference frame is represented by the identity matrix, one obtains the "correlation": $\langle I, F \rangle = \mathrm{tr}\, F = (1 + 2\cos\theta)$. The frame with the minimal $\theta \in [0, \pi]$ is the one with the maximal correlation. The map f assigning this optimal frame to each equivalence class is a complete invariant [2], and it follows that the set of frames $\{F_c\} = f(F)$ is a set of canonical forms for F.

The selection algorithm for the canonical representation of a given frame F proceeds then by optimizing $\mathrm{tr}\, FG$ over the elements G of the groups G_e or G_{se} generated by P, Sa, Sb, and Q, or Sa and Sb alone for the canonical forms, respectively, under Equivalence and Strong Equivalence in the nondegenerate case. The pseudocode is provided in the appendix.

Appendix

The Optimal Frame Algorithm for the Nondegenerate Case

Given a right-handed frame represented by the vectors of the direction cosines $[x_a, x_b, x_c]$, the Canonical Form under Equivalence is obtained via the following algorithm (for Strong Equivalence omit the loops in i and j).

```
Begin

    For $i$ := 0 to 2 do
      begin
          $[x_a, x_b, x_c] := [x_a, x_b, x_c]P^i$
          For $j$ := 0 to 1 do
            begin
              $[x_a, x_b, x_c] := [x_a, x_b, x_c]Q^j$
              Find diagonal elements $(x_{a1}, x_{b2}, x_{c3})$.
              If not all signs are positive,
                then find the two columns whose sign change maximizes the trace.
                $[x_a, x_b, x_c] := [x_a, x_b, x_c]\, S$, where $S$ is one of $S_a, S_b, S_c$
              $F(2i + j) := [x_a, x_b, x_c]$           { Store a potentially optimal frame }
              $T(2i + j) := x_{a1} + x_{b2} + x_{c3}$  { Store its correlation with $I$ }
            end
      end
    $k_{max}$ := maximum$^{-1}$ ( $T(k)$ ; $i$ = 0 to 5 )   { Search for maximum }
    Frame := $F(k_{max})$                            { Output the optimal frame }

End.
```

ACKNOWLEDGMENT

The author wishes to thank Dr. T. Gaylord of the Georgia Institute of Technology for suggesting this problem, and for providing some helpful comments.

REFERENCES

[1] H. Goldstein, *Classical Mechanics*. New York, NY: Addison-Wesley, 1980.
[2] S. MacLane and G. Birkhoff, *Algebra*. New York, NY: Macmillan, 1967.
[3] G. G. Hall, *Applied Group Theory*. London: Longmans, 1967.
[4] E. I. Verriest, "On three dimensional rotations, coordinate frames, and canonical forms for it all," Rep. EIV-10-6-87, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA.

An Improved Algorithm for Low-Pass to Bandpass Transformations

STEPHEN A. DYER

An algorithm is presented for computing the coefficients of a continuous-time bandpass transfer function, obtained by applying the standard transformation to a normalized low-pass prototype. The method has all the desirable features of a recently described algorithm while achieving increased computational efficiency through use of a recursion relation.

I. INTRODUCTION

In a recent letter [1], an algorithm was presented for performing the standard low-pass (LP) to bandpass (BP) transformation. The method is quite general, being both independent of filter order and applicable to prototypes having both poles and zeros. Since it is algebraic in nature, it provides excellent accuracy independent of choice of scaling factor.

The algorithm presented in the following shares all the desirable traits of that in [1]. However, while the method in [1] requires the evaluation of a set of binomial coefficients, the present algorithm employs a recursion relation, resulting in decreased computational effort.

II. DEVELOPMENT OF THE ALGORITHM

We wish to obtain the coefficients of the BP transfer function $R(s)$, obtained from the normalized LP transfer function $P(s)$ by the transformation

$$R(s) = P(s)\Big|_{s=[(s^2+\omega_0^2)/Ws]=\alpha s+\beta s^{-1}} \tag{1}$$

where ω_0 is the desired center frequency, in r/s, of $R(s)$; W is the desired bandwidth, in r/s, of $R(s)$; $\alpha = 1/W$; and $\beta = \omega_0^2/W$.

The LP transfer function $P(s)$ in (1) is assumed to have the general form

$$P(s) = \frac{\sum\limits_{n=0}^{N} a_n s^n}{\sum\limits_{m=0}^{M} b_m s^m}. \tag{2}$$

The BP transfer function $R(s)$ has the form

$$R(s) = \frac{\sum\limits_{n=0}^{\nu} \hat{a}_n s^n}{\sum\limits_{m=0}^{\mu} \hat{b}_m s^m}. \tag{3}$$

The transformation (1), when applied to (2), yields

$$R(s) = \frac{\sum\limits_{n=0}^{N} a_n[\alpha s + \beta s^{-1}]^n s^\gamma}{\sum\limits_{m=0}^{M} b_m[\alpha s + \beta s^{-1}]^m s^\gamma} \tag{4}$$

where $\gamma = \max(N, M)$. The factor s^γ/s^γ is included to obtain a form for $R(s)$ which contains only nonnegative powers of s.

We concentrate for the moment on the numerator polynomial in (4), writing it in the form of a power series, as in (3). We have, then,

$$\sum_{k=0}^{\nu} \hat{a}_k s^k = \sum_{n=0}^{N} a_n[\alpha s + \beta s^{-1}]^n s^\gamma \tag{5}$$

$$= \sum_{n=0}^{N} a_n \sum_{k=0}^{\nu} c_{k,n} s^k \tag{6}$$

$$= \sum_{k=0}^{\nu} \left[\sum_{n=0}^{N} a_n c_{k,n}\right] s^k \tag{7}$$

where, from the right-hand sides of (5) and (6), $\nu = \gamma + N$. Thus, the $\hat{a}_k$ can be found as

$$\hat{a}_k = \sum_{n=0}^{N} a_n c_{k,n}, \qquad k = 0, 1, \cdots, \nu. \tag{8}$$

Similarly, the $\hat{b}_k$ of (3) can be found as

$$\hat{b}_k = \sum_{m=0}^{M} b_m c_{k,m}, \qquad k = 0, 1, \cdots, \mu \tag{9}$$

where $\mu = \gamma + M$. Actually, the range of n and m in (8) and (9), respectively, can be restricted further. This matter is discussed in Section IV.

We need, however, to determine the $c_{k,n}$ before (8) and (9) can be applied. From (5) and (6),

$$\sum_{k=0}^{2\gamma} c_{k,n} s^k = (\alpha s + \beta s^{-1})^n s^\gamma \tag{10}$$

$$= (\alpha s + \beta s^{-1})(\alpha s + \beta s^{-1})^{n-1} s^\gamma$$

$$= (\alpha s + \beta s^{-1}) \sum_{k=0}^{2\gamma} c_{k,n-1} s^k$$

$$= \sum_{k=0}^{2\gamma} \alpha c_{k,n-1} s^{k+1} + \sum_{k=0}^{2\gamma} \beta c_{k,n-1} s^{k-1}. \tag{11}$$

Here, the upper limit on the sums is set to 2γ so that (10) can be applied to either (8) or (9) as needed.

After changes of variables, (11) becomes

$$\sum_{k=0}^{2\gamma} c_{k,n} s^k = \sum_{k=1}^{2\gamma+1} \alpha c_{k-1,n-1} s^k + \sum_{k=-1}^{2\gamma-1} \beta c_{k+1,n-1} s^k. \tag{12}$$

So, upon equating coefficients of like powers of s in (12), we obtain the recursion relation

$$c_{k,n} = \alpha c_{k-1,n-1} + \beta c_{k+1,n-1}, \qquad \begin{array}{l} k = 1, \cdots, 2\gamma - 1 \\ n = 1, \cdots, \gamma. \end{array} \tag{13}$$

For $n = 0$, (10) gives

$$c_{k,0} = \begin{cases} 1, & k = \gamma \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Also, (12) yields

$$c_{0,n} = \beta c_{1,n-1}, \qquad n = 1, \cdots, \gamma \tag{15}$$

and

$$c_{2\gamma,n} = \alpha c_{2\gamma-1,n-1}, \qquad n = 1, \cdots, \gamma. \tag{16}$$

III. THE ALGORITHM

The BP coefficients $\hat{a}_k$ and $\hat{b}_k$ of the $R(s)$ in (3) are computed as follows:

1) Accept N; M; ω_0; W; a_n, $n = 0, \cdots, N$; b_m, $m = 0, \cdots, M$.

Manuscript received November 3, 1987; revised February 8, 1988.

The author is with the Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA.

IEEE Log Number 8821053.

APPENDIX R

E. I. Verriest, "Alternating Discrete Time Systems: Invariants, Parametrization and Realization", <u>Proc. Annual Conference on Information Sciences and Systems</u>, Princeton University, pp. 952-957, March 1988.

ALTERNATING DISCRETE TIME SYSTEMS: INVARIANTS, PARAMETRIZATION AND REALIZATION

Erik I. Verriest
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
(404)894-2949

Abstract

Periodic discrete time systems are analyzed. In particular we investigate the Invariants, Parametrizations, Canonical Forms, and Realization from input/output data for such systems. It was found that the classical realization theory for time invariant systems carries over very nicely to such systems. For notational simplification, some results are worked out for the alternating (i.e. period two) single input single output discrete time system only. A novel definition for an Operational Transferfunction is given, which is useful in studying reductions, realizations and interconnections of such systems.

1. Introduction

This paper deals with periodic discrete time systems of period N. To fix the ideas, a state space realization of such systems is of the form

$$\begin{aligned} x_{k+1} &= A_{p(k)}x_k + B_{p(k)}u_k \\ y_k &= C_{p(k)}x_k \\ p(k) &= k \bmod N \end{aligned}$$

The N-tuple $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_{N-1}\}$ where Σ_i is the triple (A_i, B_i, C_i) will refer to such a realization. These systems arise for instance by discretization of periodically, non-uniformly sampled continuous time systems, and more general periodically switched systems. In order to simplify the ideas, we shall sometimes look at the special case of alternating (i.e. period two) single input single output discrete time system. The main ideas for the general case are not different, but only more complex in notation.

While these systems are in many ways more complex than ordinary time- invariant systems, they have still much more structure than general time varying discrete time systems analyzed by Kamen [3], or even the multi-mode systems described by Stanford et al. [2], and Helmke [1], and one can develop a parametrization theory for these systems which is in close analogy to the known geometric theory for stationary systems (Hazewinkel [4]).

In particular, the input/output behavior of such systems is left invariant by the transformation group

$$GL_n(R) \times \ldots \times GL_n(R) \ (N\text{copies}),$$

and the orbit space of the controllable systems is a manifold which can be decomposed into generalized Kronecker cells which form a cellular patch complex. The canonical forms act as local coordinate systems.

Our next main result involves the realization of such a system from the knowledge of the impulse response sequences

$$h_{i,j}\ ; i > j,\ j = 0, 1, ..$$

2. I/O Equivalent Time-Invariant Representations for period-N Systems

Some preliminary definitions and notations will be given in this section. Also, the observability, reachability and stability properties will be discussed. The properties and representations are the key to the realization given in section 4. We shall discuss the general case for N-periodic systems in this section.

Given the N-periodic system $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_{N-1}\}$, let the response of the system to a pulse occurring at instant $j < N$ be the sequence $h_{i,j}$; $i > j$. The system response is readily seen to be (where $[k]$ indicates k mod N)

$$\begin{aligned} h_{i,j} &= C_{[i]}A_{[i-1]}A_{[i-2]}\cdots A_{[j+1]}B_{[j]} \qquad i > j \\ &= 0 \qquad\qquad \text{else} \end{aligned} \tag{1}$$

Define the "Hankel" Matrices for this Periodic System as the matrices H_{j+1} whose (a, b)-element is $h_{j+a,j-b+1}$. This matrix does not have the (block) Hankel structure as in time invariant systems. However, it still allows a factorization in an observability and a reachability matrix (as defined in the time-varying case).

$$H_{j+1} = O_{j+1}R_j \tag{2}$$

e.g. the a-th block entry in O_{j+1} and the b-th block of R_j are respectively

$$[R_j]_b = A_{[j]}A_{[j-1]}\cdots A_{[j+2-b]}B_{[j+1-b]} \tag{3}$$

$$[O_{j+1}]_a = C_{[j+a]}A_{[j+a-1]}\cdots A_{[j+1]} \tag{4}$$

For fixed j in $1, \ldots, N$, the derived sequence $h_k = h_{j+k,j}$; $k > 0$ is also the response to a unit pulse, of the following augmented time invariant system of order nN. (Note that $\Sigma_N \equiv \Sigma_0$)

$$A_{ca} = \begin{bmatrix} 0 & 0 & \ldots & A_N \\ A_1 & 0 & \ldots & 0 \\ 0 & A_2 & 0 & 0 \\ \ldots & & & \\ 0 & 0 & \ldots A_{N-1} & 0 \end{bmatrix} .$$

$$C_{ca} = \begin{bmatrix} C_1 & C_2 \ldots & C_N \end{bmatrix} \tag{5}$$

with read-in matrix $[0, \ldots 0, B'_j, 0, \ldots 0]'$ where the nonzero block B_j occurs in the $(j+1)$-th block position. Such a time invariant representation of the pulse response sequence $H_{i,j}$; $i > j$ will be called an Adiabatic representation. The corresponding Adiabatic Hankelmatrices $\hat{H}_j$ with (a, b)-element $h_{j+a+b-1,j}$, will have the true Hankel structure. The subscript "ca" refers to "cyclically augmented". The above representation is in general not minimal. A minimal realization of the adiabatic Hankel matrix $\hat{H}_j$ will be denoted by $(\hat{A}_j, \hat{B}_j, \hat{C}_j)$.

In order to treat all $h_{i,j}$'s at once, an equivalent composite system (the Cyclically Augmented System) of Nn states, Nm inputs and p outputs, is defined as the realization (A_{ca}, B_{ca}, C_{ca}) where A_{ca} and C_{ca} are as in (2), and defining a B_{ca}-matrix as

$$B_{ca} = \begin{bmatrix} B_N & 0\ldots & 0 \\ 0 & B_1\ldots & 0 \\ \ldots & & \\ 0 & 0\ldots & B_{N-1} \end{bmatrix} \tag{6}$$

Letting $\hat{H}_j(z)$ denote the Zee-transform of the shifted sequence $h_{k+j,j}\ :\ k > 0$, then the transfermatrix of the cyclically augmented system is simply

$$H_{ca}(z) = [\hat{H}_0(z), \hat{H}_1(z), \ldots, \hat{H}_{N-1}(z)] \qquad (7)$$

The $\hat{H}_j(z)$ are the transfermatrices of the ADIABATIC systems, and it follows from the previous discussion that they are realized in a nonminimal way by $(A_{ca}, [0, \ldots, 0, B_j', 0, \ldots, 0]', C_{ca})$, the nonzero element in the B-matrix occurring in the $(j+1)$st block position.

Remarks

1. Dually, we can also work with an equivalent (nN, m, Np) system, thus treating the periodic system as an equivalent stationary Np-output and m- input system.

2. Classical realization theory for multivariable time-invariant systems enables us to find a minimal realization (F, G, H) for the above Hankel matrix. This minimal realisation is then the key to the rest of our development. In particular, since the equivalent stationary system captures all of the input/output information of the periodic one, so will its minimal realization (F, G, H). A parametrization for the periodic systems follows then directly from the parametrization of the multivariable system (F, G, H). At once, we see that even a scalar periodic system leads to multivariable equivalent systems. The restriction to scalar systems mentionned at the onset is thus not restrictive, but permits simpler notation and examples.

We indicate some particular results which will be usefull in the realization problem

Theorem 1: The minimal realizations $(\hat{A}_i, \hat{B}_i, \hat{C}_i)$ of the adiabatic Hankelmatrices $\hat{H}_i$ have the property that $\det(zI - \hat{A}_i)$ divides $\det(z^N I - A_0 \ldots A_{N-1})$

Proof (for $N = 2$): The Hankelmatrix $\hat{H}_0$ is obtained from the pulseresponse $h_{i,0}$. Its Z-transform equals

$$\begin{aligned}\hat{H}_0(z) &= C_0(z^2 I - A_1 A_0)^{-1} A_1 B_0 + C_1(z^2 I - A_0 A_1)^{-1} z B_0 \\ &= [N_0(z^2) + N_1(z^2)]/\det(z^2 I - A_0 A_1)\end{aligned}$$

for some polynomial matrices N_0 and N_1. Clearly then the minimal realizations of $\hat{H}_0$ and $\hat{H}_1$ have the above stated property.

In fact, it is easy to show that the realizations of H_0 and H_1 must be very closely related. Indeed, by rewriting H_1 in the form

$$\begin{aligned}\hat{H}_1 &= [C_1, C_0] \begin{bmatrix} z^2 I - A_0 A_1 & 0 \\ 0 & z^2 I - A_1 A_0 \end{bmatrix}^{-1} \begin{bmatrix} A_0 \\ zI \end{bmatrix} B_1 \\ &= [C_0, C_1] \begin{bmatrix} z^2 I - A_1 A_0 & 0 \\ 0 & z^2 I - A_0 A_1 \end{bmatrix}^{-1} \begin{bmatrix} zI \\ A_0 \end{bmatrix} B_1\end{aligned}$$

The first factors on the left also appear in the expansion of H_0

$$\hat{H}_0 = [C_0, C_1] \begin{bmatrix} z^2 I - A_1 A_0 & 0 \\ 0 & z^2 I - A_0 A_1 \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ zI \end{bmatrix} B_0$$

Hence, if we define the following transfermatrix:

$$\hat{H}_{01} = [C_0, C_1] \begin{bmatrix} z^2 I - A_1 A_0 & 0 \\ 0 & z^2 I - A_0 A_1 \end{bmatrix}^{-1} \begin{bmatrix} A_1 & zI \\ zI & A_0 \end{bmatrix}$$

then $\hat{H}_0 = \hat{H}_{01}[B_0', 0]'$ and $\hat{H}_1 = \hat{H}_{01}[0, B_1']'$.

This observation leads directly to the following theorem:

Theorem 2: There exists an observable pair $(\bar{A}, \bar{C})$ and matrices $\bar{B}_0$ and $\bar{B}_1$ such that $(\bar{A}, \bar{B}_0, \bar{C})$, and $(\bar{A}, \bar{B}_1, \bar{C})$ realize respectively the adiabatic transfer matrices $\hat{H}_0$ and $\hat{H}_1$.

Proof: Let $(\bar{A}, \bar{B}, \bar{C})$ be a minimal (observable is sufficient) realization for $\hat{H}_{01}$, then $\bar{B}_0 = \bar{B}[B_0', 0]'$, and $B_1 = B[0, B_1']'$.

The importance of this theorem lies in its use to find the realizations for an alternating system. Given the pulse response sequences $h_{i,0}$ and $h_{i,1}$, we can use the realization algorithm from time invariant systems to determine minimal realizations of either sequence. By the theorem, these realizations can be extended by addition of uncontrollable states if necessary, to observable realizations with the same A and C matrix.

3. Reachability, Observability and Stability

Definitions:

- The N-periodic system $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_{N-1}\}$ is said to be uniformly p-reachable (reachable in p steps), iff every state can be reached in p steps, independently of the starting event (= initial time and initial state). The system is said to be uniformly reachable, iff there exists a $p > 0$, such that it is uniformly p-reachable.

- The system is said to be uniformly observable in p steps iff the initial state x_j can be uniquely determined from p consecutive outputs $y_j, \ldots, y_{j+p-1}$, independently of the starting time j. The system is said to be uniformly observable iff it is p-observable for some p.

Theorem 3: The period-N system $\{\Sigma_0, \ldots, \Sigma_{N-1}\}$ is uniformly reachable iff the reachability matrices (3) have full rank for all j. The system is uniformly observable iff the observability matrices (4) have full rank for all j.

The proof is easily established by a standard argument [5]. Since the adiabatic systems of at most order nN, provide an underlying time-invariant structure in the problem, at most nN steps need to be considered for checking uniform reachability and observability, by virtue of the Cayley- Hamilton Theorem. Some direct corollaries of the theorem are:

i) The Cyclically Augmented system (A_{ca}, B_{ca}, C_{ca}) is reachable iff the period-N realization $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}$ is uniformly reachable.

ii) $\{\Sigma_1, \Sigma_2 \ldots, \Sigma_0\}$ is uniformly observable (reachable) iff $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_{N-1}\}$ is uniformly observable (reachable), whence the invariance of uniform observability and reachability under a cyclic shift.

iii) Using the backward propagation, we can write the output at time i in terms of the previous inputs. i.e., we look at $h_{i,j}$ for fixed i, and define the equivalent stationary systems with the above A-matrix and $C = [0, \ldots 0, C_i, 0, \ldots, 0]$, the nonzero block occuring in the i-th block position, and $B = (B_0', B_1', \ldots, B_{N-1}')'$. We then have the "duality"-property: $\{\Sigma_1, \ldots, \Sigma_{N-1}, \Sigma_N\}$ is unif. observable iff $\{\Sigma_N^d, \Sigma_{N-1}^d, \ldots \Sigma_1^d\}$ is unif. reachable, where the "dual" system is obtained by time reversal of the sequence of the duals Σ_i^d of the realizations Σ_i, where $(A_i, B_i, C_i)^d$ is the triple (A_i', C_i', B_i'). We are thus led to the definition:

$$\{\Sigma_1, \Sigma_2, \ldots, \Sigma_0\}^{\text{dual}} = \{\Sigma_0^d, \Sigma_{N-1}^d, \ldots, \Sigma_1^d\} \qquad (8)$$

Finally, we remark that if all A_i are nonsingular, as for instance in the important case of the discretization of a continuous system, the criterion of Theorem 1 can be simplified by virtue of the following

Lemma: If the A_j are nonsingular for all j, then the full rankness of one of the reachability matrices R_i (observability matrices O_i) implies the full rankness of all others, and hence reachability (observability).

As an example, a siso alternating system Σ_0, Σ_1 will be uniformly reachable iff the stationary systems $(A_1 A_0,\ [b_1, A_1 b_0])$ and

$(A_0A_1, [b_0, A_0b_1])$ are reachable. If the system is uniformly reachable, no more than $2n$ steps are required to reach any desired end-state. If the product A_0A_1 is nonsingular, then $(A_1A_0, [b_1, A_1b_0])$ and $(A_0A_1, [b_0, A_0b_1])$ are either both reachable or both non-reachable. By applying inputs before 0, one gets the reachability relation at time 0:

$$x_0 = R_1[u_0, u_{-1}, \ldots]'$$

where $R_1 = [b_1, A_1b_0, A_1A_0b_1, \ldots]$ is the time varying reachability matrix [3]. Observation of the output sequence after time 0, with no input applied leads then to the observability relation:

$$[y_0, y_1, y_2, \ldots]' = O_0x_0$$

where $O_0 = [c_0', A_0'c_1', A_0'A_1'c_0', \ldots]'$ is the time varying observability matrix. Similarly, we construct the reachability and observability matrices, R_0 and O_1, relating to the reference time 1. The products O_0R_1 and O_1R_0 are then the alternating (period-2) Hankelmatrices defined in (2).

We also have the following important stability theorem:

Theorem 4: The N-period system (0) is stable if the eigenvalues of the product $A_0A_1 \ldots A_{N-1}$ have modulus less than 1.

Proof: The convergence properties of the periodic systems are determined by the convergence properties of the equivalent time invariant system (A_{ea}, B_{ea}, C_{ea}). The latter is completely determined by the characteristic polynomial
$\det(z^N I - A_1A_2A_3 \ldots A_N) = 0$.

The problem with this approach is that the resulting time-invariant system has order Nn if n is the order of the individual realizations R_1. The original periodic system is only of n-th order, so that a "hidden modes"-phenomenon occurs.

4. Canonical Forms, Parametrization and Topological Structure

The first object in this study is to find the transformations on the realizations that leave the input-output behavior (i.e. all adiabatic transfermatrices and the the periodic system "Hankel" matrices (2)) invariant.

Let $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_{N-1}\}$ be a realization of an N-period system. Denote an element of the group $Gl_n(R)^N$, denoted by Gl_n^N for short, by $(P_0, P_1, \ldots, P_{N-1})$. The group action is defined by

$$(P_0, \ldots, P_{N-1}) : (A_0, B_0, C_0), \ldots, (A_{N-1}, B_{N-1}, C_{N-1}) \longrightarrow$$
$$(P_1A_0P_0^{-1}, P_1B_0, C_0P_0^{-1}), (P_0A_{N-1}P_{N-1}^{-1}, P_0B_{N-1}, C_{N-1}P_{N-1}^{-1}) \quad (9)$$

The states transform as

$$\begin{aligned} x_{Nk} &\longrightarrow P_0x_{Nk} \\ x_{Nk+i} &\longrightarrow P_ix_{Nk+i} \qquad i = 1, \ldots, N-1 \end{aligned} \quad (10)$$

The following property is readily shown:

Theorem 5: Equivalence of State Space Representations.

The product group $Gl_n(R)^N$ action on the set of period-N systems leaves the I/O properties invariant.

Once the symmetry group, (i.e. the group whose action leaves the I/O behavior invariant) is established, we can look at the question of canonical forms: Any property of the original system can be described as a map from the set of systems Σ, to some suitable set S. After introducing canonical forms, the study of the original function f is then replaced by the study of some "simpler" function $\hat{f} : C \longrightarrow S$, such that $f = \hat{f} \circ \pi$, where π is the canonical projection $\pi : \Sigma \longrightarrow C$ on the set of canonical forms.

For notational simplicity, the rest of this section will be restricted to alternating systems. The above development should give enough insight to realize that the general principles remain the same. Canonical forms for the unifomly reachable systems are obtained by the usual Kronecker selection procedure. i.e. among the $2n$ columns of R_0, select n linear independent ones, which form a basis β_0 for the state space. In particular, a unique "nice" selection may be chosen according to the Young or "crate"-diagram. Similarly, let β_1 be another basis, chosen by a nice selection among the columns of R_1. Now express the system with respect to the basis which is alternating between β_0 and β_1. e.g. A_0 is represented by the a new matrix whose j-th column is the representation in terms of the basis β_1 of A_0 operating on the j-th basisvector from the other basis β_0.

The effect is that the new representation is of the form $b_0 = b_1 = [1, 0, \ldots 0]'$. (By assumption of reachability neither b_0 nor b_1 are zero). If $v_i(k)$ denotes the position of the k-th basis vector from R_i, then we refer to the sequences $w_i = v_i(1), \ldots, v_i(n)$ as a multi-index. The k-th column of the new A-matrices are

$$\begin{aligned} A_0^r e_k &= R_0^r e_{v1(k)+1} \\ A_1^r e_k &= R_1^r e_{v0(k)+1} \end{aligned} \quad (11)$$

However straightforward the previous extension of the known scheme may be, a particular nice form is obtained as follows if A_1 is nonsingular. Search the columns of R_0 in their natural order, i.e. from left to right, and reordered in chains, as in the usual "scheme II" search [5]. Now observe that if A_1 is nonsingular, then the result of A1 operating on the above basis, is also a basis, and in fact each of these new basis vectors will be a column in R_1, except perhaps for the last new basisvector. In that case, it may be substituted for b_1 as new last basis vector. Note that this all corresponds to a scheme II search in R_1, but STARTING AT A_1b_0. It follows that, unless there was a full length $(= n)$ chain $A_0b_1, \ldots, (A_0A_1)^{n-1}A_0b_1$ in R_0, (which only happens if b_0 is zero), the operator A_1 is represented by I. The b_1 vector is full in general. The pair (A_0, b_0) has a canonical controllability form [5] representation. In the latter special case, it follows that A_0 is represented by a cyclically down shifted identity matrix, and A_1 by a cyclically down shifted right companion (controllability canonical form) matrix. The new b_0 (obviously) remains zero, and the new b_1 is a full vector. As the c_0 and c_1 have no particular structure, there are $4n$ free parameters in this canonical form. By analogy to the stationary realizations, we shall refer to this form as the controllability canonical form. Note that because the search extends over $2n$ columns, the new A-matrices will in general not be in the usual companion form themselves, unless the system is uniformly reachable in n steps, but then this is also the case with the time invariant multivariable systems. In fact, it is exactly because of such a reduction from the timevarying to the multivariable time-invariant case that all the topological properties of these systems are expected to carry over. In particular, for multivariable systems we have:

Theorem 6: The orbit space of the reachable systems is an analytic manifold, which can be decomposed into generalized Kronecker cells which form a cellular patch complex. The state space canonical forms act as local coordinates.

The number of canonical forms that is required to cover the space of all reachable alternating systems is also equal to the number of pairs of nice multi-indices that can be chosen. The information given here is rather sketchy, but the details will be presented in a forthcoming paper [6].

5. Operational Transfer Function

Because of space limitations, very little will be said here. The essential ingredient is the introduction of a sampling operator π, with $\pi^2 = \pi$, and which does not commute with the shift operator z^{-1}. In fact, we have $z^{-1}\pi z = 1 - \pi$, from which $z\pi + z(1 - \pi) = z$. The equation $x_{2k+1} = A_0x_{2k} + B_0u_k$ is then transformed to $z(1 - \pi)X(z) = A_0\pi X(z) + B_0\pi U(z)$. Similarly, the complementary equation transforms to $z\pi X(z) = A_1(1 - \pi)X(z) + B_1(1 - \pi)U(z)$.

Adding yields the form $X(z) = (zI - A(\pi))^{-1}B(\pi)U(z)$, where $A(\pi) = A_0\pi + A_1(1-\pi)$ and $B(\pi) = B_0\pi + B_1(1-\pi)$. Defining also $C(\pi) = C_0\pi + C_1(1-\pi)$, we get the Operational Transfer function as

$$H(z,\pi) = C(\pi)(zI - A(\pi))^{-1}B(\pi)$$

i.e. a transfermatrix with coefficients in the polynomial ring $R[\pi|\pi^2 = \pi]$. Using the noncommutative relation, this formalism is very helpful in deriving all sorts of results and transferfunction operations. For instance, defining the odd and even part of $G(z)$

$$\pi[G(z)] = G_e(z) = \frac{G(z)+G(-z)}{2}$$

$$(1-\pi)[G(z)] = G_o(z) = \frac{G(z)-G(-z)}{2}$$

then the commutation relation implies for the <u>operators</u>

$$G_e(z)\pi = \pi G_e(z) \qquad G_o(z)\pi = (1-\pi)G_o(z)$$

and

$$G(z)\pi = \pi G_e(z) + (1-\pi)G_o(z)$$

For instance:

$$(zI - M)^{-1}\pi = [(1-\pi)zI + \pi M][z^2 I - M^2]^{-1}$$

This last rule allows to write the OTF of the period-2 system as

$$\begin{gathered} C_o[z^2 I - A_1A_0]^{-1}[A_1B_0\pi + zB_1(1-\pi)] - \\ +C_1[z^2 I - A_0A_1]^{-1}[zB_0\pi + A_0B_1(1-\pi)] \end{gathered}$$

The following duality is also very helpful in reduction.

$$\begin{aligned} H(z,\pi) &= R_1(z)\pi + R_2(z)(1-\pi) \\ &= \pi S_1(z) + (1-\pi)S_2(z) \end{aligned}$$

where

$$\begin{bmatrix} S_1(z) \\ S_2(z) \end{bmatrix} = \begin{pmatrix} \pi & \pi - 1 \\ \pi - 1 & \pi \end{pmatrix} \left(\begin{bmatrix} R_1(z) \\ R_2(z) \end{bmatrix} \right)$$

$$\begin{bmatrix} R_1(z) \\ R_2(z) \end{bmatrix} = \begin{pmatrix} \pi & \pi - 1 \\ \pi - 1 & \pi \end{pmatrix} \left(\begin{bmatrix} S_1(z) \\ S_2(z) \end{bmatrix} \right)$$

Finally, the reachability and observability conditions derived in the previous section are also readily obtained with this formalism. It can also readily be extended for use with period N systems. The commutation relation is different i.e. $Z^{N-1}\pi + Z^{N-2}\pi Z + \ldots + \pi Z^{N-1} = Z^{N-1}$.

6. Realization

In this section we show how the above results can lead to a relatively simple realization algorithm for periodic systems. We shall assume that the order n of the minimal N-period system is known. There is then an underlying uniformly reachable system $\Sigma_0, \ldots, \Sigma_{N-1}$ of order n. The realization is given in 4 steps:

Step 1: Let the pulse responses $h_{i,0}; i > 1, \ldots, h_{i,N-1}; i > N$ be collected. With this data, the adiabatic Hankelmatrices $\hat{H}_j$, $j = 0, \ldots, N-1$ are formed. There are now two possible routes: realise each adiabatic system separately, or realize the compositte system with transfermatrix $[\hat{H}_0(z), \ldots, \hat{H}_{N-1}(z)]$.

Step 2: Obtain a minimal realization $(A_m, [B_m, 0, B_{m,1}, \ldots, B_{m,N-1}], C_m)$ of the system with composite transfer matrix $[\hat{H}_0(z), \ldots, \hat{H}_{N-1}(z)]$, where $\hat{H}_i(z)$ is the transfer matrix of the adiabatic system $(\hat{A}_i, \hat{B}_i, \hat{C}_i)$.

There are two ways in which this can be obtained. The first is to reduce the composite transfermatrix, and obtain a minimal realization of it, by standard multivariable techniques [5]. The second method consists in first obtaining minimal realizations $(\hat{A}_i, \hat{B}_i, \hat{C}_i)$ of the adiabatic systems. These minimal realizations may be of different dimensions. However, by the theorem 1, there exists a maximal characteristic polynomial, of order q say, in the sense that the characteristic polynomials of the other realizations divide it, and each non-minimal adiabatic realization can be extended by adding a non-reachable, but observable subsystem, so that all extended realizations of the adiabatic susbsystems have the same order $(= q)$, and the same A and C matrix. The composite realization $(A_m, [B_{m,0}, \ldots, B_{m,N-1}], Cm)$ is then the desired form and is minimal since (A_m, C_m) is observable, and at least one of the $B_{m,i}$ forms together with A_m a reachable pair.

Step 3: Extend the realization of order q obtained in step 2, to one whose order is a multiple of N, by adding a non-observable but reachable subsystem. Indeed, since the minimal system $(A_m, [B_{m,0}, \ldots, B_{m,N-1}], C_m)$ and the cyclically augmented system (A_{ca}, B_{ca}, C_{ca}) realize the same transfer matrix, and since the cyclically augmented system is uniformly reachable by assumption, the latter must be an extension (A_e, B_e, C_e) of $(A_m, [B_{m,0}, \ldots, B_{m,N-1}], C_m)$ by a non-observable but reachable subsystem of order $Nn - q$. This implies the existence of matrices $X, Y, Z_0, \ldots, Z_{N-1}$ so that

$$\begin{bmatrix} A_m & 0 \\ X & Y \end{bmatrix} \begin{bmatrix} B_{m0}, & \ldots, & B_{m,N-1} \\ Z_0, & \ldots, & Z_{N-1} \end{bmatrix} \begin{bmatrix} C_m & 0 \end{bmatrix}$$

is similar to the cyclically augmented system. This augmentation must not impair the reachability of the realization. The necessary reachability of the subsystem implies that none of the rows of the matrix $[X, Z_0, \ldots, Z_{N-1}]$ can be zero. This follows easily by contradiction. If $[X, Z_0, \ldots, Z_{N-1}]$ had a zero row, then the realization could be partitioned as

$$\begin{bmatrix} A_m & 0 & 0 \\ X_1 & X_2 & Y_1 \\ 0 & 0 & Y_2 \end{bmatrix} \begin{bmatrix} B_{m0}, & \ldots, & B_{m,N-1} \\ Z_0, & \ldots, & Z_{N-1} \\ 0, & \ldots, & 0 \end{bmatrix}$$

which has the un-reachable subsystem $[Y_2, 0, C_m]$.

The X, Y, and Z_i are chosen so that the reachability matrices $R(A_e^2; [B_{e0}, A_eB_{e1}])$, and $R(A_{e2}; [B_{e0}, A_eB_{e1}])$ have rank less than n.

Step 4: Determine the similarity transformation that transforms the extended system $(A_e, [B_{e1}, \ldots, B_{e,N-1}], C_e)$ to a cyclic form. For notational convenience, we shall again discuss the latter for alternating (i.e. period- 2) systems. The ideas for period-N systems are similar.

The reachability matrix for the cyclically augmented matrix has the structure

$$R_{ca} = \begin{bmatrix} B_0, & 0, & 0, & A_0B_1, & A_0A_1B_0, & 0 \\ 0, & B_1, & A_1B_0, & 0, & 0, & A_1A_0B_1 \end{bmatrix}$$

whereas R_e has entries in all positions in general. The desired similarity maps R_e into $TR_e = R_{ca}$. Partitioning T into $[T_1', T_2']'$, it is seen that the zero locations in the above equation leads to the identities:

$$\begin{aligned} T_1R(A_e^2; [B_{e1}, A_eB_{e0}]) &= 0 \\ T_2R(A_e^2; [B_{e0}, A_eB_{e1}]) &= 0 \end{aligned}$$

where the reachability matrices extend to n (block)columns only, and are thus square for single input periodic systems. On the condition that the test matrices $R_{e0} = R(A_e^2; [B_{e0}, A_eB_{e1}])$, and $R_{e1} = R(A_{e2}; [B_{e0}, A_eB_{e1}])$, have rank less than n, it is possible to find n linearly independent rowvectors in the left nullspaces of the above reachability matrices. Since the overall realizations are both reachable, there exists T_1 and T_2 such that $T = [T_1', T_2']'$ is nonsingular with T_1 and T_2 satisfying the above conditions.

We summarize then with:

Theorem 7: An invertible transformation T can always be found, bringing the extended system Σ_e of order $2n$ to the cyclically augmented form $\Sigma_{ea}(5)$ of the same order if it is reachable and if the reachability matrices $R(A_e^2; [B_{e0}, A_e B_{e1}])$ and

$R(A_e^2;$

$[B_{e1}, A_e B_{e0}])$, both have rank at most equal to n.

Theorem 8: The realization of $[\hat{H}_0(z), \hat{H}_1(z)]$ can always be augmented so that the extended system satisfies the conditions of theorem 4.

Remarks:

1. The minimal adiabatic realizations completely specify the I/O behavior of the alternating system, and may therefore lead to new canonical representations for such systems.
2. Other identification methods for periodic systems exist. One can "stagger" the impulse responses, by looking at every N-th sample, for which the system looks like a time invariant one. However, the solution for the individual realizations in $(\Sigma_1, \ldots, \Sigma_N)$ require difficult nonlinear equation solvers. Furthermore since the data samples with such a scheme are not "convoluted", a large number of data needs to be collected (roughly $2nN$) before the system starts to unfold. The scheme presented here already presents a lot of information about the system after $2n$ steps, and is therefore more "holographic".

7. Examples

Example 1. Let $h_{j,0_{j>0}} = c, a, ca, a^2, ca^2, \ldots$, and $h_{j,1_{j>1}} = b, cb, ab, cab, a^2b, \ldots$. The adiabatic systems are realized by

$$\begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}, \begin{bmatrix} b \\ cb \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix}$$

If a and b are not both zero and c^2 is different from a, then a minimal realization of the transfermatrix $[\hat{H}_0(z), \hat{H}_1(z)]$ in observability canonical form [5] is

$$\begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}, \begin{bmatrix} c & b \\ a & cb \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Since the order of this realization is even, we check first the rank conditions on the test matrices

$$R_{e0} = \begin{bmatrix} c & cb \\ a & ab \end{bmatrix}, R_{e1} = \begin{bmatrix} b & a \\ cb & ac \end{bmatrix}$$

Hence chosing T_1 from span $[-c, 1]$ and T_2 from span $[-a, c]$ leads to a transformation (after introducing suitable parameters)

$$T = \begin{bmatrix} t_1 & 0 \\ 0 & t_2 \end{bmatrix} \begin{bmatrix} -c & 1 \\ -a & c \end{bmatrix}$$

which in turn transforms to the above realization to

$$\begin{bmatrix} 0 & -t_1/t_2 \\ -at_2/t_1 & 0 \end{bmatrix}, \begin{bmatrix} t_1(a-c^2) & 0 \\ 0 & -t_2 b(a-c^2) \end{bmatrix},$$
$$\left[c/[(a-c^2)t_1], -1/[(a-c^2)t_2\right]$$

The period-2 realisation is now read out by inspection. Suitably reparametrized, we find

$$(r, t, r/t), \ (a/r, bt/r, c/t)$$

If in the above example we have $c^2 = a$, then the adiabatic realizations are $(c, c, 1)$ and $(c, b, 1)$. This leads to a reachable realization $(c, [c, b], 1)$, provided that c and b are not both 0, of the $[\hat{H}_0(z), \hat{H}_1(z)]$. As its order is odd, augmentation is required. The augmented system is

$$\begin{bmatrix} c & 0 \\ x & y \end{bmatrix}, \begin{bmatrix} c & b \\ z_1 & z_2 \end{bmatrix}, [1, 0]$$

The test matrices are

$$R_{e0} = \begin{bmatrix} c & cb \\ z_1 & xb + yz_2 \end{bmatrix}, R_{e1} = \begin{bmatrix} b & c^2 \\ z_2 & cx + yz_1 \end{bmatrix}$$

For $z_2 = 0, z_1 = x$ and $y = -c$ for instance, the rank of both test matrices is 1. Taking then $T_1 = t_1[0, 1]$, and $T_2 = t_2[-x, c]$, the cyclically augmented realization

$$\begin{bmatrix} 0 & -t_1/t_2 \\ -c^2 t_2/t_1 & 0 \end{bmatrix}, \begin{bmatrix} t_1 x & 0 \\ 0 & -t_2 xb \end{bmatrix}, [c/(xt_1), \ -1/(xt_2)]$$

results, from which the (reparametrized) period-2 system $(r, t, r/t)$,

$(c^2/r, bt/r, c/t)$ follows.

Example 2. Let $h_{j,0_{j>1}} = 2, 2, 1+a, 1+a, 1+a^2, 1+a^2, 1+a^3, \ldots$ and $h_{j,1_{j>2}} = 1+b, 1+qb, 1+ab, 1+a^2b, 1+a^2b, 1+a^3, \ldots$. The adiabatic systems have realizations

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & a & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 1+a \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \text{ and}$$
$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a & a & 1 \end{bmatrix} \begin{bmatrix} 1+b \\ 1+b \\ 1+ab \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

Extending with a first order no-observable state, the system

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -a & a & 1 & 0 \\ x_1 & x_2 & x_3 & y \end{bmatrix} \begin{bmatrix} 2 & 1+b \\ 2 & 1+ab \\ 1+a & 1+ab \\ z_1 & z_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

realizes the augmented adiabatic system $[\hat{H}_0(z), \hat{H}_1(z)]$. It can be checked that the choice $x_1 = x_2 = x_3 = 0$, $y = -1$, $z_1 = 1, z_2 = -1$ gives a system for which the matrices R_{e0} and R_{e1} have rank 2 (if a differs from 1). The left nullspace of R_{e1} is spanned by $[a, 0, -1, a-1]$ and $[0, 1, -1, 0]$, while the left nullspace of R_{e0} is spanned by $[1, -1, 0, 0]$ and $[a, 0, -1, 1-a]$. The special choice for T:

$$\begin{bmatrix} 1 & 0 & 1 & a-1 \\ 0 & -2 & 2 & 1 \\ a & 0 & -1 & 1-a \\ -2 & 2 & 0 & \end{bmatrix} \frac{1}{2(a-1)}$$

yields then an equivalent cyclically augmented realizaton, from which the period-2 system can be identified by inspection as:

$$\begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ b \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}$$

References

1. Helmke, U., "Parametrizations for Multi-Mode Systems and Yang-Mills Instantons", *Proc. 25th Conf. on Decision and Control*, Athens Greece, Dec. 1986.
2. Stanford, D.P., and Conner, L.T., "Controllability and Stabilizability in Multi-Pair Systems", *SIAM J. Control and Optimization*, Vol. 18, No. 5, September 1980, 488-497.
3. Kamen, E.W., and K.M. Hafez, "Algebraic Theory of Linear Time-Varying Systems", *SIAM J. of Control and Optimization*, Vol. 17, 500-510.

4. Hazewinkel, M., "Moduli and Canonical Forms for Linear Dynamical Systems: The Topological Case", *Mathematical Systems Theory*, 10, 363-385, 1977.

5. Kailath, T., *Linear Systems*, Prentice-Hall, 1980.

6. Verriest, E. I., "The Operational Transfer Function and Parametrizations of N-Periodic Systems," submitted to *27th IEEE Conference on Decision and Control*.

Acknowledgement: This research is supported by the U.S. Air Force under Contract No. AFOSR-87-0308.

APPENDIX S

E. I. Verriest, "The Operational Transfer Function and Parametrization of N-periodic Systems", Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, pp. 124-128, Dec. 1988.

THE OPERATIONAL TRANSFER FUNCTION AND PARAMETERIZATION OF N-PERIODIC SYSTEMS

Erik I. Verriest
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

ABSTRACT

Periodic discrete time systems are analyzed. In particular, we investigate the Invariants, Parameterizations, Canonical Forms, and Realization from input/output data for such systems. It was found that the classical realization theory for time invariant systems carries over very nicely to such systems. A novel definition for an Operational Transfer Function is given, which is useful in studying reductions, realizations, and interconnections of such systems.

1. INTRODUCTION

This paper deals with periodic discrete time systems of period N. To fix the ideas, a state space realization of such systems is of the form

$$\begin{aligned} x_{k+1} &= A_{p(k)}x_k + B_{p(k)}u_k \\ y_k &= C_{p(k)}x_k \qquad (0) \\ p(k) &= k \bmod N \end{aligned}$$

The N-tuple $\{\Sigma_0,\Sigma_1,\ldots,\Sigma_{N-1}\}$ where Σ_i is the triple (A_i,B_i,C_i) will refer to such a realization. These systems arise, for instance, by discretization of periodically, nonuniformly sampled continuous time systems, and more general periodically switched systems. In order to simplify the ideas, we shall sometimes look at the special case of alternating (i.e. period two) single input-single output discrete time system. The main ideas for the general case are not different, but only more complex in notation.

While these systems are in many ways more complex than ordinary time-invariant systems, they have still much more structure than general time-varying discrete time systems analyzed by Kamen [3], or even the multi-mode systems described by Stanford, et al. [2], and Helmke [1], and one can develop a parameterization theory for these systems which is in close analogy to the known geometric theory for stationary systems (Hazewinkel [4]).

In particular, the input-output behavior of such systems is left invariant by the transformation group $GL_n(R) \times \ldots \times GL_n(R)$ (N copies), and the orbit space of the controllable systems is a manifold which can be decomposed into generalized Kronecker cells which form a cellular patch complex. The canonical forms act as local coordinate systems.

Our next main result involves the realization of such a system from the knowledge of the impulse response sequences $\{h_{i,j};i>j,j=0,1\}$.

2. I/O EQUIVALENT TIME-INVARIANT REPRESENTATIONS FOR PERIOD-N SYSTEMS

Some preliminary definitions and notations will be given in this section. Also, the observability, reachability, and stability properties will be discussed. The properties and representations are the key to the realization given in Section 4. We shall discuss the general case for N-periodic systems in this section.

Given the N-periodic system $\{\Sigma_0,\Sigma_1,\ldots,\Sigma_{N-1}\}$, let the response of the system to a pulse occurring at instant j<N be the sequence $\{h_{i,j};i>j\}$. The system response is readily seen to be (where [k] indicates k mod N)

$$\begin{aligned} h_{i,j} &= C_{[i]}A_{[i-1]}A_{[i-2]}\cdots A_{[j+1]}B_{[j]} \qquad i>j \\ &= 0 \qquad\qquad \text{else} \qquad (1) \end{aligned}$$

Define the "Hankel" matrices for this periodic system as the matrices H_{j+1} whose (a,b)-element is $h_{j+a,j-b+1}$. This matrix does not have the same (block) Hankel structure as for time-invariant systems. However, it still allows a factorization in an observability and a reachability matrix (as defined in the time-varying case).

$$H_{j+1} = O_{j+1}R_j \qquad (2)$$

e.g. the a-th block entry in O_{j+1} and the b-th block of R_j are, respectively

$$[R_j]_b = A_{[j]}A_{[j-1]}\cdots A_{[j+2-b]}B_{[j+1-b]} \qquad (3)$$

$$[O_{j+1}]_a = C_{[j+a]}A_{[j+a-1]}\cdots A_{[j+1]} \qquad (4)$$

For fixed j in $\{0,\ldots,N\}$, the derived sequence $\{h_k=h_{j+k,j};k>0\}$ is also the response to a unit pulse of the following augmented time-invariant system of order nN.

$$A_{ca} = \left[\begin{array}{llll} 0 & 0\ldots. & & A_0 \\ A_1 & 0\ldots. & & 0 \\ 0 & A_2 0 & & 0 \\ \cdot & \cdot\ \cdot\ \cdot & & \\ 0 & 0\ldots. A_{N-1} & & 0 \end{array}\right]$$

$$C_{ca} = [C_1 \quad C_2\ldots \qquad C_0]$$

with read-in matrix $[0,\ldots 0,B_j',0,\ldots 0]'$ where the nonzero block B_j occurs in the (j+1)-th block position. Such a time-invariant representation of the pulse response sequence $\{H_{i,j};i>j\}$ will be called an Adiabatic representation. The corresponding Adiabatic Hankel matrices $\hat{H}_{ij}$ with (a,b)-element $h_{j+a+b-1,j}$, will have the true Hankel structure. The subscript "ca" refers to "cyclically augmented." The above representation is, in general, not minimal. A minimal realization of the Adiabatic Hankel matrix $\hat{H}_j$ will be denoted by $(\hat{A}_j,\hat{B}_j,\hat{C}_j)$.

In order to treat all $h_{i,j}$'s at once, an equivalent composite system (the Cyclically Augmented System) of Nn states, Nm inputs, and p outputs, is defined as the realization (A_{ca},B_{ca},C_{ca}) where A_{ca} and C_{ca} are as in (2), and defining a B_{ca}-matrix as

$$B_{ca} = \begin{bmatrix} B_0 & 0 \;\ldots. & 0 \\ 0 & B_1 \ldots. & 0 \\ \ldots\ldots & \ldots & \\ 0 & 0 & B_{N-1} \end{bmatrix} \qquad (6)$$

Letting $\hat{H}_j(z)$ denote the Z-transform of the shifted sequence $\{h_{k+j,j};k>0\}$, then the transformation of the cyclically augmented system is simply

$$H_{ca}(z) = \{\hat{H}_0(z),\hat{H}_1(z),\ldots,\hat{H}_{N-1}(z)\} \qquad (7)$$

The $H_j(z)$ are the transfer matrices of the Adiabatic systems, and it follows from the previous discussion that they are realized in a nonminimal way by $(A_{ca},[0,\ldots,0,B_j',0,\ldots,0]',C_{ca})$, the nonzero element in the B-matrix occurring in the (j+1)st block position.

Remarks.

1. Dually, we can also work with an equivalent (nN,m,Np) system, thus treating the periodic system as an equivalent stationary Np-output and m-input system.

2. Classical realization theory for multivariable time-invariant systems enables us to find a minimal realization (F,G,H) for the above Hankel matrix. This minimal realization is then the key to the rest of our development. In particular, since the equivalent stationary system captures all of the input-output information of the periodic one, so will its minimal realization (F,G,H). A parameterization for the periodic system follows then directly from the parameterization of the multivariable system (F,G,H). At once we see that even a scalar periodic system leads to multivariable equivalent systems. The restriction to scalar systems mentioned at the onset is thus not restrictive, but permits simpler notation and examples.

We indicate some particular results which will be useful in the realization problem:

Theorem 1. The minimal realizations $(\hat{A}_i,\hat{B}_i,\hat{C}_i)$ of the Adiabatic Hankel matrices $\hat{H}_i$ have the property that $\det(zI-\hat{A}_i)$ divides $(z^N I-A_0A_1\ldots A_{N-1})$.

Proof (for N=2). The Hankel matrix $\hat{H}_0$ is obtained from the pulse response $\{h_{i,0}\}$. Its Z-transform equals

$$\begin{aligned} H_0(z) &= C_0(z^2I-A_1A_0)^{-1}A_1B_0 + C_1(z^2I-A_0A_1)^{-1}zB_0 \\ &= [N_0(z^2)+zN_1(z^2)]/\det(z^2I-A_0A_1) \end{aligned}$$

for some polynomial matrices N_0 and N_1. Clearly then the minimal realizations of H_0 and H_1 have the above stated property. ●

In fact, it is easy to show that the realizations of $\hat{H}_0$ and $\hat{H}_1$ msut be very closely related. Indeed, by writing $\hat{H}_1$ in the form

$$\begin{aligned} \hat{H}_1 &= [C_1,C_0]\begin{bmatrix} z^2I-A_0A_1 & 0 \\ 0 & z^2I-A_1A_0 \end{bmatrix}^{-1}\begin{bmatrix} A_0 \\ zI \end{bmatrix} B_1 \\ &= [C_0,C_1]\begin{bmatrix} z^2I-A_1A_0 & 0 \\ 0 & z^2I-A_0A_1 \end{bmatrix}^{-1}\begin{bmatrix} zI \\ A_0 \end{bmatrix} B_1 \end{aligned}$$

The first factors on the left also appear in the expansion of $\hat{H}_0$

$$\hat{H}_0 = [C_0,C_1]\begin{bmatrix} z^2I-A_1A_0 & 0 \\ 0 & z^2I-A_0A_1 \end{bmatrix}^{-1}\begin{bmatrix} A_1 \\ zI \end{bmatrix} B_0$$

Hence, if we define the following transfer matrix:

$$\hat{H}_{01} = [C_0,C_1]\begin{bmatrix} z^2I-A_1A_0 & 0 \\ 0 & z^2I-A_0A_1 \end{bmatrix}^{-1}\begin{bmatrix} A_1 & zI \\ zI & A_0 \end{bmatrix}$$

then $\hat{H}_0=\hat{H}_{01}[B_0',0]'$ and $\hat{H}_1=\hat{H}_{01}[0,B_1']'$.

This observation leads directly to the following theorem:

Theorem 2. There exists an observable pair $(\bar{A},\bar{C})$ and matrices $\bar{B}_0$ and $\bar{B}_1$ such that $(\bar{A},\bar{B}_0,\bar{C})$, and $(\bar{A},\bar{B}_1,\bar{C})$ realize, respectively, the Adiabatic transfer matrices $\hat{H}_0$ and $\hat{H}_1$.

Proof. Let $(\bar{A},\bar{B},\bar{C})$ be a minimal (observable is sufficient) realization for $\hat{H}_{01}$, then $\bar{B}_0=\bar{B}[B_0',0]'$, and $\bar{B}_1=\bar{B}[0,B_1']'$. ●

The importance of this theorem lies in its use to find the realizations for an alternating system [6]. Given the pulse response sequences $\{h_{i,0}\}$ and $\{h_{i,1}\}$, we can use the realizations of either sequence. By the theorem, these realizations can be extended by addition of uncontrollable states if necessary, to observable realizations with the same A and C matrix.

3. REACHABILITY, OBSERVABILITY, AND STABILITY

Definitions:

- The N-periodic system $\{\Sigma_0,\Sigma_1,\ldots,\Sigma_{N-1}\}$ is said to be uniformly p-reachable (reachable in p steps), iff every state can be reached in p steps, independently of the starting event (= initial time and initial state). The system is said to be uniformly reachable, iff there exists a p>0, such that it is uniformly p-reachable.

- The system is said to be uniformly observable in p steps iff the initial state x_j can be uniquely determined from p consecutive outputs $\{y_j,\ldots,y_{j+p-1}\}$, independently of the starting time j. The system is said to be uniformly observable iff it is p-observable for some p.

Theorem 3. The period-N system $\{\Sigma_0,\ldots,\Sigma_{N-1}\}$ is uniformly reachable iff the reachability matrices (3) have full rank for all j. The system is uniformly observable iff the observability matrices (4) have full rank for all j.

The proof is easily established by a standard argument [5]. Since the Adiabatic systems of, at most, order nN provide an underlying time-invariant structure in the problem, at most nN steps, need to be considered for checking uniform reachability and observability, by virtue of the Cayley-Hamilton Theorem. Some direct corollaries of the theorem are:

(1) The Cyclicially Augmented System (A_{ca},B_{ca},C_{ca}) is reachable iff the period-N realization $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_N\}$ is uniformly reachable.

(2) $\{\Sigma_1,\Sigma_2,\ldots,\Sigma_0\}$ is uniformly observable (reachable) iff $\{\Sigma_0,\Sigma_1,\ldots,\Sigma_{N-1}\}$ is uniformly observable (reachable), whence the invariance of uniform observability and reachability under a cyclic shift.

(3) Using the backward propagation, we can write the output at time i in terms of the previous inputs, i.e. we look at $\{h_{i,j}\}$ for fixed i, and define the equivalent stationary systems with the above A matrix and $C=[0,\ldots,0,C_i,0,\ldots,0]$, the nonzero block occurring in the i-th block position, and $B=B(B_0',B_1',\ldots,B_{N-1}')'$. We then have the "duality" property:

$\{\Sigma_1,\ldots,\Sigma_{N-1},\Sigma_N\}$ is uniformly observable

iff

$\{\Sigma_N^d,\Sigma_{N-1}^d,\ldots,\Sigma_1^d\}$ is uniformly reachable,

where the "dual" system is obtained by time reversal of the sequence of the duals Σ_i^d of the realizations Σ_i, where $(A_i,B_i,C_i)^d$ is the triple (A_i',C_i',B_i'). We are thus led to the definition:

$$\{\Sigma_1,\Sigma_2,\ldots,\Sigma_0\}^{dual} = \{\Sigma_0^d,\Sigma_{N-1}^d,\ldots,\Sigma_1^d\} \qquad (8)$$

Finally, we remark that if all A_i are nonsingular, as for instance in the important case of the discretization of a continuous system, the criterion of Theorem 1 can be simplified by virtue of the following:

Lemma: If the A_j are nonsingular for all j, then the full rankness of one of the reachability matrices R_i (observability matrices O_i) implies the full rankness of all others, and hence reachability (observability).

As an example, a siso alternating system $\{\Sigma_0,\Sigma_1\}$ will be uniformly reachable iff the stationary systems $(A_1,A_0,[b_1,A_1,b_0])$ and $(A_0,A_1,[b_0,A_0,b_1])$ are reachable. If the system is uniformly reachable, no more than 2n steps are required to reach any desired endstate. If the product A_0A_1 is nonsingular, then $(A_1,A_0,[b_1,A_1,b_0])$ and $(A_0,A_1,[b_0,A_0,b_1])$ are either both reachable or both nonreachable. By applying inputs before 0, one gets the reachability relation at time 0:

$$x_0 = R_1[u_0,u_{-1},\ldots]'$$

where $R_1=[b_1,A_1b_0,A_1A_0b_1,\ldots]$ is the time-varying reachability matrix [3]. Observation of the output sequence after time 0, with no input applied, leads then to the observability relation:

$$[y_0,y_1,y_2,\ldots]' = O_0x_0$$

where $O_0=[c_0',A_0'c_1',A_0'A_1'c_0',\ldots]'$ is the time-varying observability matrix. Similarly, we construct the reachability and observability matrices, R_0 and O_1, relating to the reference time 1. The products O_0R_1 and O_1R_0 are then the alternating (period-2) Hankel matrices defined in (2).

We also have the following important stability theorem:

Theorem 4. The N-period system (0) is stable if the eigenvalues of the product $A_0A_1\ldots A_{N-1}$ have modulus less than 1.

Proof. The convergence properties of the periodic systems are determined by the convergence properties of the equivalent time-invariant system (A_{ca},B_{ca},C_{ca}). The latter is completely determined by the characteristic polynomial $\det(z^N I-A_1A_2A_3\ldots A_N)=0$. ●

The problem with this approach is that the resulting time-invariant system has order Nn if n is the order of the individual realizations Σ_i. The original periodic system is only of n-th order, so that a "hidden modes" phenomenon occurs.

4. CANONICAL FORM, PARAMETERIZATION AND TOPOLOGICAL STRUCTURE

The first object in this study is to find the transformations on the realizations that leave the input-output behavior (i.e. all Adiabatic transfer matrices and the periodic system "Hankel" matrices (2)) invariant.

Let $\{\Sigma_0,\Sigma_1,\ldots,\Sigma_{N-1}\}$ be a realization of an N-period system. Denote an element of the group $Gl_n(R)^N$, denoted by Gl_n^N for short, by $(P_0,P_1,\ldots,P_{N-1})$. The group action is defined by

$$(P_0,\ldots,P_{N-1}) : \{(A_0,B_0,C_0),\ldots,(A_{N-1},B_{N-1},C_{N-1})\} \longrightarrow \qquad (9)$$

$$\{(P_1A_0P_0^{-1},P_1B_0,C_0P_0^{-1}),\ldots,(P_0A_{N-1}^{-1}P_{N-1}^{-1},P_0B_{N-1},C_{N-1}P_{N-1}^{-1})\}$$

The states transform as

$$\begin{aligned} x_{Nk} &\longrightarrow P_0x_{Nk} \\ x_{Nk+1} &\longrightarrow P_ix_{Nk+i} \end{aligned} \qquad i = 1,\ldots,N-1 \qquad (10)$$

The following property is readily shown:

Theorem 5. Equivalence of State Space Representations. The product group $Gl_n(R)^N$ action on the set of period-N systems leaves the I/O properties invariant.

Once the symmetry group (i.e. the group whose action leaves the I/O behavior invariant) is established, we can look at the question of canonical forms: Any property of the original system can be described as a map from the set of systems Σ, to some suitable set S. After introducing canonical forms, the study of the original function f is then replaced by the study of some "simpler" function $\hat{f}:C \longrightarrow S$, such that $f=\hat{f}\circ\pi$, where π is the canonical projection $\pi:\Sigma \longrightarrow C$ on the set of canonical forms.

For notational simplicity, the rest of this section will be restricted to alternating systems. The above development (N=2) should give enough insight to realize that the general principles remain the same. Canonical forms for the uniformly reachable systems are obtained by the usual Kronecker selection procedure, i.e. among the 2n columns of R_0, select n linear independent ones, which form a basis $\{\beta_0\}$ for the state space. In particular, a unique "nice" selection may be chosen according to the Young or "crate" diagram. Similarly, let $\{\beta_1\}$ be another basis, chosen by a nice selection among the columns of R_1. Now express the system with respect to the basis which is alternating between $\{\beta_0\}$ and $\{\beta_1\}$, e.g. A_0 is represented by the new matrix whose j-th column is the representation in terms of the basis $\{\beta_1\}$ of A_0 operating on the i-th basis vector from the other basis $\{\beta_0\}$.

The effect is that the new representation is of the form $b_0=b_1=[1,0,\ldots,0]'$. (By assumption of reachability, neither b_0 nor b_1 are zero.) If $v_i(k)$ denotes the position of the k-th basis vector from R_i, then we refer to the sequences $w_i=\{v_i(1),\ldots,v_i(n)\}$ as a multi-index. The k-th column of the new A matrices are

$$A_0^r e_k = R_0^r e_{v1(k)+1}$$

$$A_1^r e_k = R_1^r e_{v0(k)+1}$$

However straightforward the previous extension of the known scheme may be, a particular nice form is obtained as follows if A_1 is nonsingular. Search the columns of R_0 in their natural order, i.e. from left to right, and reordered in chains, as in the usual "scheme II" search [5]. Now observe that if A_1 is nonsingular, then the result of A_1 operating on the above basis is also a basis, and in fact, each of these new basis vectors will be a column in R_1, except perhaps for the last new basis vector. In that case, it may be substituted for b_1 as new last basis vector. Note that this all corresponds to a scheme II search in R_1, but STARTING AT A_1b_0. It follows that, unless there was a full length (=n) chain $\{A_0b_1,\dots,(A_0A_1)^{n-1}A_0b_1\}$ in R_0 (which only happens if b_0 is zero), the operator A_1 is represented by I. The b_1 vector is full in general. The pair (A_0,b_0) has a canonical controllability form [5] representation. In the latter special case, it follows that A_0 is represented by a cyclically down shifted identity matrix, and A_1 by a cyclically down shifted right companion (controllability canonical form) matrix. The new b_0 (obviously) remains zero, and the new b_1 is a full vector. As the c_0 and c_1 have no particular structure, there are 4n free parameters in this canonical form. By analogy to the stationary realizations, we shall refer to this form as the controllability canonical form.

Note that because the search extends over 2n columns, the new A matrices will, in general, not be in the usual companion form themselves, unless the system is uniformly reachable in n steps, but then this is also the case with the time-invariant multivariable systems. In fact, it is exactly because of such a reduction from the time-varying to the multivariable time-invariant case that all the topological properties of these systems are expected to carry over. In particular, for multivariable systems, we have:

Theorem 6. The orbit space of the reachable systems is an analytic manifold, which can be decomposed into generalized Kronecker cells which form a cellular patch complex. The state space canonical forms act as local coordinates.

The number of canonical forms that is required to cover the space of all reachable alternating systems is also equal to the number of pairs of nice multi-indices that can be chosen.

5. OPERATIONAL TRANSFER FUNCTION

The essential ingredient is the introduction of a sampling operator, π_N, taking sequences into sequences, defined via

$$\pi_N\{u_i\ ;\ i\geqslant 0\} \longrightarrow \{y_i\ ;\ i\geqslant 0\}$$

$$\begin{aligned} y_{Nk} &= u_{Nk} \\ y_{Nk+i} &= 0 \qquad \text{for } i \in \{1,\dots,N-1\} \end{aligned}$$

Clearly, $\pi_N^2=\pi_N$ so that π_N is a projection operator. Letting $U(z)$ denote the usual Z-transform of the sequence u(z), then π_N induces an operator in the Z-domain which we shall denote, with a slight abuse, by the same notation π_N. Note that then

$$\begin{aligned} \pi_N z^{-Nk-i} &= 0 \qquad \text{for } i \in \{1,\dots,N-1\} \\ &= z^{-Nk} \qquad \text{if } i = 0 \end{aligned}$$

The space of formal power series in z^{-1} can then be decomposed into N orthogonal subspaces, each of which induces in turn another projection operator. The set of subspaces and the set of projection operators are isomorphic structures. Thus, define $\pi_N^{(i)}$ by $z^{-i}\pi_N z^i$ for $i \in \{1,\dots,N-1\}$. It follows at once that these operators are all generated by z and π_N, clearly though, the operator algebra will be a noncommutative one. The union of all these subspaces is the whole space, so that we have the relator

$$1 = \pi_N + z^{-1}\pi_N z + z^{-2}\pi_N z^2 + \dots + z^{-N+1}\pi_N z^{N-1}$$

or, equivalently

$$z^{N-1} = z^{N-1}\pi_N + z^{N-2}\pi_N z + \dots + z\pi_N z^{N-2} + \pi_N z^{N-1} .$$

Thus the above can be formalized as follows: Period-N systems of order n can be represented by an n-th order realization $(A(\pi_N),B(\pi_N),C(\pi_N))$, whose coefficients are in the multivariable polynomial quotient ring

$$R[\pi_N^{(i)}\ ;\ i\varepsilon\{1,\dots,N\}]/(\pi_N^{(i)}p_N^{(j)}=\pi_N^{(i)}\delta_{ij}) .$$

The periodic state space realization equations (0) are then transformed to

$$\pi_N^{(i+1)}x = A_i\pi_N^{(i)}x + {}_iB_N\pi^{(i)}u \qquad \text{for } i \neq N-1$$

$$\begin{aligned} \pi_N^{(0)}x &= A_{N-1}\pi_N^{(N-1)}x + B_{N-1}\pi_N^{(N-1)}u \\ \pi_N^{(i)}y &= C_i\pi_N^{(i)}x \end{aligned}$$

Upon Z-transforming, we find

$$z\pi_N^{(i+1)}X(z) = A_i\pi_N^{(i)}X(z) + B_i\pi_N^{(i)}U(z) \qquad \text{for } i \neq N-1$$

$$z\pi_N^{(0)}X(z) - z^N x_0 = A_{N-1}\pi_N^{(N-1)}X(z) + B_{N-1}\pi_N^{(N-1)}U(z)$$

$$\pi_N^{(i)}Y(z) = C_i\pi_N^{(i)}X(z)$$

Note the appearance of the initial condition (x_0) term. Adding the left hand sides, taking account of the above relation between the projection operator, yields

$$\begin{aligned} zX(z) - z^N x_0 = {} & A(\pi_N^{(1)},\dots,\pi_N^{(N)})X(z) \\ & + B(\pi_N^{(1)},\dots,\pi_N^{(N)})U(z) \end{aligned}$$

$$Y(z) = C(\pi_N^{(1)},\dots,\pi_N^{(N)})X(z)$$

Hence, we get, assuming zero initial conditions, and substituting $\pi_N^{(i)}$ by the combination $z^{-i}\pi_N z^i$, the Operational Transfer Function (OTF)

$$H(z,\pi_N) = C(z,\pi_N)[zI-A(z,\pi_N)]^{-1}B(z,\pi_N)]$$

where now very simply:

$$A(z,\pi_N) = A_0\pi_N + A_1z^{-1}\pi_N z + \dots A_{N-1}z_{-N+1}\pi_N z_{N-1}$$

$$B(z,\pi_N) = B_0\pi_N + B_1z_{-1}\pi_N z + \dots B_{N-1}z^{-N+1}\pi_N z^{N-1}$$

$$C(z,\pi_N) = C_0\pi_N + C_1z_{-1}\pi_N z + \dots C_{N-1}z_{-N+1}\pi_N z_{N-1}$$

In order to illustrate the ideas for period-2 systems (N=2), we have: $z\pi+z(1-\pi)=z$. The equation $x_{2k+1}=A_0x_{2k}+B_0u_k$ is then transformed to $z(1-\pi)X(z)=A_0\pi X(z)+B_1\pi U(z)$. Similarly, the complementary equation transforms to $z\pi X(z)=A_1(1-\pi)X(z)+B_1(1-\pi)U(z)$. Addition yields the form $X(z)=(zI-A(\pi))^{-1}B(\pi)U(z)$, where $A(\pi)=A_0\pi+A_1(1-\pi)$. Using the noncommutative relation, this formalism is very helpful in deriving all sorts of results and transfer function operations.

For period-2 systems, some of the ideas on Operational Transfer Matrix Reductions were explored. For instance, connections (parallel, series, and feedback) can be performed with the same formal rules as for stationary systems, as long as the noncommutativity is taken into account during the reduction.

We shall here also explore the possibility of connecting systems of DIFFERENT periodicity. So one system may have the OTF $H_1(z,\pi_N)$ and another $G(z,\pi_M)$. The series connection is then simply $G(z,\pi_M)H(z,\pi_N)$. Clearly, the combination involves now three generators: z, π_N, and π_M. So one needs to define the composite transfer matrix as a rational division ring (Noncommutative Field) extension of the polynomial ring with three generators (z,π_N,π_M). Clearly additional commutation rules (relators in the division ring) need to be invoked. This work is, as of this writing, in progress and will be reported in the final version of this paper.

Finally, we report that some interesting realization related properties can be developed from within the OTF framework as well. In particular (for N=2):

5.1 Reachability Problem via the OTF

With zero initial conditions, and input sequence $\{u_k\}$, the Z-transform of the state sequence is given by

$$X(z) = [zI-A(\pi)]^{-1}B(\pi)U(z)$$

$$= z^{-1}[I-A(\pi)z^{-1}]^{-1}B(\pi)U(z)$$

$$= z^{-1}[I+A(\pi)z^{-1}+\ldots+(A(\pi)z^{-1})^kB(\pi)z^{-k}+\ldots]U(z)$$

Noting that the commutation of π and z^{-1} involves an involution, i.e.

$$z^{-1}A(\pi) = A(1-\pi)z^{-1} ,$$

we get the series expansion

$$zX(z) = [B(\pi)+A(\pi)B(1-\pi)z^{-1}+A(\pi)A(1-\pi)B(\pi)+\ldots]U(z)$$

$$= [B(\pi),A(\pi)B(1-\pi),A(\pi)A(1-\pi)B(\pi),\ldots]\begin{bmatrix}U(z)\\ z^{-1}U(z)\\ z^{-2}U(z)\\ \vdots\end{bmatrix}$$

$$= R(\pi)U(z)$$

The operator reachability matrix decomposes into two parts:

$$R(\pi) = \pi R_0 + (1-\pi)R_1$$

where, in terms of the system components:

$$R_0 = [B_1,A_1B_0,A_1A_0B_1,A_1A_0A_1B_0,\ldots]$$

$$R_1 = [B_0,A_0B_1,A_0A_1B_0,A_0A_1A_0B_1,\ldots]$$

As the operators π and $1-\pi$ select complimentary parts of the vector $U(z)$, we find for the condition of reachability that both R_0 and R_1 should have full rank. We shall say that then the operational reachability matrix $R(\pi)$ has full rank, so that the usual criterion for reachability is retrieved.

For instance, upon identifying the coefficients of z^{-3}, we obtain

$$x_4 = A_1A_0A_1B_0u_0 + A_1A_0B_1u_1u_1 + A_1B_0u_2 + B_1u_3$$

5.2 The Observability Problem via the OTF

Here, the inputs are zero, and a nonzero initial condition x_0 is assumed. The system output is given in the transform domain by

$$Y(z) = C(\pi)[zI-A(\pi)]^{-1}x_0$$

$$= C(\pi)[I-z^{-1}A(\pi)]^{-1}z^{-1}x_0$$

$$= C(\pi)[I+z^{-1}A(\pi)+[z^{-1}(A\pi)]^2+\ldots]z^{-1}x_0$$

$$= C(\pi)z^{-1}x_0 + C(\pi)A(1-\pi)z^{-2}x_0 + \ldots$$

$$= O(\pi)X_0(z)$$

where $O(\pi)=O_0\pi+O_1(1-\pi)$. As for the reachability problem, the matrix $O(\pi)$ is said to have full rank if both O_0 and O_1 have full rank. The condition for observability follows them as from the full rankness of $O(\pi)$.

6. REALIZATION

In this section we show how the above results can lead to a relatively simple realization algorithm for periodic systems. We shall assume that the order n of the minimal N-period system is known. There is then an underlying uniformly reachable system $\{\Sigma_0,\ldots,\Sigma_{N-1}\}$ of order n. The realization is given is 4 steps:

Step 1. Let the pulse responses $\{h_{i,0};i>1\},\ldots,\{h_{i,N-1};i>N\}$ be collected. With this data, the Adiabatic Hankel matrices H_j, $j=0,\ldots,N-1$ are formed. There are now two possible routes: realize each Adiabatic system separately, or realize the composite system with transfer matrix $[H_0(z),\ldots,H_{N-1}(z)]$.

Step 2. Obtain a minimal realization $(A_m,[B_{m,0},B_{m,1},\ldots,B_{m,N-1}],C_m)$ of the system with composite transfer matrix $[H_0(z),\ldots,H_{N-1}(z)]$, where $H_i(z)$ is the transfer matrix of the Adiabatic system (A_i,B_i,C_i).

There are two ways in which this can be obtained. The first is to reduce the composite transfer matrix, and obtain a minimal realization of it, by standard multivariable techniques [5]. The second method consists in first obtaining minimal realizations (A_i,B_i,C_i) of the Adiabatic systems. These minimal realizations may be of different dimensions. However, by Theorem 1, there exists a maximal characteristic polynomial, of order q say, in the sense that the characteristic polynomials of the other realizations divide it, and each nonminimal Adiabatic realization can be extended by adding a nonreachable, but observable subsystem so that all extended realizations of the Adiabatic subsystems have the same order (=q), and the same A and C matrix. The composite realization $(A_m,[B_{m,0},\ldots,B_{m,N-1}],C_m)$ is then the desired form and is minimal since (A_m,C_m) is observable, and at least one of the $B_{m,i}$ forms together with A_m a reachable pair.

Step 3. Extend the realization of order q, obtained in Step 2, to one whose order is a multiple of N by adding a nonobservable but reachable subsystem. Indeed, since the minimal system $(A_m,[B_{m,0},\ldots,B_{m,N-1}],C_m)$ and the cyclically augmented system (A_{ca},B_{ca},C_{ca}) realize the same transfer matrix, and since the cyclically augmented system is uniformly reachable by assumption, the latter must be an augmentation (A_e,B_e,C_e) of $\left(A_m,[B_{m,0},\dots,B_{m,N-1}],C_m\right)$ by a nonobservable but reachable subsystem of order Nn-q. This implies the existence of matrices $X,Y,Z_0,\dots,Z_{N-1}$ so that

$$\begin{bmatrix} A_m & 0 \\ X & Y \end{bmatrix} \begin{bmatrix} B_{m0},\dots,B_{m,N-1} \\ Z_0,\dots,Z_{N-1} \end{bmatrix} \quad [C_m \quad 0]$$

is similar to the cyclically augmented system. This augmentation must not impair the reachability of the realization. The necessary reachability of the subsystem implies that none of the rows of the matrix $[X,Z_0,\dots,Z_{N-1}]$ can be zero. This follows easily by contradiction. If $[X,Z_0,\dots,Z_{N-1}]$ had a zero row, then the realization could be partitioned as:

$$\begin{bmatrix} A_m & 0 & 0 \\ X_1 & X_2 & Y_1 \\ 0 & 0 & Y_2 \end{bmatrix} \quad \begin{bmatrix} B_{m0},\dots,B_{m,N-1} \\ Z_0,\dots,Z_{N-1} \\ 0,\dots,0 \end{bmatrix}$$

which has the unreachable subsystem $[Y_2,0,C_m]$.

The X,Y, and Z are chosen so that the reachability matrices

$$R\left(A_e^N;[B_{e0},A_eB_{e1},\dots,A_e^{N-1}B_{e,N-1}]\right),$$

$$R\left(A_e^N;[B_{e1},A_eB_{e2},\dots,A_e^{N-1}B_{e,0}]\right),\dots,$$

$$R\left(A_e^N;[B_{e,N-1},A_eB_{e0},\dots,A_e^{N-1}B_{e,N-2}]\right)$$

all have rank less than n.

Step 4. Determine the similarity transformation that transforms the extended system $\left(A_e,[B_{e1},\dots,B_{e,N-1}],C_e\right)$ to a cyclic form. For notational convenience, we shall again discuss the latter for alternating (i.e. period-2) systems. The ideas for period-N systems are similar.

The reachability matrix for the cyclically augmented matrix has the structure

$$R_{ca} = \begin{bmatrix} B_0, & 0, & 0, & A_0B_1, & A_0A_1B_0, & 0 \\ 0, & B_1, & A_1B_0, & 0, & 0, & A_1A_0B_1 \end{bmatrix}$$

whereas R_e has entries in all positions in general. The desired similarity maps R_e into $TR_e=R_{ca}$. Partitioning T into $[T_1',T_2']'$, it is seen that the zero locations in the above equation leads to the identities:

$$T_1R\left(A_{e2};[B_{e1},A_eB_{e0}]\right) = 0$$

$$T_2R\left(A_{e2};[B_{e0},A_eB_{e1}]\right) = 0$$

where the reachability matrices extend to n (block) columns only, and are thus square for single input periodic systems. On the condition that the test matrices

$$R_{e0}=R(A_e^2;[B_{e0},A_eB_{e1}]),$$

and

$$R_{e1}=R(A_e^2;[B_{e1},A_eB_{e0}]),$$

have rank less than n, it is possible to find n linearly independent row vectors in the left nullspaces of the above reachability matrices. Since the overall realizations are both reachable, there exists T_1 and T_2 such that $T=[T_1',T_2']'$ is nonsingular with T_1 and T_2 satisfying the above conditions.

We summarize then with

Theorem 7. An invertible transformation T can always be found, bringing the extended system (F,g) of order 2n to the cyclically augmented form Σ_{ca} (5) of the same order if it is reachable and if the reachability matrices $F\left(F_2,[g_0,Fg_1]\right)$ and $F\left(F_2,[g_1,Fg_0]\right)$ both have rank at most equal to n.

Theorem 8. The realization of $\left[H_0(z),H_1(z)\right]$ can always be augmented so that the extended system satisfies the conditions of Theorem 4.

Some simple illustrative examples are given in [6].

Remarks.

1. The minimal Adiabatic realizations completely specify the I/O behavior of the alternating system, and may therefore lead to new canonical representations for such systems.

2. Other identification methods for periodic systems exist. One can "stagger" the impulse responses, by looking at every N-th sample for which the system looks like a time-invariant one. However, the solution for the individual realizations in $(\Sigma_1,\dots,\Sigma_N)$ require difficult nonlinear equation solvers. Furthermore, since the data samples with such a scheme are not "convoluted," a large number of data needs to be collected (roughly 2nN) before the system starts to unfold. The scheme presented here already presents a lot of information about the system after 2n steps, and is therefore more "holographic."

ACKNOWLEDGEMENT

This research is supported by the U.S. Air Force under Contract No. AFOSR-87-0308.

REFERENCES

[1] U. Helmke, "Parameterizations for Multi-Mode Systems with Yang-Mills Instantons," Proc. 25th Conf. on Decision and Control, Athens, Greece, December 1986.

[2] D.P. Stanford, L.T. Conner, "Controllability and Stabilizability in Multi-Pair Systems," SIAM J. Control and Optimization, vol. 18, no. 5, p. 488-497, September 1980.

[3] E.W. Kamen, P.P. Khargonekar, "A Transfer Function Approach to Linear Time-Varying Discrete Time Systems," 23d Conf. on Decision and Control, Orlando, Florida, December 1982.

[4] M. Hazewinkel, "Moduli and Canonical Forms for Linear Dynamical Systems: The Topological Case," Mathematical Systems Theory, vol. 10, pp. 363-385, 1977.

[5] T. Kailath, Linear Systems, Prentice-Hall, 1980.

[6] E.I. Verriest, "Alternating Discrete Time Systems: Invariants, Parameterization and Realization," Proc. Annual Conf. on Information Sciences and Systems, Princeton, March 1988, pp. 952-957.