

AD-A194 685

APPROXIMATIONS AND IMPLEMENTATIONS OF NONLINEAR
FILTERING SCHEMES(U) GEORGIA INST OF TECH ATLANTA
SCHOOL OF ELECTRICAL ENGINEERING A H HADDAD ET AL.

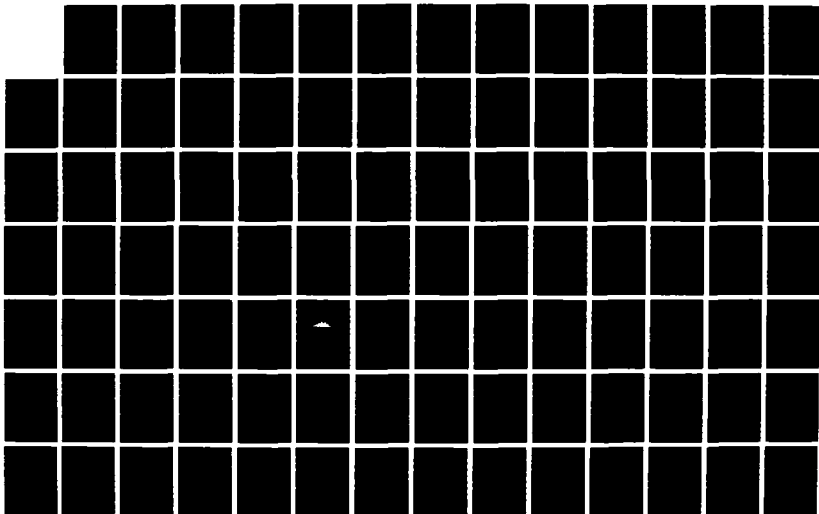
1/4

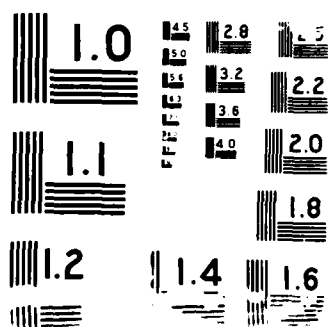
UNCLASSIFIED

FEB 88 AFATL-TR-87-73 F08635-84-C-0273

F/G 12/3

NL





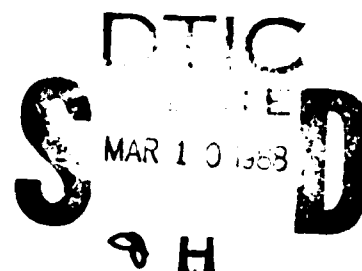
AFATL-TR-87-73

Approximations and Implementations of Nonlinear Filtering Schemes

AD-A194 685

A H Haddad
E I Verriest

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30322-0250



FEBRUARY 1988

FINAL REPORT FOR PERIOD AUGUST 1984-SEPTEMBER 1987

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

AIR FORCE ARMAMENT LABORATORY

Air Force Systems Command ■ United States Air Force ■ Eglin Air Force Base, Florida

ERRATA

AFATL-TR-87-73

APPROXIMATIONS AND IMPLEMENTATIONS OF NONLINEAR
FILTERING SCHEMES

AIR FORCE ARMAMENT LABORATORY
EGLIN AIR FORCE BASE, FLORIDA
32542-5434

1. Replace initial page 12 (Bibliography) with attached page 12 (Bibliography).
2. This errata is unclassified.


JAMES E. KRUG II
Chief, Technical Reports Section

BIBLIOGRAPHY

Ackerson, G. A. and Fu, K. S., "On State Estimation in Switching Environment," IEEE Trans. on Automatic Control, Vol. AC-15, pp. 10-17, February 1970.

Johnson, T. L., "Synchronous Switched Linear Systems," Proc. 24th IEEE Conf. on Decision and Control, Ft. Lauderdale, pp. 1699-1700, December 1985.

Ladde, G. S. and Siljak, D. D., "Multiplex Control Systems: Stochastic Stability and Dynamic Reliability," Int. J. Control, Vol. 38, pp. 515-524, 1983.

Mullis, C. T. and Roberts, R. A., "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," IEEE Trans. Circuits and Systems, Vol. CAS-23, pp. 551-562, 1976.

Tugnait, J. K., "Detection and Estimation for Abruptly Changing Systems," Automatica, Vol. 18, pp. 607-615, September 1982.

Willems, J. C., "From Time Series to Linear Systems," Automatica, 1987.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A174655

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for Public Release; distribution unlimited		
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S) AFATL-TR-87-73		
6a. NAME OF PERFORMING ORGANIZATION Georgia Institute of Technology		6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION Guidance and Control Branch Aeromechanics Division	
6c. ADDRESS (City, State, and ZIP Code) School of Electrical Engineering Atlanta, GA 30332-0250			7b. ADDRESS (City, State, and ZIP Code) Air Force Armament Laboratory Eglin AFB, FL 32542-5434		
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Aeromechanics Division		8b. OFFICE SYMBOL (If applicable) AFATL/FXG		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F08635-84-C-0273	
8c. ADDRESS (City, State, and ZIP Code) Air Force Armament Laboratory Eglin AFB, FL 32542-5434			10. SOURCE OF FUNDING NUMBERS		
			PROGRAM ELEMENT NO. 2304	PROJECT NO. E1	TASK NO. 37
11. TITLE (Include Security Classification) Approximations and Implementations of Nonlinear Filtering Schemes					
12. PERSONAL AUTHOR(S) A. H. Haddad and E. I. Verriest					
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM Aug 84 to Sep 87		14. DATE OF REPORT (Year, Month, Day) February 1988	
				15. PAGE COUNT 378	
16. SUPPLEMENTARY NOTATION Availability of report is specified on verso of front cover					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Markov Approximation Filters Finite Word Length Effects Hybrid Systems Models		
FIELD	GROUP	SUB-GROUP			
19. ABSTRACT (Continue on reverse if necessary and identify by block number) The objective of this program is to develop approximate models that may be used for the implementation of nonlinear filtering schemes. The approach is based on two major concepts. The first is concerned with the approximation of nonlinear models for air-to-air missile scenarios by a Markov linear approximate model. The second concept is concerned with models for the errors due to the finite word-length implementation of the filters. In addition, the project is involved with the study of hybrid system models that result when nonlinear or linear models such as these are quantized and are subject to switches among a set of linear models. The switching can either be deterministic or stochastic and may involve fast and slow dynamics.					
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified		
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. J. Cloutier			22b. TELEPHONE (Include Area Code) (904) 882-2961		22c. OFFICE SYMBOL AFATL/FXG

PREFACE

This is the final report for Contract No. F08635-84-C-0273 entitled "Approximations and Implementations of Nonlinear Filtering Schemes" with the U. S. Air Force Armament Laboratory at Eglin Air Force Base, Florida 32542. The work was performed at the School of Electrical Engineering of the Georgia Institute of Technology in Atlanta, Georgia 30332-0250.

The work reported herein was performed during the period 11 August 1984 to 30 September 1987, under the direction of Dr. A. H. Haddad and Dr. E. I. Verriest. Dr. J. Cloutier (AFATL/FXG) managed the program for the Air Force Armament Laboratory.

Several graduate students at the Georgia Institute of Technology contributed to this work and their work is hereby acknowledged by the authors. The students are: J. Ezzine, S. Gray, B. Heck, M. Ingram, M. Jose, and P. West. The support and guidance of Mr. Johnny Evers and Dr. J. Cloutier of the Air Force Armament Laboratory is greatly appreciated.



iii/iv (Blank)

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Avail and/or	
Dist	Special
A-1	

TABLE OF CONTENTS

Section	Title	Page
I	INTRODUCTION.	1
II	MARKOV APPROXIMATION FILTERS.	2
III	FINITE WORD LENGTH EFFECTS.	4
IV	HYBRID SYSTEMS MODELS	6
V	CONCLUSIONS	8
	REFERENCES.	9
	BIBLIOGRAPHY.	12
Appendix		
A	Linear Markov Approximations for Piecewise Linear Stochastic Systems	13
B	Linear Markov Approximations of Piecewise Linear Stochastic Systems	21
C	Piecewise Linear Modeling of Multidimensional Stochastic Nonlinear Systems	55
D	On the Modeling and Filtering for Piecewise Linear Stochastic Systems.	59
E	Approximate Nonlinear Filters for Piecewise Linear Models . .	67
F	Approximate Nonlinear Filtering for Piecewise Linear Systems	73
G	Linear Filters for Linear Systems with Multiplicative Noise and Nonlinear Filters for Linear Systems with Non-Gaussian Additive Noise.	85
H	Stochastic Reduced Order Modeling of Deterministic Systems. .	91
I	Uncertainty Equivalent Reduced Order Models for Discrete Systems	95
J	Model Reduction via Balancing, and Connections with Other Methods	99
K	Reduced Order LQG Design: Conditions for Feasibility.	133

TABLE OF CONTENTS (CONCLUDED)

Appendix	Title	Page
L	On Redefining the Optimal Least Squares Filter Under Floating Point Operations.	141
M	Error Analysis of Linear Recursions in Floating Point	147
N	A Bilinear Model for Linear Recursive Computations Using Floating Point Arithmetic	153
O	Gain Correction in Optimal Filtering Using Floating Point Arithmetic.	161
P	Subspace Correlation Measures and Applications to the Stochastic Realization Problem.	165
Q	A Unified RV-Coefficient Approach for Solving the Covariance based Stochastic Realization Problem.	183
R	A Note on the Cross Riccatian and Related Properties for Symmetric Stochastic Realizations	233
S	Projection Techniques for Model Reduction	249
T	Optimality Properties of Balanced Realizations: Minimum Sensitivity	267
U	Robust Design Problems: A Geometric Approach.	275
V	On Singularly Perturbed Switched Parameter Systems.	295
W	On Linear Singularly Perturbed Systems with Quantized Control	299
X	Singular Perturbation in Piecewise Linear Systems	307
Y	On the Controllability and Observability of Hybrid Systems. .	329
Z	Optimal and Suboptimal Filtering for Linear Systems Driven by Self-Excited Poisson Processes.	357

LIST OF FIGURES

Figure	Title	Page
B-1	The Distribution of the Stationary Density for $\sigma = 3$	53
E-1	Block Diagram of the Filter Stages.	71
F-1	Block Diagram of the Filter Stages.	81
F-2	System and Measurement Model for the Scalar Case.	82
F-3	Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = .5$, $H_0 = 5$, $H_1 = -.1$	82
F-4	Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = .5$, $H_0 = 5$, $H_1 = .1$	83
F-5	Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = 2$, $H_0 = 1$, $H_1 = -.1$	83
F-6	Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = 2$, $H_0 = 1$, $H_1 = .1$	84
J-1	Open Loop Representation of the Optimal LQG System.	120
J-2	The $(\dot{x}, x)$ Representation of the Optimal LQG System.	120
J-3	The $(\ddot{x}, \dot{x})$ Representation of the Optimal LQG System.	120
K-1	The Closed Loop System.	137
K-2	The Error-Driven Closed Loop Decomposition.	137
K-3	The Error-Driven Estimator Decomposition.	137
Q-1	An Hierarchy of Solutions to the Stochastic Realization Problem	232
W-1	Quantizer Function for $n = 7$	306
W-2	η_s as a Function of $K_1 \xi$	306
W-3	Phase Plane Plot of η_2 versus η_1 for the Actual System. . . .	306
W-4	Response of x_1 to Initial Condition for Actual System (solid line) and Approximate System.	306
W-5	Response of x_2 to Initial Condition for Actual System (solid line) and Approximate System.	306

LIST OF FIGURES (CONCLUDED)

Figure	Title	Page
W-6	Response of z_1 to Initial Condition for Actual System (solid line) and Approximate System.	306
W-7	Response of z_2 to Initial Condition for Actual System (solid line) and Approximate System.	306
X-1	Response of x_1 to Initial Condition for Actual System (solid line) and Approximate System.	324
X-2	Response of x_2 to Initial Condition for Actual System (solid line) and Approximate System.	325
X-3	Response of z_1 to Initial Condition for Actual System (solid line) and Approximate System.	326
X-4	Response of z_2 to Initial Condition for Actual System (solid line) and Approximate System.	327
Z-1	Example State Trajectory.	368
Z-2	Estimate Output	368

LIST OF TABLES

Table	Title	Page
B-1	The Transition Matrix Π and the Stationary Probabilities P. .	52
E-1	The Sample Variance of the Filters.	72
F-1	Sample Variance and Means for the Combined Filter and the EKF	81

SECTION I

INTRODUCTION

The geometry of the engagement of an air-to-air missile guidance scenario involves several elements that make the problem difficult to solve using straightforward techniques. The problem involves highly nonlinear geometry as well as nonlinear guidance and control models. The scenario involves uncertainties in the trajectory of the maneuvering target as well as in the magnitude and type of the maneuvers that make the tracking problem more complex. Finally, the digital implementation of the filtering and guidance algorithms have to reside in an airborne computer and hence problems of quantization errors need to be accounted for in the algorithms.

The objectives of this work are to develop an approximate model for nonlinear dynamic systems that can serve as a model for the air-to-air engagement scenario. Such a model is based on piecewise linear approximation of the nonlinearities and then the resulting model is further approximated by a set of linear models which are switched according to a Markov law. Such a model is then used to derive an implementable nonlinear filtering scheme for the tracking of such targets. This segment of the work represents the major part of this report. The extension of such models to higher dimensions and to practical scenarios are under continuing investigation.

A second major segment deals in the finite word-length implementation of the resulting filters. An error model for such filters is derived that considers the quantization effects. Two aspects of the model are considered: The first derives bounds on the errors due to the quantization, and the second derives corrections to the filters to improve the performance subject to the quantized implementations. A systematic design procedure for such systems is under continuing study.

Since the two approximations described above involve both quantization and systems that exhibit switched behavior among linear models, the third part of the report is concerned with the general behavior of quantized or switched systems. These may be modeled by what is known as hybrid system models. Since the original approximation of the nonlinear model involves switching behavior that exhibits fast and slow dynamics, this research concentrated on two general aspects of the models. First, general properties of hybrid systems for estimation and control purposes were derived. Second, the study of such systems when subjected to slow and fast dynamics has allowed the derivation of decoupled multiple time-scale implementations of controllers for such systems.

The results derived in these three parts of the project will be described in the body of the report, with the major derivations given in the Appendices.

SECTION II

MARKOV APPROXIMATION FILTERS

The primary objective of the research was to develop approximate filtering schemes for nonlinear dynamic models. The basic approach is based on approximating the nonlinear dynamics by piecewise linear segments. Such an approximation can be made arbitrarily accurate as the number of segments is increased. The next step in the approximation is to assume that the transition from one linear segment to another is based on a Markov transition law. In addition it is assumed that each linear submodel extends over the entire space so that we obtain a switched Markov linear approximation. This approximation is then used to develop implementable nonlinear filtering structure.

The research on the new filter structure may be divided into three primary phases. In the first phase, the mathematical theory associated with the switched Markov approximation to piecewise linear systems was developed, and all necessary constraints and approximations were clarified, References 1 and 2. These details are given in Appendix A and B. The extension of the model to higher dimensions is given in Reference 3 (see Appendix C). In the next phase of the research, the basic structure of the new filtering scheme was developed, References 4 and 5 (see Appendix D and E). During the final and most recent phase, attention has been directed toward improving the performance of the basic structure (albeit at the expense of complexity), Reference 6 and is given in Appendix F. Throughout the project, the development of Monte-Carlo digital computer simulations has accompanied the theoretical research.

Given the basic underlying piecewise linear model (or any model that may be accurately modelled as piecewise linear), the basic filter is developed by assuming that the single (N -segment) piecewise linear system may be approximated by N linear models with each (time series) sample chosen at random from one of the N systems running in parallel (the switched Markov model). The assumption of Markov switching leads to a well known optimal filtering scheme which is asymptotically unrealizable due to its exponentially increasing complexity with time. The research performed under this contract sought to develop a new realizable (albeit suboptimal) filtering scheme based on the optimal result. In essence, the goal was to decide how best to reduce the required computational complexity of the optimal filter, while still preserving its basic structure.

In phase two of the research, a structure was developed where a unique consistency update was devised whereby the estimates of filters tuned to certain linear models (macro-states) are checked to assure that the estimate is appropriate for the specific filter, Reference 5. When a filter and its estimate are not consistent, less weight is put upon that estimate in the overall calculation. Further, the filter uses aggregation techniques to reduce the number of filters (computational complexity) at each time step to a pre-determined value.

The filtering scheme discussed above has only a one step memory (with respect to the system macro-state). It was felt that superior results could be obtained by conditioning the consistency and aggregation calculations on a longer chain of macro-states. That is to say, the basic filter structure bases each consistency decision on only the current macro-state, but it is possible to look at a longer chain of macro-states for the consistency update. This scheme would have a memory of, say k , time samples. A collection of N^k sequences of macro-state trajectories would then exist for each time step. In turn, each consistency, and aggregation, calculation would be conditioned upon the macro-state sequence, rather than just the current macro-state. This new result allows a systematic technique for improving filter performance by expanding the memory of the filter. The development is given in Reference 6, where the switched Markov model has also been extended to the observation process as well.

Direct mathematical analysis of the filter is thought to be untractable due to its complexity, necessitating analysis through simulation techniques. Several detailed digital computer simulations were developed to simulate the new filter structure, and to allow comparison with conventional Extended Kalman filtering (EKF) techniques. At the present time, the simulation provides for the analysis of scalar systems with an arbitrary number of macro-states, and with the nonlinearity in both the state propagation function, as well as the measurement function. As currently implemented, the model does not fully support the multiple-level memory. The results of exercising the computer model indicate that the single stage filter shows considerable advantages over the EKF, particularly when the nonlinearities are not one-to-one.

Several questions exist concerning the advantages of the multi-memory filter. First, it is felt that since it is assumed that the system jumps from macro-state to macro-state with Markov jumps, then why should a filter require a multi-step memory? One response is that since the true macro-state is never known, adding the additional memory makes the detection problem (estimating the macro-state) more reliable. Further, it is believed that the filtering scheme will be applicable to many systems which are described by general nonlinearities rather than Markov switched piecewise linear models (i.e. the additional memory may improve filter performance in the presence of modeling errors). An approximate filter for smooth nonlinear systems was derived as an alternative to the present scheme in Reference 7 and Appendix G. Another issue concerns evaluation of the tradeoff between complexity and memory. Adding additional memory increases complexity, so it will be necessary to compare to increasing the number of filters (N) allowed at any time. Model reduction techniques may be used to simplify the systems in each macro-state. A technique for evaluating and consequently incorporating the loss of information based on balancing has been developed in References 8 and 9 and Appendices H and I. A related result on reduced order LQG design was derived in References 10 and 11 (see Appendices J and K).

SECTION III

FINITE WORD LENGTH EFFECTS

The main thrust of this part of the research has been in the analysis of the performance of filtering schemes subject to finite word-length implementations using floating point arithmetics. The results of the analysis are then used to derive improved design techniques for such filters. Finally these results are related to the modeling and realization aspects, with applications to switched Markov systems.

The bilinear model for finite word-length implementation of recursive computations using floating point arithmetic was first developed in Reference 12 (see Appendix L). The effect of the word-length on different implementations of recursive filters is discussed in References 13 and 14 and provided in Appendix M and N. The optimal derivation of filter gain to compensate for the quantization error due to the finite word-length is given in Reference 15 and Appendix O. The results show that there is quite a good agreement between the model and the actual error obtained via simulation. Furthermore, the results indicate the possibility of swapping model size for word-length size in order to arrive at an optimal choice for the filter computational complexity. The resulting related topics involve a more general approach to realization theory and optimal design using geometric approaches.

More recent results involve the improvement of the floating-point error model by including impulsive type errors that may occur as a result of subtraction operations in filter implementation. This has been accomplished by using an additive Poisson error model in addition to the bilinear (multiplicative) Gaussian error model.

As for realization theory we first distinguish between realization and identification. In the first all measurements are supposed to be error free. In the stochastic context for instance, this means that the joint distribution of all relevant variables are exactly known. The identification problem, we see then as the more realistic problem, given inaccurate measurements. The realization problem is therefore a fundamental idealization, and it is natural to study this first. Moreover any identification problem may be envisioned as the realization problem of a perturbed system. Approximation problems relate then to the previous as the realization of a system that is deliberately perturbed, so as to trade off with some complexity measure. The main question we considered is the following: Suppose a sequence of data vectors is obtained, how can we know if it is a sample path from a stochastic system, or a set of input/output data for a deterministic system? The early results of this theory are given in References 16, 17, and 18 and shown in Appendix P, Q, and R. Reduction techniques for stochastic models are discussed in References 19 and 20 and Appendix S.

The work on realization is still incomplete in many ways. More work is needed to strengthen some arguments. The idea that the data is fundamental and which leads to the measure theoretic (the measures are directly determined

from the data) approach will also enable us to generalize to a realization procedure for deterministic and stochastic automata. In fact, it turns out that the discussed setup is conceptually much simpler for systems over a discrete state space, which is appropriate for the piecewise linear Markovian approximation model for nonlinear systems discussed in Section I.

A related topic is the use of a geometric approach to the sensitivity problem that has the potential to lead to an optimal selection of a design of a filter given a finite word-length quantizer. In particular, we investigated the problem of approximating a system model using elements or parameters from a finite set. The preliminary results of this effort is provided in References 21 and 22 and Appendices T and U.

SECTION IV

HYBRID SYSTEMS MODELS

Hybrid systems are system models that include both discrete and continuous dynamics. The study of such systems is directly related to the switched Markov approximations of nonlinear systems. The macro-states provide the discrete states while the trajectory is represented by continuous states. We also obtain a hybrid system model if we consider a quantized implementation of guidance and control algorithms. The research into hybrid systems may be divided into three distinct parts.

The first considers the properties of systems subject to quantization and general hybrid systems subject to fast and slow dynamics. Such multiple time-scale dynamic behavior is inherent in the switched Markov approximation in which the system remains in the contracting macro-states for a long time and in the expanding macro-states only a brief time. Three major results were derived. The first derives the limiting behavior of hybrid systems when both the discrete process and the continuous process can display fast and slow dynamics. The results are given in Reference 23 and Appendix V. Next, the control algorithm for a quantized linear system subject to fast and slow dynamics is examined and an approximate method is used to implement the control using singular perturbation theory Reference 24 (see Appendix W). Finally, the results in Reference 24 are extended to a general multimodel system that is piecewise linear in different regions which is the basic approximation of interest throughout this investigation. These are given in Reference 25 and Appendix X.

The second topic is involved in investigating the general properties of hybrid systems. Properties such as observability, controllability, stability, and stabilizability of such systems are derived Reference 26 (see Appendix Y). The result are being used to derive guidance and control algorithms for such systems with direct applications to the control of the switched Markov systems used in this report.

The third topic involves the study of incompletely known hybrid systems. The incomplete knowledge may stem from, say fluctuations of the operating characteristic. For example, in radar, it may be the precise form of the radiation pattern of an antenna. Typically, the exact location of the many nulls in the pattern are unknown, or the exact pattern is too complex to describe especially when multiple scatterers are used. We derived an approximation method for this problem, which is exact if only one-step prediction is used. It is possible to do so for both discrete and continuous time systems, as the problem lies more with device characteristics than with algorithms.

An additional approach that is also a hybrid in nature involves the modeling of the maneuvers of the tracked vehicle by a dynamic system whose input is both Poisson and Gaussian noise. The approach assumes that the Poisson process is dependent on the continuous state of the system. Optimal

filters and an approximate implementation for such models are derived in Reference 27 and shown in Appendix Z. The approach is based on simultaneous detection of the incident actions of the Poisson input as well as the estimation of the state of the system which is the primary objective. The combination of techniques derived in these various approaches should yield a comprehensive procedure for tracking, guidance, and control of air-to-air systems satisfying the hybrid models assumptions.

SECTION V

CONCLUSIONS

The results of this work indicate that an implementable nonlinear filter can be developed for a scenario that includes the air-to-air guided missile as a special case. In particular, the approximate filter can be derived by a consistent set of approximating procedures that can be improved in accuracy as needed at the expense of complexity. The filter performs better than typical filters using different approximation especially for ambiguous and sharp nonlinearities. The implementation can be simplified if the special two-time scales behavior of the resulting approximation is utilized. The filter implementation may be improved if the finite word-length of the quantizers used in the computations are used to optimally select the filter gains. Finally, the general properties of hybrid systems and the general theory of realization and sensitivity can be used to design an improved guidance and control algorithms for such scenarios.

REFERENCES

- [1] A. H. Haddad and E. I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 202-206, Princeton University, March 1984.
- [2] 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.
- [3] A. H. Haddad and E. I. Verriest, "Piecewise Linear Modeling of Multidimensional Stochastic Nonlinear Systems," Proc. Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, pp. 777-778, October 1985.
- [4] A. H. Haddad, "On the Modeling and Filtering for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 44-49, Johns Hopkins University, March 1985.
- [5] 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.
- [6] 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.
- [7] E. I. Verriest, "Linear Filters for Linear Systems with Multiplicative Noise and Nonlinear Filters for Linear Systems with Non-Gaussian Additive Noise", Proc. 1985 American Control Conference, Boston, pp. 182-183, June 1985.
- [8] E. I. Verriest, "Stochastic Reduced Order Modeling of Deterministic Systems", Proc. 1985 American Control Conference, Boston, pp. 1003-1004, June 1985.
- [9] E. I. Verriest, "Uncertainty Equivalent Reduced Order Models for Discrete Systems", Proc. 24th IEEE Conf. on Decision and Control, Ft. Lauderdale, pp. 1238-1239, December 1985.
- [10] E. I. Verriest, "Model Reduction via Balancing, and Connections with Other Methods", in Modeling and Application of Stochastic Processes, (U. B. Desai, Ed.), Kluwer Academic Publishers, pp. 123-154, 1986.
- [11] E. I. Verriest, "Reduced Order LQG Design: Conditions for Feasibility", Proc. 25th IEEE Conf. on Decision and Control, Athens, Greece, pp. 1765-1769, December 1986.

- [12] E. I. Verriest, "On Redefining the Optimal Least Squares Filter Under Floating Point Operations," Proc. International Conf. on ASSP, San Diego, pp. 30.9.1-4, March 1984.
- [13] E. I. Verriest, "Error Analysis of Linear Recursions in Floating Point," Proc. International Conf. on ASSP, Tampa, pp. 44.1.1-4, March 1985.
- [14] E. I. Verriest, "A Bilinear Model for Linear Recursive Computations Using Floating Point Arithmetic," Proc. 19th Annual Conf. on Information Sciences and Systems, Johns Hopkins University, pp. 96-100, March 1985.
- [15] E. I. Verriest, "Gain Correction in Optimal Filtering Using Floating Point Arithmetic," Proc. 23rd IEEE Conf. on Decision and Control, Las Vegas, pp. 528-529, December 1984.
- [16] E. I. Verriest, "Subspace Correlation Measures and Applications to the Stochastic Realization Problem," Proc. Conference on Mathematical Theory in Networks and Systems, Tempe, AZ, June 1987.
- [17] J. A. Ramos, E. I. Verriest, and S. G. Rao, "A Unified RV-Coefficient Approach for Solving the Covariance based Stochastic Realization Problem," submitted to IEEE Trans. Automatic Control, 1987.
- [18] J. A. Ramos, and E. I. Verriest, "A Note on the Cross Riccattian and Related Properties for Symmetric Stochastic Realizations", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, December 1987.
- [19] E. I. Verriest, "Projection Techniques for Model Reduction", in Modeling, Identification and Robust Control, (C. I. Byrnes, Ed.), North-Holland, pp. 381-396, 1987.
- [20] 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.
- [21] W. S. Gray and E. I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity", Proc. 26th IEEE Conf. on Decision and Control, Los Angeles, December 1987.
- [22] E. I. Verriest and W. S. Gray, "Robust Design Problems: A Geometric Approach", Proc. Conference on Mathematical Theory in Networks and Systems, University of Arizona, Tempe AZ, June 1987.
- [23] M. V. Jose and A. H. Haddad, "On Singularly Perturbed Switched Parameter Systems," Proc. 1987 American Control Conference, Minneapolis, pp. 426-427, June 1987.
- [24] 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.

- [25] B. S. Heck and A. H. Haddad, "Singular Perturbation in Piecewise Linear Systems", submitted to 1988 American Control Conference, Atlanta, June 1988.
- [26] J. Ezzine and A. H. Haddad, "On the Controllability and Observability of Hybrid Systems", submitted to 1988 American Control Conference, Atlanta, June 1988.
- [27] 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.

BIBLIOGRAPHY

- [1] J. K. Tugnait, "Detection and Estimation for Abruptly Changing Systems," Automatica, Vol. 18, pp. 607-615, September 1982.
- [2] G. A. Ackerson and K. S. Fu, "On State Estimation in Switching Environment," IEEE Trans. on Automatic Control, Vol. AC-15, pp. 10-17, February 1970.
- [3] T. L. Johnson, "Synchronous Switched Linear Systems," Proc. 24th IEEE Conf. on Decision and Control, Ft. Lauderdale, December 1985, pp. 1699-1700.
- [4] G. S. Ladde and D. D. Siljak, "Multiplex Control Systems: Stochastic Stability and Dynamic Reliability", Int. J. Control, Vol. 38, 1983, pp. 515-524.
- [5] J. C. Willems, "From Time Series to Linear Systems," Automatica, to appear in 1987.
- [6] C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," IEEE Trans. Circuits and Systems, vol. CAS-23, 1976, pp. 551-562.

APPENDIX A

LINEAR MARKOV APPROXIMATIONS FOR PIECEWISE LINEAR STOCHASTIC
SYSTEMS

LINEAR MARKOV APPROXIMATIONS FOR PIECEWISE LINEAR STOCHASTIC SYSTEMS

A. H. Ezzadd and E. I. Verriest

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

Abstract

This paper is concerned with an approximate linear switched-parameter Markov model for discrete-time systems whose nonlinear homogeneous part is piecewise linear. A scalar system with white Gaussian noise input is considered and it is shown that a steady-state approximation is valid for two extreme cases. The first is the case when all the slopes of the piecewise linear model are stable, and the regions are large relative to the noise variance. The second is the case when there are unstable regions adjoining stable regions, and the unstable regions are small relative to the noise variance.

I. INTRODUCTION

There has been a large amount of work devoted to estimation and filtering in switched-environments (e.g., [1-7]), which involves linear systems with unknown parameters or models which can take different values at every observation instant based on a Markov chain model. Such models have been used to represent systems with time-varying but unknown parameters observed in noise with unknown but slowly changing covariance matrix. Such schemes are applicable to nonlinear systems when represented as a Markov transition among linear models [8]. Since many nonlinear problems defy systematic analytical procedures, this paper is concerned with exploring the approximation of a nonlinear dynamic system by a set of linear models with Markov transitions among these linear models. The objective is to derive the approximate model, and investigate the conditions under which it is a valid representation of the nonlinear system. For simplicity we consider a scalar discrete-time system

$$x_{k+1} = g(x_k) + w_k \quad (1)$$

where x_k is the state at time t_k , $g(x)$ is a zero-memory nonlinearity, and $\{w_k\}$ is a white Gaussian noise sequence with zero mean and variance σ^2 . The primary assumption is that $g(x)$ is given as a piecewise linear function, or that it can be adequately represented by such an approximation:

$$g(x) = \beta_i + \frac{\beta_{i+1} - \beta_i}{\alpha_{i+1} - \alpha_i} (x - \alpha_i), \quad \alpha_i < x < \alpha_{i+1} \\ i = 1, 2, \dots, N \\ \approx a_i x + b_i \quad (2)$$

Let the macrostate S_i define the region $\alpha_i < x < \alpha_{i+1}$, then from (1) and (2) we may derive transition probabilities

$$\Pi_{ij}(x_k) = P\{x_{k+1} \in S_j | x_k \in S_i, x_k\}$$

The approximation considered in this paper replaces $\Pi_{ij}(x_k)$ by

$$\Pi_{ij} = E\{\Pi_{ij}(x_k) | x_k \in S_i\} \quad (3)$$

which removes the dependence of the transition probabilities on the actual value of x_k , and also assumes that in macrostate S_i the system satisfies the linear equation

$$x_{k+1} = a_i x_k + b_i + w_k \quad (4)$$

Furthermore, it is assumed that the transition from state to state follows a Markov chain rule with transition matrix $\{\Pi_{ij}\}$.

The objective of this paper is to analyze the validity of the approximation depending on the assumptions on the parameters (α_i , a_i , b_i , σ). In this section the notations of the transition and other density functions of the system state are defined and derived. Then the analysis is carried out for the special case of an odd symmetry in $g(x)$ with $N = 3$ to simplify the analysis. The special cases of stable systems (4) is first considered, and then the unstable case falling between two stable regions is investigated. The resulting expressions, while specialized to $N = 3$, may be generalized without major difficulty to the arbitrary case.

Let $f_k(x)$ denote the probability density function of x_k , then we define two new quantities:

$$p_{ki}(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\alpha_i}^{\alpha_{i+1}} e^{-\frac{(x - a_i y - b_i)^2}{2\sigma^2}} f_k(y) dy \quad (5)$$

and

$$P_{ki} = \int_{\alpha_i}^{\alpha_{i+1}} f_k(y) dy \quad (6)$$

The transition probabilities and the recursive relation of the density functions are given by

$$f_{k+1}(x) = \sum_{i=1}^N p_{ki}(x) \quad (7)$$

and

$$\Pi_{ij}(k) = \frac{1}{P_k} \int_{\alpha_j}^{\alpha_{j+1}} p_{ki}(x) dx \quad (8)$$

Assuming a steady state distribution and transition matrix exist, then they must satisfy:

$$f(x) = \sum_{i=1}^N p_i(x) \quad (9)$$

$$p_i = \int_{a_i}^{a_i+1} f(x) dx \quad (10)$$

$$\Pi_{ij} = \frac{1}{p_i} \int_{a_j}^{a_j+1} p_i(x) dx \quad (11)$$

The transition matrix $\{\Pi_{ij}\}$ in turn leads to a steady state probability, Π_i , of the system being in state S_i . The remaining parts of the paper are devoted to exploring the conditions under which for the model defined by (3) and (4) one obtains $\Pi_i = p_i$ and the $p_i p_i(x)$ are approximately equal to the density obtained for the system (4) under state S_i . The derivation is performed for an odd-symmetric $g(x)$ with $N=3$ as discussed above.

II. THE STABLE CASE

For the stable case we consider for convenience $g(x)$ of the form

$$g(x) = \begin{cases} ax & , |x| < a \\ bx + \beta \operatorname{sgn} x & , |x| > a \end{cases} \quad (12)$$

where $\beta = a(a-b)$, and assume that $|a| < 1$ and $|b| < 1$, so that the system is stable in each region. The system has three states denoted S_0 : $|x| < a$, S_+ : $x > a$ and S_- : $x < -a$. Due to symmetry, the steady-state densities and transitions are given by:

$$f(x) = p_0(x) + p_+(x) + p_+(-x)$$

$$p_0(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-a}^a e^{-\frac{(x-y)^2}{2\sigma^2}} f(y) dy$$

$$p_+(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_a^\infty e^{-\frac{(x-by-\beta)^2}{2\sigma^2}} f(y) dy$$

$$p_0 = \int_{-a}^a f(x) dx, \quad p_+ = \int_a^\infty f(x) dx = \frac{1}{2} (1-p_0) \quad (13)$$

$$\Pi_{00} = \frac{1}{p_0} \int_{-a}^a p_0(x) dx$$

$$\Pi_{0+} = \Pi_{0-} = \frac{1}{p_0} \int_a^\infty p_0(x) dx$$

$$\Pi_{+0} = \Pi_{-0} = \frac{2}{1-p_0} \int_a^\infty p_+(x) dx$$

$$\Pi_{++} = \Pi_{--} = \frac{2}{1-p_0} \int_a^\infty p_+(x) dx$$

$$\Pi_{+-} = \Pi_{-+} = \frac{2}{1-p_0} \int_a^\infty p_+(-x) dx \quad (14)$$

It is obvious from these expressions that

$$\int_{-\infty}^{\infty} p_0(x) dx = \int_{-a}^a f(x) dx, \quad \int_{-\infty}^{\infty} p_+(x) dx = \int_a^\infty f(x) dx \quad (15)$$

If we now assume that $a/\sigma \gg 1$ then it appears that the transition probabilities to other states are relatively small and hence an approximate solution for the Markov linear model becomes one of weighted sum of the steady-state densities in each region multiplied by the corresponding probability of being in that macrostate. The steady state for the probability densities solution of (4) for the nonlinearity (12) in the three regions S_0 and $S_\pm$ is denoted by $q_0(x)$ and $q_\pm(x)$ respectively, and is given by

$$q_0(x) = \frac{1}{\sqrt{2\pi} \sigma_0} e^{-\frac{x^2}{2\sigma_0^2}} \quad (16)$$

$$q_\pm(x) = \frac{1}{\sqrt{2\pi} \sigma_1} e^{-\frac{(x \mp u)^2}{2\sigma_1^2}} \quad (17)$$

$$\sigma_0^2 = \frac{\sigma^2}{1-a^2}, \quad \sigma_1^2 = \frac{\sigma^2}{1-b^2}, \quad u = \frac{\beta}{1-b} = \frac{a(a-b)}{(1-b)} \quad (18)$$

If we denote the steady-state probabilities of being in macrostate S_0 , $S_\pm$ by Π_0 and $\Pi_\pm$ respectively, we arrive at the approximation

$$p_0(x) = \Pi_0 q_0(x), \quad p_\pm(x) = \Pi_\pm q_\pm(x) \quad (19)$$

where $\Pi_0 = (1 + 2 \frac{\Pi_+}{\Pi_0})^{-1}$, $\Pi_+ = \Pi_- = \frac{1}{2} (1 - \Pi_0)$ (20)

It remains to show that this heuristic approximation is indeed valid as first-order approximation to the functions as defined by the integral operators (13). The substitution of $q_0(x)$ and $q_\pm(x)$ in the integral equations (13) yields the following error terms in $p_0(x)$ and $p_+(x)$ respectively:

$$-\frac{\Pi_0}{\sqrt{2\pi} \sigma} \left\{ \int_{-a}^a e^{-\frac{(x-y)^2}{2\sigma^2}} q_0(y) dy + \int_a^\infty e^{-\frac{(x+y)^2}{2\sigma^2}} q_0(y) dy \right\} + \frac{\Pi_+}{\sqrt{2\pi} \sigma} \int_{-a}^a e^{-\frac{(x-y-\beta)^2}{2\sigma^2}} (q_+(y) + q_+(-y)) dy \quad (21)$$

$$\frac{1}{\sqrt{2\pi} \sigma} \left\{ \int_a^\infty e^{-\frac{(x-by-\beta)^2}{2\sigma^2}} (\Pi_0 q_0(y) + \Pi_+ q_+(y)) dy - \Pi_+ \int_{-\infty}^a e^{-\frac{(x-by-\beta)^2}{2\sigma^2}} q_+(y) dy \right\} \quad (22)$$

These terms can be expressed explicitly by using the Gaussian density and distribution functions. It can be shown after lengthy manipulations that these terms can become negligible if $a/\sigma \gg 1$ provided that the other parameters are sufficiently bounded as follows:

$$|a| < K_1 < 1, \quad |b| < K_2 < 1, \quad |b| - |a| > K_3 > 0.$$

The derivation for this case does not yield an expansion for the appropriate densities so that correction terms may be used. The generalization to $N > 3$ and to nonsymmetric nonlinearities appears straightforward even though tedious and the condition in this case are related to a relatively large linear regions, and sufficiently different slopes which are bounded away from unity.

III. THE UNSTABLE CASE

This section is concerned with cases when the nonlinearity $g(x)$ has slopes greater than unity, and thus leading to unstable models in (4). Since in this case the system leaves the regions with probability one, in order to obtain a meaningful model, we have to assume that unstable regions are bordered by stable ones. Furthermore, it is assumed that the external regions are stable so that the overall system will not diverge. Again, for convenience we consider the case $N=3$ given in (12) with the added assumption that $|b| < 1$, while $|a| > 1$. If the notations for the densities and transition probabilities of Section II are used, then for this case (13) and (14) continue to be valid. However, while we may consider a steady state for $f(x)$ we cannot postulate the existence of a steady state for $q_0(x)$ which represents the steady state density for the system in region S_0 . In order to obtain reasonably valid approximations, it is assumed that $a/\sigma < 1$ so that the transition probability $\Pi_{00} < 1$. In this case asymptotic expansions of the steady-state solution for $f(x)$ are possible, and thus $p_0(x)$ and $p_+(x)$ can be identified and compared to the linear Markov models. However, due to the size of the unstable region, the two stable regions behave in the limit as a single region. In order to avoid this rather sundane case, the additional assumption that $|b|/\sigma \gg 1$ is made. It can be shown that these conflicting requirements can be met if the slope of the unstable region is large.

We start by obtaining a series expansion for the solution of (12) and (13). Let $p_1(x) = p_+(x) + p_-(x)$ then we have

$$p_0(x) = \hat{G}(a)[p_0 + p_1] \quad (23)$$

$$p_1(x) = \hat{H}(a)[p_0 + p_1] \quad (24)$$

where $\hat{G}(a)$ and $\hat{H}(a)$ are integral operators defined by

$$\begin{aligned} \hat{G}(a)\phi &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-a}^a e^{-\frac{(x-ay)^2}{2\sigma^2}} \phi(y) dy \\ &= \int_{-a}^a G(x,y) \phi(y) dy \end{aligned} \quad (25)$$

$$\begin{aligned} \hat{H}(a)\phi &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-a}^a \left[e^{-\frac{(x-by-\beta)^2}{2\sigma^2}} + e^{-\frac{(x+by+\beta)^2}{2\sigma^2}} \right] \phi(y) dy \\ &= \int_{-a}^a H(a;x,y) \phi(y) dy \end{aligned} \quad (26)$$

Since the L_∞ norm of $\hat{G}(x)$ is bounded by (a/σ) , then it is possible to expand it in a series with respect to the parameter a/σ . Using the assumptions $a/\sigma \ll 1$ and $a/b \cdot a/\sigma \gg 1$ we may use as a starting point the steady state solutions of (4) in regions $S_\pm$ as an approximation to $p_1(x)$, namely

$$p_1(x) \approx \Pi_+ [q_+(x) + q_-(-x)] \quad (27)$$

In view of the fact that $|p_0(x)| < a/\sigma \sqrt{2/\pi} K$, where K is finite, (the value of K will be verified in the sequel), the error terms of the approximations (27) when substituted in (24) can be derived in the same manner as for the stable case of Section II. It remains to derive the approximation for $p_0(x)$ by substituting (27) in (23). The first-order term for $p_0(x)$ can be obtained by integration as:

$$p_0(x) \approx \frac{\Pi_+}{\sqrt{2\pi}\sigma_2} \left[\psi(x) e^{-\frac{(x-au)^2}{2\sigma_2^2}} + \psi(-x) e^{-\frac{(x+au)^2}{2\sigma_2^2}} \right] \quad (28)$$

where

$$\sigma_2^2 = a\sigma^2 + \sigma_1^2 \quad (29)$$

and

$$\psi(x) = \phi\left(\frac{ax\sigma_1^2 + a\sigma_2^2 + u\sigma^2}{\sigma_1\sigma_2\sigma}\right) - \phi\left(\frac{ax\sigma_1^2 - a\sigma_2^2 + u\sigma^2}{\sigma_1\sigma_2\sigma}\right) \quad (30)$$

where ϕ is the unit Gaussian distribution. The function $\psi(x)$ achieves its maximum at

$$x = -\frac{u\sigma^2}{a\sigma_1^2} = -a\left(1 - \frac{b}{a}\right) (1+b) \quad (31)$$

and the maximum value is given by

$$|\psi(x)| < \phi\left(\frac{a\sigma_2^2}{\sigma_1\sigma}\right) - \phi\left(-\frac{a\sigma_2^2}{\sigma_1\sigma}\right) = K_1 < 1 \quad (32)$$

However, the maximum of $\psi(x)$ occurs inside $(-a,a)$ while of the multiplying exponential is in (a,∞) . Consequently, the following bounds for $p_0(x)$ in the two intervals may be obtained: For $|x| < a$, we have

$$\begin{aligned} |p_0(x)| &< \frac{\Pi_+}{\sqrt{2\pi}\sigma_2} K_1 \left[e^{-\frac{(x-au)^2}{2\sigma_2^2}} + e^{-\frac{(x+au)^2}{2\sigma_2^2}} \right] \\ &< \frac{2\Pi_+}{\sqrt{2\pi}\sigma_2} e^{-\frac{(au-a)^2}{2\sigma_2^2}} < 1. \end{aligned}$$

and for $|x| > a$

$$|p_0(x)| < \frac{\Pi_+}{\sqrt{2\pi}\sigma_2} \psi(a) \left[e^{-\frac{(x-au)^2}{2\sigma_2^2}} + e^{-\frac{(x+au)^2}{2\sigma_2^2}} \right]$$

$$< \frac{2\Pi_+}{\sqrt{2\pi} \sigma_2} \psi(a) < 1$$

These bounds validate the original assumptions made about $p_0(x)$. Furthermore, it can be easily shown that the steady state probability of macro-state S_0 is given by

$$\begin{aligned} \Pi_0 &\approx \int_{-\infty}^{\infty} p_0(x) dx = \Pi_+ \int_{-a}^a [q_+(x) + q_+(-x)] dx \\ &= \frac{\Pi_+ 4a}{\sqrt{2\pi} \sigma_1} e^{-\frac{(\mu-a)^2}{2\sigma_1^2}} < \frac{\sqrt{2}}{\pi} \frac{a}{\sigma} \end{aligned} \quad (33)$$

which confirms the assumptions that $\Pi < 1$ and the size of the bound on $p_0(x)$. Additional manipulations of (28) allow the expression for $p_0(x)$ to be approximated by

$$p_0(x) \approx \frac{\Pi_+}{\sqrt{2\pi} \sigma_2} \left(\Pi_{+0} e^{-\frac{(x-au)^2}{2\sigma_2^2}} + \Pi_{-0} e^{-\frac{(x+au)^2}{2\sigma_2^2}} \right) \quad (34)$$

where

$$\Pi_{+0} \approx \psi\left(\frac{\mu+a}{\sigma_1}\right) - \psi\left(\frac{\mu-a}{\sigma_2}\right) < 1 \quad (35)$$

The density obtained in (34) is equivalent to that derived for the system state x_{k+1} in one step

$$x_{k+1} = ax_k + w_k, \quad x_k \in S_k \quad (36)$$

unifying that the transition $\Pi_{00} < 1$ so that higher order terms of the approximations of $p_0(x)$ are negligible. The bounds on the transition probabilities are obtained from their approximate expressions by using (14), (28), (34) and (27).

We now consider the degenerate case of $\epsilon < 1$ and $\epsilon a/b < 1$. It is convenient to reformulate the problem as a perturbation problem. Upon defining the perturbation (or scattering) operator

$$\tilde{H}(a) = \hat{H}(a) - \hat{H}(0) \quad (37)$$

and expanding in a Taylor series about $a=0$, one gets

$$\tilde{H}(a) = \sum_{n=1}^{\infty} a^n \tilde{H}_{(n)} \quad (38)$$

where

$$\begin{aligned} \tilde{H}_{(n)} \psi &= \frac{1}{n!} \int_0^a \left[\left(\frac{\partial}{\partial a} \right)^n \tilde{H}(a; x, y) \right]_{a=0} \psi(y) dy + \\ &- \frac{1}{(n-1)!} \left\langle \left[\left(\frac{\partial}{\partial a} \right)^{n-1} \tilde{H}(a; x, y) \right]_{a=0} \psi(y) \right\rangle \end{aligned} \quad (39)$$

$\langle \psi(y) \rangle$ denotes the average of ψ over $(0, a)$ i.e.

$$\langle \psi(y) \rangle \triangleq \frac{1}{a} \int_0^a \psi(y) dy \quad (40)$$

The first order scattering operator is

$$\tilde{H}_1 \psi = -\frac{a-b}{\sigma^2} \hat{L} \psi - \langle \tilde{H}(0; x, y) \psi(y) \rangle$$

where

$$\begin{aligned} \hat{L} \psi &= \int_0^a L(x, y) \psi(y) dy \\ &= \int_0^a \frac{(by-x)e^{-\frac{(by-x)^2}{2\sigma^2}} + (by+x)e^{-\frac{(by+x)^2}{2\sigma^2}}}{\sqrt{2\pi} \sigma} \psi(y) dy \end{aligned} \quad (41)$$

For what follows, it will be convenient to write the unperturbed operator $\tilde{H}(0)$ and the scattering operator $\hat{L}$ in a symmetric form. Indeed by substituting $-y$ for y in $\tilde{H}(0; x, y)$ and $L(x, y)$ and adding we get

$$\hat{H} \psi = \int_{-\infty}^{+\infty} \tilde{H}(0; x, y) \text{Ev}[\psi(y)] dy \quad (42)$$

$$\hat{L} \psi = \int_{-\infty}^{+\infty} L(x, y) \text{Odd}[\psi(y)] dy \quad (43)$$

where $\text{Ev}[\psi]$ and $\text{Odd}[\psi]$ are respectively the even and odd parts of $\psi(\cdot)$. The L_∞ norm of the perturbing operator $c\tilde{H}_{(1)}$ is bounded by

$$\frac{\sqrt{2}}{\pi} \frac{a}{\sigma} \frac{a}{b} = \epsilon \frac{a}{b}$$

The assumptions made at the beginning of the paragraph validate therefore a perturbation expansion. Letting

$$p_0(x) = p_0^{(0)}(x) + c p_0^{(1)}(x) + c^2 p_0^{(2)}(x) \dots$$

$$p_1(x) = p_1^{(0)}(x) + c p_1^{(1)}(x) + c^2 p_1^{(2)}(x) \dots$$

and substituting in

$$p_0 = c\tilde{G}(p_0 + p_1) \quad (44)$$

$$p_1 = (\tilde{H}(0) + c\tilde{H}) (p_0 + p_1) \quad (45)$$

where we wrote explicitly $c\tilde{G}$ for $\hat{G}$ and $c\tilde{H}$ for $\hat{H}_{(1)}$. We get, up to 1st order

$$p_0^{(0)} = 0 \quad (46a)$$

$$p_1^{(0)} = \tilde{H}(0) p_1^{(0)} \quad (46b)$$

$$p_0^{(1)} = \tilde{G} p_1^{(0)} \quad (46c)$$

$$p_1^{(1)} = \tilde{H}(0) p_1^{(1)} + \tilde{H} p_1^{(0)} + \tilde{H}(0) p_0^{(1)} \quad (46d)$$

The solution of (46b) is readily found to be

$$p_1^{(0)}(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \quad (47)$$

Note that this is the normalized solution. Clearly, when the perturbation terms are to be taken into account, the overall solution needs to be renormalized, i.e.

$$f(x) = p_0(x) + p_1(x) = p_1^{(0)}(x) + c(p_0^{(1)}(x) + p_1^{(1)}(x)) = (1 + c\tilde{G} + c(1-H)^{-1}(\tilde{H} + \tilde{H}\tilde{G}))p_1^{(0)}(x)$$

must integrate to one. The successive solutions for $p_0^{(i)}$ and $p_1^{(i)}$ are then computed as a series in the orthonormal Hermite functions

$$\phi_n(t) = \frac{1}{\sqrt{2^n n!} \sqrt{\pi}} H_n(t) e^{-\frac{t^2}{2}} \quad (48)$$

Some algebra shows that for

$$f(x) = \sum_{n=0}^{\infty} f_n \phi_n\left(\frac{x}{\sigma_1}\right) \quad (49)$$

$$\psi^{(k)}(x) = \sum_{n=0}^{\infty} \psi_n^{(k)} \phi_n\left(\frac{x}{\sigma_1}\right) \quad (50)$$

the equations

$$\begin{aligned} \psi^{(1)} &= \tilde{G}f, \quad \psi^{(2)} = \tilde{H}f, \\ \psi^{(3)} &= \tilde{H}f, \quad \psi^{(4)} = \tilde{H}\psi^{(4)} + f \end{aligned}$$

yield respectively

$$\begin{aligned} \psi_k^{(1)} &= \frac{\sqrt{\pi}}{2} \frac{1}{\sigma_1^k} \sum_{m=0}^{\infty} f_m \frac{\sigma_1}{a} \int_{-a}^a \phi_n\left(\frac{y}{\sigma_1}\right) \phi_k\left(\frac{ay}{\sigma_1^2}\right) dy \\ \psi_{2k}^{(2)} &= -\frac{\sqrt{\pi}}{2} (a-b) 4\sqrt{k} b^{2k} f_{2k-1} - \sqrt{2\pi} b^{2k} f_{2k} \phi_{2k}\left(\frac{b\theta}{\sigma_1}\right) \\ \psi_{2k+1}^{(2)} &= \sqrt{2\pi} b^{2k+1} f_{2k} \phi_{2k+1}\left(\frac{b\theta}{\sigma_1}\right) \\ \psi_{2k}^{(3)} &= \frac{\sqrt{2}}{\pi} \left(\frac{b}{\sigma_1}\right)^{2k} f_{2k} \\ \psi_{2k+1}^{(3)} &= 0 \\ \psi_{2k}^{(4)} &= \frac{1}{1 - \frac{\sqrt{2}}{\pi} \left(\frac{b}{\sigma_1}\right)^{2k}} f_{2k} \\ \psi_{2k+1}^{(4)} &= f_{2k+1} \end{aligned}$$

The θ appearing in the coefficients for $\psi^{(2)}$ is contained in the interval $(0, a)$ and arises from the fact that $\tilde{H}$ contains an "average" of Hf which by Rolle's theorem equals the value of Hf for some point in that interval $(0, a)$. Further the expansion

$$H(0; x, \theta) = \sum_{n=0}^{\infty} 2(-b)^n \phi_n\left(\frac{b\theta}{\sigma_1}\right) \phi_n\left(\frac{x}{\sigma_1}\right) \quad (51)$$

is used. The use of these expansions allows then the computation of the steady state probabilities and the transition probabilities up to first order for the macrostates.

IV. SUMMARY AND CONCLUSIONS

This paper showed that under certain restrictions a piecewise nonlinear discrete-time dynamic system may be approximated by Markov

transitions among several linear models, as far as the steady-state distribution is concerned. It remains to generalize the approach to higher order systems as well as multiple nonlinearities. The effect of the approximation on estimation and filtering schemes needs to be further explored. Finally, the continuous time version requires a different approach because of the high density of the level crossings at the boundary. One approach is to model the noise not as white noise, but as having a finite nonzero correlation length. The additional aspects of the Stratonovich integral definitions are expected to be relevant.

V. REFERENCES

1. D. G. Lainiotis, "Optimal adaptive estimation: Structure and parameter adaptation," IEEE Trans. Automat. Contr., vol. AC-16, pp. 160-170, April 1971.
2. A. B. Haddad and J. K. Tugnait, "On state estimation using detection estimation schemes for uncertain systems," Proc. JACC, pp. 514-519, Denver, CO, June 1979.
3. G. A. Ackerson and K. S. Fu, "On state estimation in switching environments," IEEE Trans. Automat. Contr., vol. AC-15, pp. 10-17, February 1970.
4. C. B. Chang and M. Athans, "State estimation for discrete systems with switching parameters," IEEE Trans. Aerosp. Electron. Syst., vol. AES-14, no. 418-425, May 1978.
5. J. K. Tugnait, "Detection and estimation for abruptly changing systems," Automatica, vol. 18, pp. 607-615, Sept. 1982.
6. H. Akashi and H. Kumamoto, "Random sampling approach to state estimation in switching environments," Automatica, vol. 13, pp. 429-434, July 1977.
7. J. K. Tugnait, "Adaptive estimation and identification for discrete systems with Markov jump parameters," IEEE Trans. on Automat. Contr., vol. AC-27, pp. 1054-1065, October 1982.
8. E. F. Van Landingham, R. L. Moose, and W. H. Lucas, "Modeling and control of nonlinear plants," Proc. 17th IEEE Conf. Decision and Control, pp. 337-344, January 1979.

APPENDIX B

LINEAR MARKOV APPROXIMATIONS OF PIECEWISE LINEAR STOCHASTIC
SYSTEMS

LINEAR MARKOV APPROXIMATIONS OF PIECEWISE LINEAR STOCHASTIC SYSTEMS

E. I. Verriest and A. H. Haddad

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

ABSTRACT

This paper is concerned with the properties of piecewise linear discrete-time dynamic systems driven by white Gaussian noise. The properties of the deterministic system are explored, and condition for the existence of invariant distributions are derived. The existence of an invariant distribution was then used to justify the approximation of the stochastic system by a switched Markov linear model if the piecewise linear regions are large "contracting" ones or small "expanding" ones relative to the input noise variance. The approach is expected to be useful for constructing approximate nonlinear filtering schemes for such systems.

1. INTRODUCTION

Piecewise linear systems can represent an approximation to general nonlinear systems. Usually, nonlinear filtering schemes for such systems are not exactly implementable. As a step leading to the development of nonlinear filtering schemes a systematic analysis of such systems is needed. Since it is not possible to obtain exact implementations, one must resort to approximations. Such approximations can be made at the filtering stage, or they can be made at the modeling stage. Consequently we are concerned in this paper in the analysis problem of piecewise linear systems driven by white Gaussian noise. However, as the analysis is

already complex in the deterministic case, we consider the properties of the deterministic case first. The model we are interested in is the switched Markov linear model for which there already exist several approximate approaches for the solution of the filtering problem. This model assumes that underlying the system there exists a finite state Markov model, such that under each state the system satisfies one linear dynamic model. The paper is concerned with the assumptions that are needed for the approximation to be a valid representation for the original nonlinear system.

In the study of the dynamics of discrete nonlinear systems, it is known that in some cases, the sensitivity of trajectories (sequences) with respect to initial condition may lead to some wildly unpredictable dynamic behavior, even though the system is completely deterministic. It appears therefore that a statistical approach may be more fruitful than a purely deterministic one. In fact, what seems to be a very significant problem in this case is the study of the flow of densities. In particular, a very important question is the existence and uniqueness of invariant measures which will be considered in the first part of this paper. Its importance stems from the fact that existence and uniqueness of invariant measures imply the ergodicity of the dynamical map. The second part will build upon the properties of the deterministic distributions to obtain the characteristics of the system for the stochastic case [1,2].

2. DETERMINISTIC NONLINEAR SYSTEMS

This section introduces and defines the terms to be used in the later sections of the paper, and describes the properties of piecewise linear dynamic discrete-time systems. In general we are concerned with scalar systems of the form

$$x_{k+1} = f(x_k) \tag{1}$$

where x_k is the state of the system at time t_k , and $f(x)$ is a nonlinear function. The properties of such a system are discussed

in general, with particular attention to the piecewise linear case.

2.1. General Maps

Let f be a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, associated with a dynamical system

$$x_{k+1} = f(x_k) \quad \text{where } x \text{ is in } \mathbb{R}^n$$

Definitions:

The map $f^k: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called the n -th iterated map of x .

The orbit of x is the sequence $x, f(x), f^2(x), \dots$

An equilibrium point is a point x in $\mathbb{R}^n$ for which $f(x) = x$. Equilibrium points are also called fixed points.

A point x is a periodic point for the map f of period p if it is a fixed point for f^p . The least positive n such that $f^n(x) = x$ is called the prime period of x .

If x is periodic with prime period p , then the part $x, f(x), f^2(x), \dots, f^{n-1}(x)$ of the orbit of x will be called the cycle of x , and will be denoted by $\langle x \rangle$.

A set S is an invariant set for f if both

$$f(S) \subset S$$

$$f^{-1}(S) \subset S$$

Alternative definitions can be given for invariant sets by virtue of the following equivalence:

Theorem 1: The following are equivalent:

- 1) The set S is invariant with respect to the map f
- 2) $f^{-1}(S) = S$
- 3) For all integer k , the set $f^k(S)$ is included in S .

Proof: We first show the equivalence of 1) and 2). By definition, we already have $f^{-1}(S) \subset S$. So we only need to show that $S \subset f^{-1}(S)$. But this is obvious, since for all x in S , $f(x)$ is in S by the first part of the definition. But then x is also in the inverse image $f^{-1}(S)$.

Conversely, if $f^{-1}(S) = S$, then $f^{-1}(S) \subset S$ is obvious, while for

all x in $S = f^{-1}(S)$, clearly $f(x)$ is in S . Hence $f(S) \subset S$.
 Next, letting k equal 1 or -1 in 3) implies the definition.
 Iteration on the conditions of the definition implies 3).•

Finally, if f is invertible, then invariance of a set S under the map f is equivalent to $f(S) = S$.

Related to the notion of an invariant set is:

A closed, simply connected region Q in $\mathbb{R}^n$ is a **trapping region** in f if $f(Q)$ is contained in the interior of Q . It is sufficient that the vector field $f(x)$ is directed everywhere inward on the boundary of Q . It follows that for a trapping region $f^k(Q)$ is contained in Q for every positive integer k . In fact it is readily shown that the sets $f^k(Q)$ are closed and nested.

2.2. General One-Dimensional Maps

In the one dimensional case, assume that f is differentiable almost everywhere. (i.e. f is not differentiable in at most a countably infinite number of points. Then the following definitions are standard:

An equilibrium point (fixed point) of a one-dimensional differentiable map is called **stable** if $|f'(x)| < 1$, and **unstable** if $|f'(x)| > 1$.

Theorem 2: If $\inf |f'(x)| > 1$ in some invariant set S , then $\inf |(f^n)'(x)| > 1$ in S . If $\sup |f'(x)| < 1$ in some invariant set S , then $\sup |(f^n)'(x)| < 1$ in S .

Proof: By induction. Assume $\inf |(f^n)'(x)| > 1$, then

$$\begin{aligned} \inf |(f^{n+1})'(x)| &= \inf |f'(f^n(x)) (f^n)'(x)| \\ &\geq \inf |f'(f^n(x))| \inf |(f^n)'(x)| \\ &> 1 \end{aligned}$$

The second part is completely analogous. •

Even for continuous maps f , the invariant sets can have a strange topology. A classic example is the logistics map $f(x) = \mu x(1-x)$ for $\mu > 4$. The invariant set for this map has the structure of the Cantor set [3]. A Cantor set is a closed, totally

disconnected, and perfect set. A set is totally disconnected if it contains no intervals, and it is perfect if every point in it is the accumulation point of other points in the set. A Cantor set has a Lebesgue measure of zero, even though it contains an uncountably infinite number of points. On the other hand, if an interval is an invariant set for f , the following can be asserted:

Theorem 3: Closed invariant intervals of piecewise continuous maps contain at least one fixed point.

Proof: Let $J = [a, b]$ be the invariant interval. If there would not be a fixed point, the graph of $f(x)$ is either above or under the graph (x, x) in J . Say $f(x) > x$ in J , then clearly $f(b)$ is not in J , contradicting the invariance of J . •

Remark: Open sets may be invariant and contain no fixed points as long as the limit points are fixed points.

An interesting result on the existence of periodic points is the following, due to Sarkowskii:

Theorem 4: (Sarkowskii) Suppose $f: \mathbb{R} \rightarrow \mathbb{R}$ is continuous. Suppose f has a periodic point of prime period k . If $k \prec 1$, where $\prec$ is the Sarkowskii ordering, then f also has a periodic point of period 1.

The Sarkowskii ordering is an ordering of the natural numbers:

$$3 \prec 5 \prec 7 \prec \dots \prec 2 \cdot 3 \prec 2 \cdot 5 \prec \dots \prec 2^2 \cdot 3 \prec 2^2 \cdot 5 \prec \dots \prec 2^3 \cdot 5 \prec 2^3 \cdot 5 \prec \dots \prec 2^3 \prec 2^2 \prec 2 \prec 1$$

A proof of this theorem can be found for instance in [3].

2.3. Markov Partitions and Markov Maps

A special class of one dimensional dynamical systems are the piecewise differentiable maps with the following restrictions:

1) There exists a partition P of the real line in a finite or countable set of disjoint open intervals $\{I_k\}$, such that the map f restricted to the interval I_k is differentiable. Denote by Ω the set of closure points of the intervals $\{I_k\}$. Clearly, Ω is a countable disjoint set.

$$\Omega = \mathbb{R} \setminus \bigcup_k I_k \quad (2)$$

2) In each interval I_k the infimum and supremum of $f(x)$ belong to the set Ω .

The above condition may look like a very severe restriction on the map f , however it seems to be crucial in order to "do the theory". It is noteworthy that any piecewise continuous map can be suitably approximated by such a map, by taking finer and finer partitions of $\mathbb{R}$. In particular, in a later chapter we shall study such an approximation. Any map satisfying the above two restrictions will be called a **Markov map**, and the corresponding partition P a **Markov Partition**. We caution that these definitions are not standard. Their motivation stems from the following property:

Theorem 5: If f is Markov with respect to the partition $P = \{I_k\}$, then for each interval I_k , the restriction of f to I_k may be extended to $\tilde{f}_k$ with domain $\tilde{I}_k$, the closure of I_k , such that

$$\tilde{f}_k(x) = f(x), \quad x \in I_k$$

$\tilde{f}_k(x) = \lim f(y_n)$ where $\{y_n\} \in I_k$ and $y_n \rightarrow x \in \partial I_k$, the boundary of I_k

then $\tilde{f}_k(\tilde{I}_k)$ is the closure of a countable union of adjacent intervals from P .

Proof: Denote by a_k and b_k respectively the infimum and supremum of f in the interval I_k . Since f is Markov with respect to P , a_k and b_k are in Ω . Hence there exist countable adjacent open intervals I_{kj} such that the set

$$[a_k, b_k] \setminus \bigcup_k I_{kj} \quad (3)$$

only contains isolated elements from Ω (in fact: $\Omega \cap (a_k, b_k)$). By construction $\tilde{f}_k$ is a continuous function on $\tilde{I}_k$. The intermediate value theorem guarantees that for any y between a_k and b_k , and in particular for the points in (3) there exists an

$x \in (\tilde{f}_k^{-1}(a_k), \tilde{f}_k^{-1}(b_k))$ in I_k such that in I_k $f(x) = y$. Hence $\tilde{f}_k(\tilde{I}_k) = [a_k, b_k]$. •

Theorem 5 indicates that some of the interesting dynamical properties of the original system can be studied by "projecting" it onto the set of intervals P . Later on we shall make this notion more explicit. It only serves to say that the evolution of the system as seen in P behaves like a Markov chain. For this reason, we shall from now on refer to the partition P as the **MACROSCOPIC STATE SPACE** and the intervals of P as the **MACROSTATES**.

Let N be the index set for the intervals in the partitioning P , N is either the set of natural numbers, or the set of integers in case the partition is countably infinite, or it will be set $\{1, 2, 3, \dots, N\}$ if the cardinality of N is N . Clearly N is a representation of the macro state space introduced earlier.

The iterated maps f^n of Markov maps are continuous on each open interval of the form

$$I_{i(1)} \cap f^{-1}(I_{i(2)}) \cap f^{-2}(I_{i(3)}) \cap \dots \cap f^{-n+1}(I_{i(n)})$$

where the $i(j)$ belong to N .

The canonical macroprojection is the map $\pi : R \rightarrow N$, defined by $\pi(x) = k \iff x$ is in I_k , i.e. $\pi(x)$ identifies the macrostate to which x belongs.

The itinerary of x is the sequence

$$\pi(x), \pi(f(x)), \pi(f^2(x)), \dots$$

Associated with the macroscopic state space, we can introduce the notion of a macrostate transition matrix Π with ij -elements

$$\Pi_{ij} = m(f^{-1}(I_i) \cap I_j) / m(I_j) \quad (4)$$

where m is some (not necessarily finite) measure. It is that fraction of I_j which is mapped into I_i . As long as m is such that each macrostate has a nonzero measure, the macroscopic state space can be decomposed into transient and recurrent macrostates.

Finally, we shall on occasion also need another transition matrix $\underline{\Pi}$ with ij -elements

$$\underline{\Pi}_{ij} = \begin{cases} 1, & \text{if } f(I_j) \cap I_i \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

i.e., $\underline{\Pi}_{ij}$ indicates whether or not it is possible to go from I_j to I_i .

The macrostate $I_e (I_c)$ is **expanding (contracting)** for the Markov map with respect to the Markov partition P , if and only if for an expanding interval I_e and a contracting interval I_c we have for the usual Borel measure

$$\begin{aligned} m(f(I_e)) &\geq m(I_e) \\ m(f(I_c)) &\leq m(I_c) \end{aligned} \tag{6}$$

In fact, the above formulas can be used to define expansiveness and contractiveness of f with respect to a more general (nonuniform) Lebesgue measure, but we shall not pursue this yet.

2.4. Affine Maps

In this section, the map $f: \mathbb{R} \rightarrow \mathbb{R}$ is restricted to be piecewise affine (also called piecewise linear). This means that there exists a partition P of the real line in a finite or countable set of disjoint open intervals $\{I_k\}$, such that the map f restricted to the interval I_k is affine, i.e. there exists constants α_k and β_k such that for each x in I_k

$$f(x) = \alpha_k x + \beta_k \tag{7}$$

The set Ω contains all the points where f is not differentiable ("knee points" and points of discontinuity). The map f is then completely specified by Ω and the set of indexed pairs

$$\{(\alpha_k, \beta_k) ; k \in \mathbb{N}\} \tag{8}$$

corresponding to the intervals $\{I_k\}$ of P .

From Ω derive the family S of closure points of the intervals I_k . i.e. if $I_k = (a_k, b_k)$ is an interval in the partition P , then S contains the points a_k and b_k . Strictly speaking, we have

$$b_k = a_{k+1} = \omega_k \text{ for all } \omega_k \text{ in } \Omega \tag{9}$$

but using this double notation will facilitate the development below. It is also customary to denote them by ω_k^+ and ω_k^- . Noting that the infimum and supremum of $f(\cdot)$ in each interval I_k occurs at points of S , it can be expected that this set will play a major role. In particular, the piecewise affine map $f(\cdot)$ will be a Markov map in the sense defined above, if for every x in S , $f(x)$ is also in S . It follows then that if $f(I_k)$ intersects I_j , then I_j lies entirely inside $f(I_k)$. Furthermore, the iterated maps f^n are also piecewise affine, and affine on each open interval of the form

$$I_{i(0)} \cap f^{-1}(I_{i(1)}) \cap f^{-2}(I_{i(2)}) \cap \dots \cap f^{-n+1}(I_{i(n-1)}) \quad (10)$$

Because of the definition of S we also have that if $f^n(\omega) = \omega$, then the derivative of the iterated map $(f^n)'(\omega) \geq 0$.

In reference to the (piecewise affine) Markov maps defined above, it is clear that the condition i) implies that piecewise affine Markov maps can only have unstable equilibria. However, if a piecewise affine map has contracting intervals, then it does not necessarily follow that this map is non Markovian. (Of course, as long as the contracting part does not intersect the graph (x,x)).

By virtue of the specific applications we have in mind, we shall further assume that the piecewise affine maps are continuous (i.e. the values $f(\omega)$ are well defined for each ω in Ω). This restriction is however not required for any mathematical reasons. Moreover, it turns out that for dealing with higher dimensional cases, we will have to do away with this assumption.

Continuous piecewise affine maps can be entirely characterized by $f(0)$, the set Ω and the set of slopes α_k in each of the intervals of the partition P . For ease of bookkeeping, augment the set Ω with 0 and let the intervals of P to the right of the origin be indexed by positive integers, and the ones to the left by negative integers. i.e., for positive k

$$I_k = (a_{k-1}, b_k) \quad (11)$$

$$I_{-k} = (a_{-k}, b_{-(k-1)})$$

Letting α_k be the slope of $f(\cdot)$ in the interval I_k , then the map $f(\cdot)$ is evaluated by

$$f(x) = f(0) + \alpha_1 (b_1 - a_0) + \alpha_2 (b_2 - a_1) + \dots + \alpha_i (x - a_{i-1}) \quad (12)$$

for positive x in the interval I_1 , and by a similar formula for negative x .

For each x , define the function $A_n(x)$ as the "cumulative slope product in n iterations" i.e.

$$A_n(x) = \prod_{k=1}^n |\alpha_{\pi(f^k(x))}| \quad (13)$$

If x is a periodic point of prime period r , then $A(\langle x \rangle)$ is a shorthand notation for $A_r(x)$. $\langle x \rangle$ denotes then one cycle of the itinerary through x . If the macro state space is infinite, it will be assumed that all cycles are finite.

A consequence is (the proof is direct and omitted):

Lemma: If x is a periodic point of f , then $A(\langle x \rangle) = A(\langle f^k(x) \rangle)$ for all positive integers k . Hence $A(\langle x \rangle)$ is an invariant of the cycle through x .

Denote by Σ_N the set of sequences of N , and by Σ_A the subset all possible itineraries of f .

$$\Sigma_A = \{ s_0, s_1, s_2, \dots \mid \prod_{i=1}^{\infty} s_i = 1 \} \quad (14)$$

Then we have the following theorem (see e.g. [3]).

Theorem 6: Σ_A is a closed subset of Σ_N , and is invariant with respect to the shift σ_A .

The study of the invariant sets is of interest to the long term behavior of these maps, in order to assess the stability properties of the system they represent. It is well known from linear system theory that there is only one equilibrium point for linear systems. This is the origin. If the system is unstable,

then any nonzero initial condition will diverge to infinity, while the invariant distribution for a stable system is a singularity at the origin. On the other hand, it is known that many, even "simple" nonlinear maps allow an invariant distribution, equivalent to Lebesgue measure. Such measures have a density function. In many cases the invariant density for a system with initial condition x_0 may very well be independent from x_0 . The Birkhoff Ergodic Theorem [4] gives a condition for this independence: f must be an endomorphism, i.e. f is onto and for each measurable set $\mu(A) = \mu(f^{-1}(A))$. The system behaves then as if it were driven by a stochastic process, even though it is in fact entirely deterministic. So some pertinent questions are the existence of invariant densities. Once this has been established, the uniqueness of the invariant densities needs to be determined. Finally, one needs to find finite algorithms to compute the invariant density or at least an approximation of it.

Due to the complexity, one might resort to computer simulations of such systems in order to solve the above problems. However, a direct approach by simulating the system equations may be doomed to failure, exactly because of the possible chaotic nature of the map.

As an example consider the following pathological case. Let f be the map

$$f : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \begin{cases} 0 & \text{if } x \text{ is in } Q \\ \alpha x & \text{if } x \text{ is in } \mathbb{R} \setminus Q \end{cases} \quad (15)$$

Computer simulation of the system $x_{k+1} = f(x_k)$ will invariably lead to a stable equilibrium at 0. The exact iteration will however diverge to infinity for almost any initial condition (with respect to the Borel measure) if $\alpha > 1$. In the stochastic case, the addition of a standard white gaussian noise w_k will destabilize it with probability one even in the case $\alpha = 1$, independently of the initial condition. The computer simulation will yield

$$x_{k+1} = [w_k]$$

where $[w_k]$ is the machine representation of the noise sample, and is therefore necessarily a rational number. The invariant distribution one would conclude from a simulation would be the standard gaussian. Finally note that the discontinuities in f are finite in number.

Another, perhaps more realistic example of such misbehaving computations is the map $f(x) = 2^k x \pmod{1}$, due to Li [5]. Restrict the domain to the interval $[0,1]$. The unique invariant measure is known to be Lebesgue measure. Consider now $x \geq 2^{-k}$ stored in a n -bit computer. After sufficiently many iterations, all initial conditions will end up at zero. It seems that this misbehavior could be avoided by combining the left shift (multiplication by 2) with an addition of a random bit in the lsb position. (after all, the transformation is producing information e.g. see Shaw [6]). As this is equivalent to the addition of noise, it seems that this may help in the computation of the invariant distribution. This paper will on theoretical grounds discuss the interplay between the exact invariant distribution, and the distorted invariant distribution.

The reason for our concern with these problems is that the knowledge of the invariant distribution is of importance in obtaining approximate filtering algorithms for nonlinear systems.

The topics of existence, uniqueness, and computation of the invariant density are dealt with in the following section.

2.5. Invariant Measures for Piecewise Affine Markov Maps

In order to find the dynamics of the distribution of the deterministic system many approaches can be taken. For continuous time systems, we may take a stochastic point of view, and consider a dynamical system, perturbed by external noise. (This is the approach that will be considered in the stochastic section of the paper). Even with a well specified initial condition, a smooth probability density for the state may evolve. The well known Fokker-Planck equation describes the evolution or flow of

the density. In fact, the density itself may be viewed as the underlying dynamical quantity, whose dynamics are then governed by the Fokker-Planck equation. Such an approach does however necessitate the use of (infinite dimensional) Banach or more general function spaces. Conceptually we can then let the variance of the external noise approach zero, so as to obtain the original deterministic system. With random initial conditions, these are usually referred to as crypto-deterministic systems, and have received a great deal of attention in statistical mechanics. In this limit case, there will obviously no longer be a diffusion term in the Fokker-Planck equation. The resulting first order partial differential equation is commonly referred to as the Liouville equation.

The evolution operator induced by the Liouville equation has its discrete analog in the Frobenius-Perron operator. If f is a measurable function, nonsingular with respect to the Lebesgue measure, mapping $[0,1]$ into $[0,1]$, then the operator

$$P_f \varphi(x) = \frac{d}{dx} \int_{f^{-1}[0,x]} \varphi(s) ds \quad (16)$$

has the properties:

- 1) P_f is positive (i.e. if $\varphi \geq 0 \implies P_f \varphi > 0$)
- 2) P_f preserves integrals

$$\int_0^1 P_f \varphi \, dm = \int_0^1 \varphi \, dm \quad \varphi \in L_1([0,1]) \quad (17)$$

- 3) $P(f^n) = (P_f)^n$
- 4) $P_f \varphi_0 = \varphi_0 \iff d\mu = \varphi_0 \, dm$ is invariant under f
(i.e. for all measurable A : $\mu(f^{-1}(A)) = \mu(A)$)

Remark: The Perron-Frobenius operator can be defined more generally, in its integrated form, for abstract measure spaces [7].

The invariant density of the map f is then nothing else than the fixed point of the Perron-Frobenius operator. The following theorem by Lasota and Yorke [8] expresses the existence of the invariant density φ^* under certain conditions.

Theorem 7: If $f: [0,1] \rightarrow [0,1]$ is

a) piecewise C^2 (with a finite partition)

b) $\inf |(f^n)'| > 1$ for some n ,

then: for all $\varphi \in L_1[0,1]$

$$(1/n) \sum_{k=0}^{n-1} P_f^k \varphi \rightarrow \varphi^* \in L_1 \quad (18)$$

where the convergence is in the norm. Furthermore φ^* has the following three properties:

1) $\varphi \geq 0 \Rightarrow \varphi^* \geq 0$

$$2) \int_0^1 \varphi^* dm = \int_0^1 \varphi dm$$

3) $P_f \varphi^* = \varphi^*$, consequently $d\mu^* = \varphi^* dm$ is invariant under f .

If instead of b), we have the more restrictive condition,

$$b') \quad \inf_x |f'(x)| > 1$$

then also 4) holds

4) φ^* has bounded variation. In fact, there is a constant c , independent of φ such that

$$\text{Variation}(\varphi^*) \leq c \|\varphi\|$$

The piecewise affine maps introduced earlier fall within the class of the theorem, as long as there are a finite number of intervals in the partition. Obviously such a map is piecewise C^2 , and

$$\inf_x |f'(x)| > 1 \Leftrightarrow \inf_N |\alpha_k| > 1$$

or

$$\inf_x |f^n(x)'| > 1 \quad \text{for some } n$$

$$\Leftrightarrow \inf_x A_n(x) = \prod |\alpha_\pi[f^k(x)]| > 1$$

The condition a) of the theorem is automatically satisfied. The condition b') seems restrictive. It means that all regions of continuity of f must be expanding. The less restrictive condition (at the cost of loosing the bounded variation property) leads to the following

Theorem 8: A (finitely) piecewise affine map has an invariant density if there exists an integer n_0 such that for all initial conditions, the average expansion over n_0 iterations is positive.

The average expansion in n steps is defined as

$$\frac{1}{n} \sum_{k=0}^{n-1} \beta_\pi[f^k(x)] \quad (19)$$

where

$$\beta_i = \log |\alpha_i| \quad (20)$$

Note that for general piecewise affine maps, still a condition needs to be checked over all initial conditions. For many iterations (n_0) the resulting iterated map rapidly becomes very jagged. It is here that the subclass of Markovian maps is extremely useful. It allows to restrict the search of the average growth condition over the macrostates N rather than the original state space. Using the earlier defined transition matrix Π the subset of all allowable sample sequences Σ_A may be constructed. If these sequences are represented as a horizontal tree, for which the slopes of the branches correspond to the "growth" factor β , then the above theorem states that for some level (of branching) n_0 , all nodes at level n_0 of the tree must be above the zero-level line.

In case the map f is piecewise C^2 , but with a countably infinite partition, the extension of the theorem of Lasota and

Yorke [8] needs to be used. Its conditions are however a lot more restrictive

Theorem 9: Let $f: [0,1] \rightarrow [0,1]$ be countably piecewise C^2 such that

$$i) \inf_x |f'(x)| > 2, \quad \sup_x |f''(x)| < \infty$$

ii) in almost all macrostates, f is onto $[0,1]$,
then the conclusions of Theorem 8 remain valid.

Clearly, in the context of piecewise affine maps, the condition ii) in particular restricts the maps to very special, uninteresting cases. A more interesting variant is based on Adler's theorem [9,10] for a map, $f: I \rightarrow I$, satisfying the properties:

There exists a partition P of the interval I into a finite or countable collection of disjoint open intervals $\{I_k\}$ such that

- 1) f is defined on $\bigcup I_k$, and $m(I \setminus \bigcup I_k) = 0$
- 2) $f|_{I_k}$ is strictly monotonic and extends to a C^2 function on the closure of I_k .
- 3) If $f(I_k) \cap I_j \neq \emptyset$, then $f(I_k)$ contains I_j .
- 4) There is an R so that

$$\bigcup_{n=1}^R f^n(I_k) \text{ contains } I_j \text{ for all } k \text{ and } j$$

Theorem 10 (Adler): Let $f: I \rightarrow I$ satisfy the above properties, and let

$$M = \sup_{I_k} \sup_{y, z \in I_k} \left| \frac{f''(z)}{f'(y)^2} \right| < +\infty \quad \inf_x |(f^n)'(x)| > 1 \quad (21)$$

then f admits an invariant finite measure $d\mu = \phi(x)dx$ with $\phi(x)$ bounded away from 0 and $+\infty$.

If this is applied to the the piecewise (finite or countable) affine Markov maps, the sufficient conditions are that the average expansion for some n_0 is strictly positive, and f is Markov map, corresponding to a positive recurrent chain. The type of map

defined above is called "Markov map" in the literature, but it is obviously more restrictive than our definition of Markov maps. Note that the theorem still requires the expansiveness of the map at all points. Bowen [11] provided a method ("inducing"), which enabled the conditions of Adler's theorem to be relaxed to certain nonexpansive maps.

Another method exists to extend the range of validity of the above theorems by introducing the equivalence of Conjugation. Two transformations $f: I \rightarrow I$ and $g: J \rightarrow J$ are conjugate if there exists a homeomorphism $h: I \rightarrow J$ such that (see e.g. [3])

$$g(x) = h(f(h^{-1}(x))) \quad (22)$$

This is a generalization of the notion of a similarity transformation in linear system theory. It is then straightforward to show that if h is differentiable, and if ϕ_f^* is the stationary density of f , then ϕ_g^* , the stationary density of g , is given by

$$\phi_g^*(x) = \phi_f^*(h^{-1}(x)) \left| \frac{dh^{-1}(x)}{dx} \right| \quad (23)$$

Hence in order to check the existence of an invariant density, it suffices to find a conjugate map which is known to have an invariant density. The conjugating function h does not need to be piecewise linear. Grossmann and Thomae [12] have shown that conjugating functions may be constructed by relating two dynamical laws to each other. The resulting conjugating function h may have the structure of a Cantor function however.

Finally, conditions for the uniqueness of the stationary density were derived by Li [5,12]. Li also provided a converging approximation method to compute the unique invariant density. The proof of convergence settled a long standing conjecture by Ulam [13]. The approximation stems from considering an arbitrary partitioning of the domain in finitely many disjoint intervals. On these a Markovian transition matrix Π is defined. With this construct, the Perron-Frobenius operator is effectively

approximated (exact for Markov Maps) by this transition matrix, Π . The fixed point is then very efficiently found by a quadratic programming problem, i.e.

$$\text{minimize } \|\Pi \varphi - \varphi\| \text{ subject to } \varphi_i \geq 0 \text{ and } \sum \varphi_i = 1$$

2.6. Reconciliation with the theory of Markov Chains

This paragraph justifies the name "Markovian Map" introduced earlier for a certain class of maps. Indeed, the transition matrix Π for the macrostates introduced in (4), satisfies

$$\begin{aligned} \Pi_{ij} &\geq 0 \text{ for all } i, j \text{ in } N \\ \sum_i \Pi_{ij} &= 1 \text{ for all } j \text{ in } N \end{aligned} \tag{24}$$

It can be shown that given the above properties for a matrix Π and given any distribution $\{\varphi_i\}$ such that

$$\begin{aligned} \varphi_i &\geq 0 \\ \sum_i \varphi_i &= 1 \end{aligned}$$

there exists a probability space, and random variables X_n , $n \geq 0$, on that space satisfying the Markovian property (a kind of "Huygens principle" obeyed by nonhereditary systems).

$$P(X_0=x_0, \dots, X_n=x_n) = \Pi_{x(n), x(n-1)} \dots \Pi_{x(1), x(0)} \varphi_{x(0)} \tag{25}$$

Therefore all of the theory of Markov chains becomes available in the theory for crypto-deterministic systems. In particular, the decomposition theorem of the state space leads to a decomposition theorem for the macrostate space.

There are recurrent and transient states. For a recurrent state, the probability that the chain starting in that state returns to that state is one, while it is strictly less than one for a transient state.

Theorem 11: The set of recurrent states is the union of a finite or countably infinite number of disjoint irreducible closed sets.

The proof can be found in any elementary book on stochastic processes, e.g. [14]. It basically expresses the fact that "two way communication" among states is an equivalence relation. If, for a recurrent state the mean return time is finite, then the state is called positive recurrent. If the mean return time is infinite, the state is called null recurrent. If S is a finite irreducible closed set, then all states in S are positive recurrent. For the set of positive recurrent states Σ_+ (strongly ergodic states), the following is well known theorem [15].

Theorem 12: i) If $\Sigma_+ = \emptyset$ then no stationary distribution exists

ii) If Σ_+ is nonempty and irreducible, then a unique stationary distribution exists

iii) If Σ_+ is nonempty and reducible, then infinitely stationary distributions exist.

The Markov Chain theory gives information about the behavior of the Markov maps introduced earlier. The problem is that it only provides information on the macrostate space, and not the microstate space. In fact, as evidenced by Sarkowskii's theorem the structure at the microscopic level can be much more rich than what can be inferred from the macroscopic picture. And, what is worse, the existence of a macroscopic stationary distribution, does not necessarily imply the existence of a microscopic stationary distribution.

2.7. Minimal Systems

In linear system theory, an important role is played by the invariant subspaces, especially in realization theory. Characterization of the invariant subspaces lead to the decomposition of the system into reachable and non-reachable subsystems, observable, and non-observable ones. A linear system is minimal if there is no lower order system realizing the same input-output map. Minimal systems are important because of their joint reachability and observability. It is known that the role of

invariant subspaces for linear systems can be extended to nonlinear system theory [16]. We shall first extend the notion of minimality. The proper way to do this turns out not to relate solely on the dimension, or order of the system. Rather, the invariant distributions are taken as the defining property.

Definition: A system $x_{k+1} = f(x_k)$ is minimal if no proper subset of the state space is invariant under the action of f .

In this section, we consider the separation of the dynamical system into minimal subsystems. By definition, each subsystem is invariant. Hence, the knowledge of the initial condition allows one to delimit the appropriate state space. (This is somewhat an abuse of terminology, since in the proper sense, the set of states does not have the structure of a vector space).

Suppose now that it is known that the disjoint intervals $J_1, J_2, \dots$ are invariant for the restrictions, respectively: $f|_{J_1}, f|_{J_2}, \dots$. Without loss of generality, we assume also that these intervals are disconnected, in the sense that for any two intervals J_i, J_j ,

$$\text{span}(J_i, J_j) \setminus J_i \cup J_j \neq \emptyset$$

$\text{Span}(J_i, J_j)$ is the closure of the convex linear combination of points of $J_i \cup J_j$. The problem then is to find the proper conditions under which the intervals, known to be invariant under the restriction of f to these intervals, remain invariant under f .

Geometrically, the picture is obvious for continuous f . If $J = (a, b)$ is invariant under $f|_J$, then by the definition of invariance, the set $f|_J^{-1}(J) \subset J$ and $f|_J(J) \subset J$. However, invariance under f also requires $f^{-1}(J) \subset J$ and $f(J) \subset J$. A necessary condition for this is that $f(a) = a$ and $f(b) = b$, as can be seen very easily from a geometric picture.

Thus the problem then boils down to the properties of the map f in the interconnecting intervals of the form

$$\text{span}(J_i, J_j) \setminus J_i \cup J_j \neq \emptyset$$

Several possibilities can occur:

- a) The above interval is itself invariant, then the whole of span (J_i, J_j) is invariant.
- b) If say the conditions $f(a) = a$ and $f(b) = b$ are not satisfied for either J_i or J_j , then "spillover" will occur, and the intervals invariant under the restricted map will no longer be invariant under f .
- c) If the interconnecting interval contains a stable equilibrium, then it will also contain a trapping set, and a smooth invariant distribution will not exist.

The above discussion leads to the following theorem:

Theorem 13: Intervals invariant under the restriction of $f|_J$, either remain invariant (for which a necessary condition is that the graph of f leaves the "box" $(J, f(J))$ at the diagonal points $(\partial J, f(\partial J))$, or can be embedded into larger invariant intervals. If the intervals remain invariant, then the interconnecting interval breaks down into alternating trapping sets, and invariant (sub)intervals.

Note that if J_1 and J_2 are adjacent open invariant intervals, then the closure of their union will be invariant, however, each constitutes a minimal system.

Finally, we shall also note that a trapping set is characterized by

$$f(\partial I_\alpha) = \partial f(I_\alpha)$$

where ∂I is the set of limit points (boundary) of I .

2.8. The Connection with the Stochastic Case

In the previous sections, we investigated the minimal systems. These were characterized by the fact that a smooth invariant density exists in the state set. As a result, a phenomenon, called **deterministic diffusion** occurs, even though no stochastics entered the system. [17]

Consider now the case where a stochastic driving term enters in the system equation. The local transition model is then:

$$x_{k+1} = f(x_k) + \sigma(x_k) u_k \quad (26)$$

where the sequence $\{u_k\}$ is a standard white gaussian noise, uncorrelated with the initial state of the system x_0 . In this paper we shall only deal with the simple case, where the variance σ^2 is constant over all of R .

The implication of this assumption is that the original invariant distributions of the (minimal) subsystems will be "clouded" over into one smooth invariant distribution [18]. Of course, we assume that such an invariant distribution exists. A sufficient condition is, roughly speaking, that the system is "eventually stable" for sufficiently large states x . More precisely:

$$\lim_{x \rightarrow \infty} \sup |f(x)|/|x| < 1 \quad (27)$$

If the Lebesgue measures of the minimal state spaces are large compared to the variance s , then the overall stochastic invariant distribution will be "close" in a sense to be made more precise later to a convex sum (mixture) of the invariant distributions of the deterministically minimal systems.

If, on the other hand, the noise variance s is much larger than the measures of the minimal systems, then it is expected that the steady state distribution will be close to the gaussian distribution, and in fact computable from a global linearization of the original nonlinear system. Clearly, the intermediate cases are the difficult ones.

Several approaches are possible for the solution of this problem. One is to consider the space of all densities, and define locally the system by an evolution in this infinite dimensional space. Its advantage is that no stochastics as such needs to enter in the picture, and the deterministic concepts can be used, albeit for an infinite dimensional system. Alternatively, and the route taken here, a new macrostate transition is computed or

approximated from consideration of the original transition probability matrix Π and the "overlap" of the s -neighborhoods with adjacent domains.

3. STOCHASTIC SYSTEMS

In this section we consider the properties of the dynamic affine system when it is driven by a white Gaussian noise sequence. The model to be considered is given by

$$x_{k+1} = f(x_k) + w_k \quad (28)$$

where $f(x)$ is a piecewise affine map as defined above, and $\{w_k\}$ is a white Gaussian noise sequence with zero mean and variance σ^2 . In the deterministic case it has been shown that the system can be characterized by invariant aggregations of the macrostates in P . It has also been found that for some contracting macrostates (namely, ones with intersections with the line $f(x) = x$) we obtain equilibrium points inside these intervals. The effect of the noise is to allow transitions out of these intervals. The probability of these transitions can be made very small if the noise variance is small. For aggregations containing expanding intervals we obtain minimal subsystems with invariant distributions without leaving the aggregation of such intervals. The addition of noise to the system will allow for interactions among these minimal states and will smooth the resulting distribution on the aggregations of macrostates. Obviously the noise should allow for more interactions among the macrostates if some restrictive assumptions are made. If the noise variance is small relative to the measure of the contracting macrostates, the results should allow for a high probability of the system staying around the equilibrium point if it exists for such macrostates, and some small probability of a transition to other macrostates. Alternatively, if the noise variance is assumed to be large relative to the measure of expanding macrostates then the system

is expected to escape these states with high probability. We formally derive the conditions that allow the approximation of such a behavior by a Markovian transition among purely linear dynamic models, for the following simple problems. This approximation has been considered for the three region scalar case in [19].

Consider the affine model for $f(x)$

$$f(x) = \alpha_i x + \beta_i, \quad a_i < x < b_i, \quad i = 1, 2, \dots, N \quad (29)$$

where the interval $I_i = (a_i, b_i)$ represents a macrostate of the system. Suppose the probability density of x_k is given by $p_k(x)$, then we can write the density as a sum of conditional densities $p_{ki}(x)$ given by the following expressions

$$p_k(x) = \sum_{i=1}^N P_{(k-1)i} p_{ki}(x) \quad (30)$$

$$p_{ki}(x) = \frac{1}{P_{(k-1)i}} \int_{y \in I_i} \frac{1}{\sigma} q\left(\frac{x - \alpha_i y - \beta_i}{\sigma}\right) p_{(k-1)}(y) dy \quad (31)$$

where the P_{ki} are the probabilities of being in macrostate I_i at time k , and are given by

$$P_{ki} = \int_{x \in I_i} p_k(x) dx, \quad (32)$$

and where $q(x)$ is the unit Gaussian density function,

$$q(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2). \quad (33)$$

It should be noted that (31) is a recursive relation that determines the evolution of the density of the state of the system x_k , assuming we started from some initial density. If we now also define the transition probability among the macrostates represented by the intervals $\{I_i\}$ by $\Pi(k)$ with ij -elements

$$\Pi_{ij}(k) = P \{ x_{k+1} \in I_j \mid x_k \in I_i \} \quad (34)$$

then the transition probabilities are given by the expression

$$\Pi_{ij}(k) = \int_{x \in I_j} P_{(k+1)i}(x) dx. \quad (35)$$

We now assume that a steady-state solution to the problem given above exists, which would under the restrictions imposed by the deterministic model, namely that an invariant distribution exists. The problem is to determine the condition for the convergence of this model to the one given by the Markov switched model, formulated as

$$x_{k+1} = \alpha_i x_k + \beta_i + w_k, \quad \text{when } S(k) = i \quad (36)$$

where $S(k)$ is the state of an underlying finite state Markov process taking values $i = 1, 2, \dots, N$, and has a transition probability matrix given by Π . The approximation is in the assumption that the Markov process is independent of the original state of the system, and in assuming that the linear models are not restricted by the values of the state. This approximation implies that the process $\{S(k)\}$ represents the macrostates of the original system.

The steady state solution satisfies the relations

$$p(x) = \sum_{i=1}^N P_i p_i(x) \quad (37)$$

$$p_i(x) = \frac{1}{P_i} \int_{y \in I_i} \frac{1}{\sigma} q\left(\frac{x - \alpha_i y - \beta_i}{\sigma}\right) p(y) dy \quad (38)$$

$$P_i = \int_{x \in I_i} p(x) dx, \quad \int p_i(x) dx = 1 \quad (39)$$

$$\Pi_{ij} = \int_{x \in I_j} p_i(x) dx. \quad (40)$$

The transition matrix Π leads in turn to steady state

probabilities of the states $S(k)$, given by $\{P_i\}$, as can be seen from (37)-(40). The question becomes therefore to find the conditions under which $p_i(x)$ is approximately equal to the stationary density of the linear model given by the parameters α_i and β_i . We are only interested in sufficient condition, hence we restrict our attention to the cases of relatively large contracting regions and small expanding regions relative to the noise variance. The effects of each of these regions will be considered in the following subsections.

First we shall postulate certain assumptions on the Transition probability matrix Π and then will show that these are satisfied for the resulting system under the imposed condition. It will be assumed that the contracting regions are large relative to the noise variance, hence we postulate that

$$1 - \Pi_{ii} \ll 1, \text{ for } I_i \text{ contracting} \quad (41)$$

which implies that the escape probability from such regions is very small. Similarly, it will be assumed that the expanding regions are small relative to the noise variance, hence we postulate that

$$\Pi_{ii} \ll 1, \text{ for } I_i \text{ expanding} \quad (42)$$

which implies that the system exits an expanding region with high probability. The resulting solutions for the stationary probabilities P_i satisfying the equation

$$P_j = \sum_i P_i \Pi_{ij} \quad (43)$$

have the properties that the P_i are very small for expanding regions. These assumptions will be used to derive the validity of the switched Markov approximation, and in turn they too will be verified.

3.1. Contracting Intervals

Since we assume that the stationary probabilities of the expanding regions are small, we may restrict our attention only to

the contributions of the contracting regions. For large contracting intervals the steady state stationary probability density of the switched Markov approximation is given by a weighted sum of Gaussian densities, $q_i(x)$, with means

$$\mu_i = \frac{\beta_i}{(1 - \alpha_i)} \quad (44)$$

and variances

$$\sigma_i^2 = \sigma^2 / (1 - \alpha_i^2). \quad (45)$$

Since the regions are contracting, the mean μ_i falls within the interval I_i ,

$$a_i < \mu_i < b_i.$$

In order for the approximation to be valid the length of the intervals must be large relative to the variances of stationary densities, i.e.

$$(b_i - a_i) > 2m \sigma_i, \quad m \geq 3 \quad (46)$$

where m may be restricted to be larger than the three-sigma range to ensure that the contributions of the tails can be made as small as desired. Under these assumptions the effect of the contracting regions on the equation for $p_i(x)$ given in (38) may be obtained by substituting the densities $q_i(x)$ in the right hand side, which will show that the resulting density is approximately equal to $q_i(x)$. The substitution in (38) yields

$$p_i(x) = q_i(x) - \int_{y \notin I_i} (1/\sigma) q[(x - \alpha_i y - \beta_i)/\sigma] q_i(y) dy \\ + \sum_{j \neq i} (P_j/P_i) \int_{y \in I_i} (1/\sigma) q[(x - \alpha_i y - \beta_i)/\sigma] q_i(y) dy \quad (47)$$

It is easy but tedious to show that due to the restrictions on the relative sizes of the intervals and the variances given in (46),

the terms in (47) except the first can be made arbitrarily small by properly selecting the value of m .

The effect of the expanding regions, which have small probabilities P_i , on (47) can also be made as small as desired.

Finally, the resulting transition probabilities Π_{ij} are given by

$$\Pi_{ij} = 1 - \Phi[-(\mu_i - a_i)/\sigma_i] - \Phi[-(b_i - \mu_i)/\sigma_i] \quad (48)$$

where $\Phi(x)$ is the cumulative unit Gaussian distribution, namely

$$\Phi(x) = \int_{-\infty}^x q(y) dy. \quad (49)$$

In view of the assumption (46) it is seen from (48) that the transition probabilities satisfy the postulated properties.

These results can be summarized by the following theorem,

Theorem 14: Let the system model be given by (28)-(29) and let the contracting intervals satisfy the assumption (46), then the stationary density on these intervals, $p_i(x)$, can be approximated by the stationary density of the switched Markov model (36).

3.2. Expanding Intervals

The expanding intervals in the switched Markov model do not exhibit a stationary density, due to their instability. In this case we can approximate the density on these regions by using a first order approximation. The approximation is based on using only the first order term of the transition probabilities to the expanding regions in the switched Markov model. The approximate density $q_i(x)$ of the expanding intervals in the Markov model is given by

$$q_i(x) = \sum_j P_j \Pi_{ji} q_{ji}(x) + \text{higher order terms} \quad (50)$$

where the $q_{ji}(x)$ are Gaussian densities with means μ_{ji} and variances σ_{ji} given by the expressions

$$\mu_{ji} = \alpha_i \mu_j + \beta_i \quad (51)$$

$$\sigma_{ji}^2 = \sigma^2 + \alpha_i^2 \sigma_j^2 \quad (52)$$

We now substitute the $q_i(x)$ for the contracting regions in the right hand side of (38) we obtain a solution for the $p_i(x)$ of the expanding regions on the left side of (38). The resulting solution can be made arbitrarily close to the solution of the Markov model provided the following two conditions are satisfied for expanding regions:

$$(b_i - a_i) \ll \sigma \quad (53)$$

$$\alpha_i (b_i - a_i) > 6 \sigma \quad (54)$$

These condition imply that the regions have to be small but with a high slope. The results indeed show that the transition probabilities Π_{ij} are indeed very small for expanding regions. The result can be summarized in the following theorem,

Theorem 15: Under the conditions of Theorem 14, if also conditions (53)-(54) are satisfied then the density function of the expanding intervals of the system (28)-(29) can be approximated by the switched Markov model.

In order to cover the case of small expanding region but with relatively small slope, it is possible to use series expansions as shown in [19]. The series expansion given in [19] indicates that the small region can be absorbed by the adjoining regions as a first order approximation. The approximation is also valid for small contracting regions. The results can be used to construct approximate nonlinear filters for such nonlinear systems as described in [20].

4. EXAMPLE

A first order example with three regions is considered. The expressions for the resulting densities are explicitly derived. The nonlinearity is assumed to be symmetric, so that only densities for positive values of x need to be displayed. The three regions are normalized with the expanding region being (-1,

+1) and a having a slope of α_0 , while the contracting regions are $(+1, +\infty)$ and $(-\infty, -1)$ and having a slope of α_1 . The equilibrium points for the contracting regions are at

$$x^* = (\alpha_0 - \alpha_1)/(1 - \alpha_1) > 1 \quad (55)$$

since $\alpha_0 > 1$, and $-1 < \alpha_1 < 1$. The value of the noise variance σ and the values of the parameters α_i 's can be selected to verify the assumptions made on the validity of the approximation. The system satisfies the assumptions discussed in the paper if $\sigma \ll (x^* - 1)$. The nonlinear system was simulated numerically and the transition probability of the actual system was computed and compared to the Markov model. The two are in agreement for σ as high as 10 when the slopes are selected as $\alpha_0 = 10$ and $\alpha_1 = -0.2$, even though the assumptions require σ to be smaller than 10 by a factor of 3. The transition probabilities for these cases and the corresponding stationary probabilities are shown in Table 1. The resulting distribution for one of the cases is shown in Figure 1. The result corroborate the theoretical results discussed in the paper.

TABLE 1.

The Transition Matrix Π and the Stationary Probabilities P

	$\sigma = 3$			$\sigma = 5$			$\sigma = 10$		
$\Pi =$	.990	.008	.002	.928	.049	.023	.773	.046	.181
	.450	.100	.450	.443	.113	.443	.469	.063	.469
	.002	.008	.990	.023	.049	.928	.181	.046	.773
$P =$	.495	.010	.495	.474	.053	.474	.476	.048	.476

5. SUMMARY AND CONCLUSIONS

In this paper an approximation for piecewise linear scalar discrete-time system was derived. It was based on the properties of such systems in the deterministic case. In the stochastic case the approximation depends on the relative size of the variance of the driving Gaussian noise. The approximation may be extended to multidimensional systems, and can be applied to the derivation of

nonlinear filtering schemes.

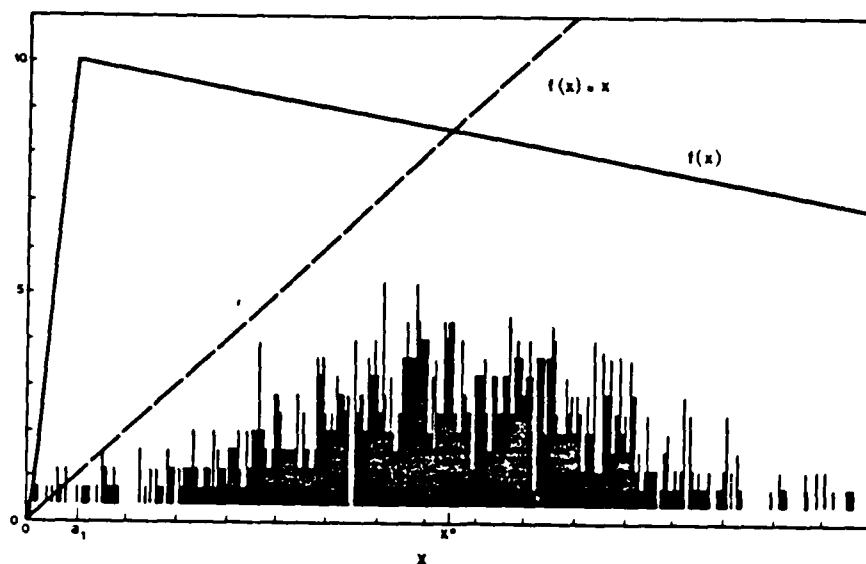


FIG. 1.

The Distribution of the Stationary Density for $\sigma = 3$.

ACKNOWLEDGEMENTS

This research is supported by the U. S. Air Force Armament Laboratory, under Contract F08635-84-C-0273.

REFERENCES

- [1] V.I. Arnold and A. Avez, "Ergodic Problems of Classical Mechanics", Benjamin, 1968.
- [2] P.R. Halmos, "Lectures on Ergodic Theory", Chelsea, 1956.
- [3] R. Devaney, "An Introduction to Chaotic Dynamical Systems, Benjamin/Cummings, 1985.
- [4] I.P. Cornfeld, S.V. Fomin and Ya. G. Sinai, "Ergodic Theory", Springer-Verlag, 1982.
- [5] T.Y. Li, "Finite Approximation for the Frobenius-Perron Operator. A Solution to Ulam's Conjecture.", J. Approx. Thy. 17, 177-186, 1976.
- [6] R. Shaw, "Strange Attractors, Chaotic Behavior, and Information Flow", Z. Naturforsch. 36a, 80-112 (1981).
- [7] A. Lasota and M. C. Mackey, "Probabilistic Properties of Deterministic Systems", Cambridge Univ. Press, 1985.

- [8] A. Lasota and J. A. Yorke, "On the Existence of Invariant Measures for Piecewise Monotonic Transformations", Trans. Am. Math. Soc. Vol. 186, 481-489, December 1973.
- [9] R. L. Adler, "F-Expansions Revisited", Springer Lecture Notes 318, pp. 1-5, 1973.
- [10] R. L. Adler, "Continued Fractions and Bernoulli Trials", in Ergodic Theory, Moser, J., Phillips, E., Varadhan, S. (eds.) Lecture Notes. Courant Inst. Math. Sci. 1975.
- [11] R. Bowen, "Invariant Measures for Markov Maps of the Interval", Commun. Math. Phys., 69, pp. 1-17, 1979.
- [12] S. Grossmann and S. Thomae, "Invariant Distribution and Stationary Correlation Functions of One-Dimensional-Discrete Processes", Z. Naturforsch. 32a, 1353-1363, 1977.
- [12] S. M. Ulam, "A Collection of Mathematical Problems", Interscience Tracts in Pure and Applied Math., 8, pp. 73, 1960.
- [14] S. Karlin and H. M. Taylor, "A First Course in Stochastic Processes", Academic Press, 1975.
- [15] P. G. Hoel, S. C. Port, and C. J. Stone, "Introduction to Stochastic Processes", Houghton Mifflin Co., 1972.
- [16] A. Isidori, "Nonlinear Control Systems: An Introduction", Springer Verlag, Lecture Notes in Control and Inf. Sci., 1985.
- [17] T. Geisel and G. Nierwetberg, "Onset of Diffusion and Universal Scaling in Chaotic Systems", Phys. Rev. Lett. 48, No. 1, 7-10, 1982.
- [18] K. Matsumoto and I. Tsuda, "Noise-Induced Order", Journal of Statistical Physics, Vol. 31, 1983, pp. 87-106.
- [19] A. H. Haddad and E. I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, Princeton University, pp. 202-206, March 1984.
- [20] 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.

APPENDIX C

PIECEWISE LINEAR MODELING OF MULTIDIMENSIONAL STOCHASTIC
NONLINEAR SYSTEMS

PIECEWISE LINEAR MODELING OF MULTIDIMENSIONAL STOCHASTIC NONLINEAR SYSTEMS

A. H. HADDAD AND E. I. VERRIEST
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

This paper extends to the multi-dimensional case an approximation of nonlinear dynamic systems with random inputs, by a set of linear models with Markovian transitions among the models.

SUMMARY

The problem of piece-wise linear approximation of nonlinear dynamic systems subject to stochastic input has been considered for the scalar case in [1-2]. The approach was used to obtain an approximate model for the discrete-time nonlinear system described by Markovian transitions among linear models. The resulting approximation was then utilized to derive an approximate filtering scheme for the nonlinear system. The analysis has been restricted to the scalar case for simplicity, and to illustrate the concepts. However, the extension to the multi-dimensional case may appear to be simple in principle, but it presents several difficulties. The objective of this paper is to provide a systematic approach to the approximate multivariable modeling problem for nonlinear systems driven by stochastic inputs. The paper also briefly describes the applications of the model to the approximate solution of the nonlinear filtering problem.

The discrete-time nonlinear system considered has the form

$$x_{k+1} = f(x_k) + w_k \quad (1)$$

where x_k is the n -dimensional state vector, and $\{w_k\}$ is a white Gaussian noise sequence with zero mean and covariance matrix Q . The nonlinearity $f(x)$ is assumed to be given by a piece-wise linear approximation

$$f(x) = A_i x + b_i, \quad x \in \Omega_i. \quad (2)$$

The properties and nature of the approximating model for the scalar case depend on the stability or instability of the linear models, and the size of their regions, Ω_i , relative to the noise standard deviation. In order to extend the notion of stability and size of the regions to the multi-dimensional case a new definition as to stability is needed, and a more systematic approach to the region definition is required.

First a measure $\mu(\cdot)$ on R^n is defined relative to noise covariance matrix Q . If the covariance matrix is singular the measure may be based on the covariance of the state of an equivalent linear system that is obtained from (1) by least squares approximations. The regions Ω_i may now be classified as to their being contracting or expanding (as opposed to stable or unstable), using modified assumptions. A region Ω_i is said to be Contracting (C) if it satisfies the assumption:

$$(C1) \quad \mu[f(\Omega_i)] \leq \mu(\Omega_i)$$

Similarly, a region Ω_i is said to be Expanding (E) if it satisfies

$$(E1) \quad \mu[f(\Omega_i)] > \mu(\Omega_i)$$

where the notation $f(\Omega_i)$ is used for the image of the region Ω_i under the transformation $f(x)$. The class of nonlinearities considered here is further restricted to satisfy:

$$(C2) \quad f(\Omega_i) \subset \Omega_i$$

$$(C3) \quad \mu(\Omega_i) \gg 1$$

$$(E2) \quad \Omega_i \subset f(\Omega_i)$$

$$(E3) \quad \mu(\Omega_i) \ll 1$$

for contracting and expanding regions, respectively. These assumptions (similar to the scalar case) restrict the class of nonlinearities, but are required for the validity of the approximate Markov model.

The next step is then to derive the Markov transition probabilities among the linear models given in (2), after the restriction of (2) to the regions Ω_i is removed. The model parameters are obtained by micro-partitioning the space R^n into cells whose measure is commensurate with unity relative to the covariance Q . The transition probabilities for a contracting region can then be approximated by considering the number of cells in an expanding region that forms the closure of the mapping of the original contracting region under f . Similarly, the transition probabilities for an expanding region can be obtained via the number of cells that are contained in the contracting region forming the mapping of the expanding region under the transformation f when the noise effects are included.

The resulting model is similar to the scalar case except that the transitions are no longer restricted to neighboring regions only as discussed in [2]. The assumptions (C) and (E) ensure that the transition matrix is relatively sparse. One problem which needs to be further resolved is the case when the matrix A_i is contracting in one direction but expanding in another. Preliminary investigation of this aspect is being considered.

The next phase involves the application of the model to nonlinear filtering problems. Here again, a departure from the scalar case occurs due to the dependence of the transition probabilities on the value of x . The state of the system is determined by x , its covariance P under the appropriate linear model, and the probabilities p_i of the i th linear model. The time update in the filtering problem solution is based on updating P corresponding to the model in effect for the given state, and then updating the probability vector p based on the ellipsoid defined by P , the state estimate, and the allowable regions for transitions. The overall estimate update is obtained as a weighted sum of the allowable transition regions resulting from the current region and P . The measurement update follows similarly by updating the estimate based on the possible models, while the update of the probability vector p is more ad hoc. The update of p is based on a weighted sum of the steady state probability obtained from the original Markov model, and the possible transitions determined from the estimate of x and its covariance under the i th model. Additional study of the resulting filter and its properties is needed.

There are two additional major problems to be considered. The first is involved in relaxing some of the restrictions that limit the types of the nonlinearities which can be used. The second is involved with the identification problem when the nonlinearity is unknown. In both cases, the dimensionality problem can be resolved as in the scalar case utilizing the sparseness of the transition matrix.

ACKNOWLEDGEMENT

This research is supported by the US Air Force Armament Laboratory, under Contract F08635-84-C-0273.

REFERENCES

- [1] A. H. Haddad and E. I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. Ann. Conf. on Inform. Sci. and Systems, pp. 202-206, March 1984, Princeton University.
- [2] A. H. Haddad, "On the Modeling and Filtering for Piecewise Linear Stochastic Systems," Proc. 19th Ann. Conf. on Inform. Sci. and Systems, pp. 44-49, March 1985, Johns Hopkins University.

APPENDIX D

ON THE MODELING AND FILTERING FOR PIECEWISE LINEAR
STOCHASTIC SYSTEMS

ON THE MODELING AND FILTERING FOR PIECEWISE LINEAR STOCHASTIC SYSTEMS*

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

ABSTRACT

This paper is concerned with the problem of nonlinear filtering for piecewise linear uncertain dynamic systems. The system is approximated by a set of linear models with Markovian transitions. The resulting multi-model nonlinear filtering scheme is implemented by allowing only slow or fast transitions among the models, based on the stability or instability of the linear models involved. This implementation reduces the exponential increase in the complexity of such systems. The scheme does not assume exact a priori knowledge of the transition probabilities involved in the model.

1. INTRODUCTION

Nonlinear discrete time models, which usually model sampled versions of continuous time systems, are very common in many applications of nonlinear filtering schemes. Many ad hoc approximations have been proposed for such systems. Piecewise linear approximations for nonlinear systems is such an ad hoc procedure, which has a potential for being amenable to systematic analysis in terms of the quality of the approximation. The systematic nature of the approach stems from the fact that once the model is approximated the error can be arbitrarily controlled by using additional terms or higher order approximations. The general idea is based on approximating the piecewise linear model by a set of linear models with Markov transitions among the models. An application of the switched Markov model to particular systems can be found in [1]. In other nonlinear filtering problems it has been primarily applied to systems with time varying and unknown parameters, as discussed in the surveys found in [2-4]. The basis of this paper is the one-dimensional case discussed in [5], in which the Markov linear approximation is analyzed. The required assumptions on the quality of the approximations are utilized to reduce the complexity of the resulting nonlinear filtering problem. For simplicity, the system will still be restricted to the scalar case, even though direct extension to the vector case is possible.

We are concerned with the following system model

$$x_{k+1} = g(x_k) + w_k, \quad k = 0, 1, 2, \dots \quad (1)$$

where the nonlinearity $g(x)$ is assumed to have the form

$$g(x) = \beta_i + \frac{\beta_{i+1} - \beta_i}{\alpha_{i+1} - \alpha_i} (x - \alpha_i), \quad \alpha_i \leq x < \alpha_{i+1}$$

$$\triangleq a_i x + b_i, \quad i = 1, 2, \dots, M \quad (2)$$

and where the $\{w_k\}$ is a white Gaussian noise sequence with variance Q . It is desired to solve the nonlinear filtering problem of estimating the state of the system while the precise values of the cutoff points in the nonlinear regions are not precisely known. One may also assume that the piecewise linear models are not precisely known, and that the expressions given in (2) are merely a quantized approximation to the unknown linear levels. The observations are assumed to be linear of the form

$$y_k = c x_k + v_k, \quad k = 0, 1, 2, \dots \quad (3)$$

where the noise sequence $\{v_k\}$ is also white Gaussian with variance R . The problem can be further generalized to allow piecewise linear observation model, in which case the Markov linear approximation can also be used for the observations as well. However, this will not be done here to conserve space.

The objective, therefore, is to obtain a nonlinear filtering scheme which does not require explicit identification of the nonlinearity. The approach is based on approximating the nonlinearity by a linear switched parameter Markov model [5]. One could also apply the same approximation to the observation process. The result will be similar to the uncertain observations case discussed, for example, in [6-8].

The next section discusses the Markov switched approximation as derived in [5], with particular emphasis on the restrictions imposed on the nonlinearity that allows the approximation to be valid. Next the general filter structure is summarized, and the need for assumptions to reduce its exponential complexity growth is discussed. The paper then discusses the approach used and the restrictions imposed on the nonlinear model that allow the reduction in complexity by using only fast and slow transitions. The paper concludes with discussion of possible extensions and generalizations.

*This work is supported by the U.S. Air Force Armament Laboratory under Contract F08635-84-C-0273.

2. LINEAR MARKOV APPROXIMATION

For simplicity we restrict the discussion to the scalar case as was done in [5], where the linear Markov approximation was investigated. Here we summarize the results in order to provide the basis for the approximations and assumptions used in the sequel. The linear Markov approximation is based on defining macro states S_i , ($i = 1, \dots, M$), such that if the system is in state S_i , then it follows the following linear model

$$x_{k+1} = a_i x_k + b_i + w_k, \quad x_k \in S_i. \quad (4)$$

If the state S_i represents the region $\{a_i < x_k < a_{i+1}\}$, then the transition probabilities among the macro states are, of course, dependent on the value of x_k . The basic approximation is involved in assuming that the sequence of the macro states form a Markov chain whose transition probability matrix has elements

$$\pi_{ij} = P(x_{k+1} \in S_j \mid x_k \in S_i) \quad (5)$$

which are not dependent on x_k , and that the model (4) is valid for all values of x_k .

Such an approximation would not be an improvement were it not for the special cases that led to simplified models. In particular, three cases can be discerned as was discussed in [5].

1. The first case occurs when in a given state the original system contains a relatively large stable region. In this case, the transition probability to neighboring regions is very small, and hence one can assume that some form of a steady state may be reached while in that state. Furthermore, when constructing all the possible transitions in a given time interval, one may assume a relatively small number of transitions.
2. The second case occurs when the original system contains a relatively small unstable region, having a high gain and bounded by two stable regions. In this case it has been shown that the probability of transition to other states is very high. Hence, one may assume that the system leaves such a state in only a few sampling instants.
3. The third case considered occurs when a stable region is relatively small and has two other neighboring stable regions. In this case one can write a series expansion for the system properties, and it can be shown that asymptotically, the small region can be absorbed by the larger neighboring regions for modeling purposes.

The only case which is not included in the three cases given above is the case of a large unstable region. However, such a case may cause other problems to the system, and thus we shall

restrict our considerations to a system model displaying these three cases of behavior.

In general, these restrictions should not pose any difficulties, as several adjacent linear stable regions may be combined and approximated by one large stable linear region. Consequently, the general form of the approximating model involves a series of large linear stable regions, with relatively small unstable linear regions separating them. It should be noted that the unstable regions require a relatively large slope for the approximation to be valid. However, if the slope is not large enough, again the asymptotic series expansion applies. In such a case, the region may also be absorbed by the neighboring stable regions.

The overall approximating model for this type of piecewise linear nonlinearity becomes a switched Markov model, with transitions between alternating stable and unstable linear systems. The approximate transition probability is very large for the unstable states, and very small for the stable states. If one also assumes that transitions beyond the immediate neighbors can be neglected (due to continuity assumptions on the original nonlinear system, and the relative sizes of the regions), then one obtains the following form for the transition probability matrix

$$\begin{aligned} \pi_{ii} &= \begin{cases} 1 - c_i & , i = 1, 3, \dots, M \\ c_i & , i = 2, 4, \dots, M-1 \end{cases} \\ \pi_{i,i+1} &= \pi_{i,i-1} = \begin{cases} \frac{1}{2} c_i & , i = 3, 5, \dots, M-2 \\ \frac{1}{2} - \frac{1}{2} c_i & , i = 2, 4, \dots, M-1 \end{cases} \\ \pi_{1,2} &= c_1, \quad \pi_{M,M-1} = c_M \end{aligned} \quad (6)$$

where the c_i 's denote small positive parameters, and where all the other π_{ij} are assumed to be negligible, or identically zero. The number of models, M , is of course odd, so that the overall system is represented by a stable model. The odd macro states represent the large regions, while the even ones represent the small unstable regions of the original nonlinearity. This simplified structure will result in significant reduction in the complexity of the nonlinear filtering scheme which is based on the switched Markov linear model.

3. GENERAL FILTER STRUCTURE

The optimal filter for the switched Markov linear model has been derived by many authors as can be found in the surveys [2-4]. It consists in general of a weighted sum of multiple Kalman filters in parallel and may be written as

$$\hat{x}(k) = \sum_{j=1}^{M_k} \hat{x}_j(k) P(I_j(k) | Y_k) \quad (7)$$

where

$$\hat{x}_j(k) = E(x_k | Y_k, I_j(k)) \quad (8)$$

$$Y_k \triangleq \{y_0, y_1, \dots, y_k\} \quad (9)$$

and where $I_j(k)$ is a specific sequence of macro states $\{S(i)\}$ of the Markov chain during the observation interval $\{i = 0, 1, \dots, k\}$. Thus

$$I_j(k) = \{j_0, j_1, \dots, j_k\}, \quad 1 \leq j_1 \leq M \quad (10)$$

where in the above sequence $S(i) = S_{j_i} =$ the state of the system at time i , and hence $M_k = M^{k+1}$. In the sequel we shall denote S_{j_i} simply by j_i for simplicity.

Here $\hat{x}_j(k)$ is implemented as a linear Kalman filter matched to a given sequence of transitions in the Markov model, and $P(I_j(k) | Y_k)$ is the generalized likelihood functional of that sequence:

$$P(I_j(k) | Y_k) = \frac{P(Y_k | I_j(k)) P(I_j(k))}{\sum_{j=1}^{M_k} P(Y_k | I_j(k)) P(I_j(k))} \quad (11)$$

where $P(Y_k | I_j(k))$ is Gaussian, due to the linearity, for a given sequence $I_j(k)$, and this is easily obtained from the Kalman filters $\hat{x}_j(k)$ and their covariances. The $P(I_j(k))$ are directly given from the transition probabilities (6).

Such a solution is, of course, optimal and depends on the entire past record. However, it is impractical due to the exponential increase in the number of filters required as the observation interval increases. Many approaches have been derived in the past to avoid the increase in complexity, including random sampling, Gaussian approximations, and other criteria for truncating the number of filters [3,7]. In this paper we utilize the properties of the model as described in the previous section to reduce the complexity of the filter. These properties follow from the fast and slow nature of the Markov chain transitions given in (6). In this case, the approximation is based on the property of the piecewise linear nonlinearity and its switched Markov linear model.

An initial reduction in the number of filters M_k is directly obtained from the tridiagonal structure of the transition probability matrix (6). In this case the total number of filters is approximately

$$M_k \approx 3^k M \quad (12)$$

However, this result does not take into account the end effects (states S_1 and S_M) or the relative size of transition probabilities.

The simplified scheme involves two sets of filters, a slow set of filters representing the slowly switching stable linear regions, and a fast set that represents the fast switching unstable regions. The length of observation interval used is also truncated to be compatible with the time constants of the slowest linear submodel. In this manner, for each filter in the slow set, it may start with any of the stable linear models, but may include at most one switching in the interval, to one of two neighboring unstable linear submodels. Similarly, each fast filter for the unstable mode may only remain a small finite number of samples in the fast state and then it switches to one of two neighboring stable linear models.

We now obtain an estimate of the number of the required filters if the above assumptions are also used. If $K/2$ is the number of samples required for the stable submodels to reach steady state, and if the unstable submodels are assumed to remain in the unstable mode at most L samples, then the number of required filters is approximately $2KL(M-1)$. This expression will be clarified when the filter structure is described in the next subsection. This approximation is derived under the assumption that no more than one switching occurs in the interval of K samples for the stable states. Consequently, for $L=1$, namely, the system switches immediately from a fast state after only one sample, then only $2K(M-1)$ filters are required. However, even this linearly increasing number may be too large. In that case, alternative approaches to the detection of a transition can be employed as in failure detection schemes, or the detection of incident processes.

3.1 The Slow Filters Structure

The slow filters are implemented by using one of the $(M+1)/2$ stable linear models. At the starting phase, the filters used are those corresponding to the assumed initial conditions. As a matter of fact, if the initial conditions are known, then only one slow filter needs to be used at the starting phase, and then it may switch to other modes at future time instants. However, if the initial condition is also assumed unknown, then the partitioning rule applies to the initial condition, so that at the start each of the slow filters assumes an initial condition equal to the steady state mean value of the model for that region.

The basic assumption used in the implementation is that during an observation interval of length K time samples, each slow filter may experience at most one jump at some arbitrary instant i . If a jump occurs, the filter switches into one of two adjacent fast macro states. However, since it is assumed that the system

escapes the fast states after one sample (this assumption can be relaxed, by allowing more than one sample in the fast states) to switch to another adjacent slow state, the resulting slow mode filters may be divided into three different basic types:

The first involves the filters with no jumps during the observation interval, and these are denoted by

$$\hat{x}_{i,0}(k) = E\{x_k | Y_k, I_{i,0}(k)\}, \quad i = 1, 3, \dots, M \quad (13)$$

where the sequence $I_{i,0}(k)$ is defined by the following states

$$I_{i,0}(k) = \{S(j) = i, 0 \leq j \leq k\} \quad (13a)$$

The second involves the filters experiencing a jump at the l -th time sample, $1 \leq l \leq k-2$, and the system returns to a different slow mode after the transition to the fast mode, and these are denoted by

$$\begin{aligned} \hat{x}_{i,1\pm 2}(k|l) &= E\{x_k | Y_k, I_{i,1\pm 2}(k|l)\}, \\ i &= 1, 3, \dots, M \end{aligned} \quad (14)$$

where the sequence $I_{i,1\pm 2}(k|l)$ is defined by the following states

$$\begin{aligned} I_{i,1\pm 2}(k|l) &= \{S(j) = i, 0 \leq j < l, S(l) = i \pm 1, \\ &\quad S(j) = i \pm 2, l < j \leq k\} \\ i &= 1, 3, \dots, M. \end{aligned} \quad (14a)$$

The third involves the filters which return to the original slow mode state after a jump to an adjacent fast state, and these are denoted by

$$\hat{x}_{i,1\pm}(k|l) = E\{x_k | Y_k, I_{i,1\pm}(k|l)\}, \quad i = 1, 3, \dots, M \quad (15)$$

where the sequence $I_{i,1\pm}(k|l)$ is defined by the following states

$$\begin{aligned} I_{i,1\pm}(k|l) &= \{S(j) = i, j \neq l, S(l) = i \pm 1\}, \\ i &= 1, 3, \dots, M. \end{aligned} \quad (15a)$$

It is seen that this last type may be obtained in two distinct ways depending on the adjacent fast states. In all the above cases, for the end states $i = 1, M$ the number of filters need to include only the allowed transitions.

The number of filters required to implement the scheme over the observation interval K , may be obtained from all the distinct possibilities defined in (13)-(15), and by including the assumption that the system switches back from the fast states after sample $l + j$, $1 \leq j \leq L$. If we start with initial conditions in only one region, then (13) yields only one filter, while (14) and (15) provide each $2K(\frac{M-1}{2}) \cdot L$ filters, where only one-sided jumps are allowed in states $i = 1, M$. Here the factor 2 stems from the two possible

transitions in each of (14) and (15), the $(\frac{M-1}{2})$ is the total number of slow states less one (to account for the end states), and the K and L represent the possible samples where jumps can occur. Hence, the approximate number of filters is given by the expression $2KL(M-1)$ as mentioned above.

The basic structure, therefore, involves a set of slow filters which are interrupted at one time sample l by a jump to a fast mode filter, where l may vary over the observation interval. The actual implementation of these filters is obtained via standard Kalman filters, with a change in initial conditions and gains after the transitions corresponding to the parameters of the new state. Furthermore, the fast filters may be replaced by an initial condition change (as obtained from the applications of singular perturbation theory), so that as a first-order approximation there may be no need to implement the fast filters.

The problem is how to prevent the growth, albeit linear growth, in the number of filters for $k > K$. It can be easily shown that the slow stable filters are stable and approach steady state faster than the state of the original system. We shall assume that the maximum time required to approach the steady state for the stable filters is $K/2$. Consequently, under the above assumptions, for $k = l + K/2$, the slow filter in (13)-(15) which correspond to the same final state, namely with $S(k) = i$, will be aggregated into a single filter. The filters are combined, and their likelihood functionals are added together. The process after the aggregation continues as at the initial state, so that the number of required filters remains stable.

3.2 The Fast Filter Structure

The fast filters occur only during the transition from one slow state to another slow state. With the assumptions that it does not stay in a given state more than one sample, the resulting filter only involves at most a single time varying stage in the transition between two slow filters. However, if a continuous time system is to be implemented, or a faster sampling rate is available, then the fast filters may be implemented at a fast stretched time scale. Namely, when the fast filters are implemented, a faster sampling rate may be used.

If only one sample is assumed in the fast mode filters, the fast filters simply provide the transitional mode between two slow states. If the transition occurs at the l -th time sample, the estimate corresponding to the fast mode may be denoted by

$$\begin{aligned} \hat{x}_{i,1\pm 1}(l+1|l) &= E\{x_{l+1} | Y_l, I_{i,1\pm 1}(l)\}, \\ i &= 1, 3, \dots, M \end{aligned} \quad (16)$$

where the sequence $I_{i,i\pm 1}(l)$ is used to denote the following state transitions

$$I_{i,i\pm 1}(l) = \{S(j) = i, 0 \leq j < l, S(l) = i \pm 1\} \quad (16a)$$

It therefore can be implemented by a Kalman filter with initial conditions $x_{i,i\pm 1}(l)$ with corresponding covariances, and gains and covariance equation obtained from the parameters of the fast state $S(l) = i \pm 1$. The new estimate obtained in (16) and its covariance, serves as initial conditions for the continued processing of the slow filters (14)-(15).

The actual estimate (16) is used in the weighted sum implementation of the overall filter only when the transition is assumed to occur at $l = k - 1$, where k is the current observation sample.

3.3 The Overall Filtering Scheme

The overall scheme involves the evaluation of the likelihood functionals (11) that are used in the weighted sum approximation (7). Since the probability of transition is assumed to be either large or small, it can be assumed that all the resulting filters are a priori equally likely. If in addition a higher order approximation is required for the fast modes by allowing jumps after more than one sample, then a second order approximation is obtained by allowing a jump after two samples with probability ϵ if the original jump probability is ϵ . Of course, this requires the use of double the number of filters. Similarly, if we need to allow more than one jump in an interval of length K , then the number of filters need to be squared, unless additional approximations are used.

The overall first order filter approximation is therefore given by a weighted sum of all the filters defined in (13)-(15) in addition to the fast filters states for $l = k$ as given in (16). The likelihood functionals are obtained by the standard expressions involving the correlation of the individual filters outputs with the residuals.

The identification aspect is involved in not precisely knowing the values of ϵ_i 's which represent the transition or nontransition probability of the model. Since the model is obtained from the system only through the piecewise linear approximation, it is not possible or unlikely to be possible in general to have precise values for the ϵ_i 's. These parameters need only be assumed as small but unknowns. If all the values are assumed to be equal, then they do appear in the expressions for the weighted sum filter in a way that they can be canceled. However, if this assumption is not valid, then alternatives, such as those discussed in [8], may have to be used at the expense of increased filter complexity.

The filter needs to be analyzed in terms of its ability to identify the state of the original piecewise linear system. Furthermore, its performance needs to be investigated via simulation. Preliminary simulation of the original system indicates that the model is quite good in so far as the autocorrelation and the density function of the Markov linear model as compared to the piecewise nonlinear system. One possible modeling assumption may be to shift the regions of validity of the piecewise linear approximations so that the transition probabilities may be equalized.

4. SUMMARY AND CONCLUSIONS

The scheme discussed here is only a preliminary one, with many of its properties yet to be investigated. In particular, the initial scalar simulation needs to be extended both to higher dimensions and to systems with more than the three linear regions. A hierarchy of ordering for the vector case needs to be developed so that the assumptions of tridiagonal transition probability matrix may still apply. Finally, applications to specific nonlinear systems remain to be carried out so as to demonstrate the utility of the approach. The approach can be made systematic in that only the error in the Markov model should be controlled while the error in the identification and filtering would follow from the modeling error. Implementation issues involved with the numerical problems using a large number of filters are investigated by Verriest (e.g. [9]).

5. REFERENCES

- [1] H.P. Van Landingham, R.L. Moose, and W.H. Lucas, "Modeling and Control of Nonlinear Plants," Proc. 17th IEEE Conference on Decision and Control, pp. 337-341, San Diego, CA, January 1979.
- [2] A.H. Haddad and J.K. Tugnait, "On State Estimation Using Detection-Estimation Schemes for Uncertain Systems," Proc. JACC, pp. 514-519, Denver, CO, June 1979.
- [3] J.K. Tugnait, "Detection and Estimation for Abruptly Changing Systems," *Automatica*, Vol. 18, pp. 607-615, September 1982.
- [4] A.H. Haddad, "On Detection-Estimation Schemes for Uncertain Systems," in *Communications and Networks: A Survey of Recent Advances*, I.F. Blake and H.V. Poor, Eds., Springer-Verlag, New York, 1985.
- [5] A.H. Haddad and E.I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. 1984 Conference on Information Sciences and Systems, pp. 202-206, Princeton University, March 1984.

- [6] M.T. Hadidi and S.C. Schwartz, "Linear Recursive State Estimators Under Uncertain Observations," IEEE Trans. on Automatic Control, Vol. AC-24, pp. 944-948, 1979.
- [7] J.K. Tugnait and A.H. Haddad, "A Detection-Estimation Scheme for State Estimation in Switching Environment," Automatica, Vol. 15, pp. 477-481, 1979.
- [8] J.K. Tugnait, "On Identification and Adaptive Estimation for Systems with Interrupted Observations," Automatica, Vol. 19, pp. 61-73, 1983.
- [9] E.I. Verriest, "Error Analysis of Linear Recursion in Floating Point," Proc. ICASSP, Tampa, FL, March 1985.

APPENDIX E

APPROXIMATE NONLINEAR FILTERS FOR PIECEWISE LINEAR MODELS

APPROXIMATE NONLINEAR FILTERS FOR PIECEWISE LINEAR MODELS

E. I. Verriest and A. M. Haddad
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

This paper is concerned with an approximate nonlinear filtering scheme for piecewise linear stochastic systems. The system is modeled by a switched Markovian transitions among linear models. The optimal estimator for the resulting system requires exponentially increasing number of filters in a combined detection estimation system. The approach proposed in this paper reduces the number of such filters by using a consistency test based on the original linear regions of the nonlinear system. In this way an improvement in the accuracy of the scheme using fixed number of filters can be obtained.

INTRODUCTION

The objective of this paper is to provide a suboptimal scheme for nonlinear filtering in systems modeled by piecewise linear nonlinearities. These systems have been approximated by multimodel linear systems with Markovian jumps among the models [1-2]. However, the resulting nonlinear models require a nonlinear filtering scheme that uses a detection estimation structure with a number of filters that is exponentially increasing [3]. An earlier approach [4] attempted to reduce the required number of filters in the scheme by utilizing the sparse nature of the Markovian transition probability matrix. However, the scheme based on the approximate model does not take into account the consistency constraints imposed by the original model of the nonlinear system. The scheme to be considered here attempts to overcome both of these limitations, by employing the constraints of the original model, and requiring a finite number (which can be selected to achieve desired accuracy) of filters for the suboptimal filter. The scheme is a modification of the finite Gaussian sum approximation used in [5], with the addition of the consistency condition imposed by the original piecewise linear model.

The system under consideration is given by the following discrete time model

$$x_{k+1} = g(x_k) + v_k \quad (1)$$

where the state of the system at time t_k is x_k , and where the noise sequence $\{v_k\}$ is assumed to be white and Gaussian. The observation model y is (for the time being) assumed to be linear with additive white Gaussian noise v

$$y_k = C x_k + v_k \quad (2)$$

The nonlinearity is assumed to have a continuous piecewise linear approximation given by the following model

$$g(x) = A_i x + b_i \quad (3)$$

for $x \in Q_i$, $i = 1, \dots, M$

where the regions $\{Q_i\}$ form a partition of the

entire space. The approximation discussed in [1-2] provides the basis for the nonlinear filter design. In this approximation the system (1) is assumed to have M different linear models as given by (3) where model i is valid when the state S (called macro-state) of an underlying Markov process is equal to S_i . The transition probability matrix Π of the Markov process is derived from the original system by considering the transitions from Q_i to Q_j . This may seem like an enormous computational effort, depending on the complexity of the nonlinearity. On the other hand this transition matrix can be precomputed. The steady state probabilities can be found in the same manner. The validity of the Markov approximation is based on some assumptions on the linear regions $\{Q_i\}$. Two types of regions are allowed for the approximation to be appropriate, depending on a measure, $u(\cdot)$, of region size defined relative to the covariance of the white process noise v . The first type is what is termed as a contracting region, satisfying the relation

$$u(g(Q_i)) < u(Q_i).$$

The second type is called an expanding region, satisfying the relation

$$u(g(Q_i)) > u(Q_i).$$

Here, the notation $g(Q)$ is used to refer to the image of Q under the mapping of the nonlinearity g . Furthermore, it is assumed that the measures of contracting regions are relatively large, while the measures of expanding regions are relatively small, in order to ensure the validity of the approximation.

The resulting model is a finite state Markov chain with macro-states $\{S_i\}$, having transition probability matrix $\{\Pi_{ij}\}$, and steady state probabilities $\{p_i\}$, where the Π_{ij} are very small for expanding regions. When the macro-state is S_i the system obeys a linear model (3). The optimal filter for such a model [3] involves a set of Kalman filters matched to all possible sequences of macro-states, and followed by a weighted sum using the generalized likelihood function of each sequence. This filter involves an exponentially increasing number of filters. An earlier approach [4] to reduce this number to polynomial growth utilized the sparseness of the transition probability matrix, Π , and the relative size of the transitions to the different types of regions. In this paper, an alternative approach is used that allows a fixed number of filters, and this number may be expanded depending on the need for accuracy. The approach is a modification of the Gaussian sum approximation in [5] that utilizes the structure of the original nonlinear model.

GENERAL FILTERING SCHEME

The scheme is assumed to have M possible filters (such a choice can be generalized to a number of filters M^k , for arbitrary k), yielding at time k , a set of M estimates $\hat{x}_i(k)$, corresponding

covariances $P_i(k)$, and estimated probabilities of the macro-states $\hat{p}_i(k)$. Hence, at each stage the total information state update involves the incorporation of the measurement $y(k+1)$ with the prior information state

$$I(k) \equiv (\hat{x}_i(k), P_i(k), \hat{p}_i(k); i = 1, 2, \dots, M)$$

to the new information state $I(k+1)$. At each stage, the first step is to combine the estimates and their covariances by a weighted sum to arrive at a single estimate $\hat{x}(k)$ and a single covariance $P(k)$ using the macro-state probabilities $\hat{p}_i(k)$. This is the estimate that is the output of the scheme at stage k . Then a consistency test is made to ensure that the estimate $\hat{x}(k)$ is consistent with the macro-state probabilities $\hat{p}_i(k)$. This consistency test involves an adjustment of the $\hat{p}_i(k)$ to conform the $\hat{x}(k)$ and its covariance $P(k)$ to the region Q_i . The consistency update generates M macro-state probabilities $\hat{p}_i(k)$ to be used in propagating the information state. In order to update the information for the next time instant, the single estimate $\hat{x}(k)$ is used together with the transition probabilities Π and the M models (3) to obtain the information state $I(k+1|k)$ prior to the next measurement. These estimates are then updated by incorporating the measurement $y(k+1)$ via the usual linear Kalman filter matched to the model governed under macro-state S_i , while likelihood functions are used to obtain the measurement update of the a posteriori macro-state probabilities. In a more general setting, more information can be carried out from one stage to the next, so that a sequence of two states can be used to update the filters, if M^2 filters are selected.

Consistency Update

Since this step is the major difference between this approach and earlier ones, it will be described first. If the variance of the estimate is small, then the information provided by the estimates of $\hat{p}_i(k)$ can be neglected. In this case, these values are changed based on the position of the estimate $\hat{x}(k)$ in the appropriate region Q_i to generate the a priori macro-state probabilities $\hat{p}_i(k)$ to be used for the transition to the next stage for the updating of $\hat{p}_i(k+1|k)$. If the covariance P is large, then the macro-state information is relied on more heavily in determining the macro-state probabilities. One ad hoc way to accomplish this is to use the following weighted update expression

$$\hat{p}_i(k) = \alpha(P) \hat{p}_i(k) + [1 - \alpha(P)] U_i\{\hat{x}(k)\}. \quad (4)$$

Here $\alpha(P)$ is a function of the norm of the covariance of the estimate, that tends to zero as the covariance becomes small, and tends to 1 as the covariance becomes large. The $U_i(x)$ is an indicator function of the region Q_i that represents the macro-state S_i , i.e. it is equal to unity if $x \in Q_i$, and zero otherwise.

Time Update of Estimates

We shall address first the question of time updating the macro-state probabilities $\hat{p}_i(k+1|k)$. These are updated by using the consistency updated values $\hat{p}_i(k)$ together with the transition probabilities. In this case we have

$$\hat{p}(k+1|k) = \Pi \hat{p}(k) \quad (5)$$

where $\hat{p}(k)$ denotes the vector of $\hat{p}_i(k)$'s. The updates of the estimates $\hat{x}_i(k+1|k)$ and their covariances $P_i(k+1|k)$ are obtained from the combined estimate at stage k and the models described by (3), to yield

$$\hat{x}_i(k+1|k) = A_i \hat{x}(k) + b_i \quad (6)$$

$$P_i(k+1|k) = A_i P(k) A_i' + Q. \quad (7)$$

where Q is the process noise covariance. This approach assumes in essence that the distribution of the state x satisfies a Gaussian sum approximation. This implies that the update is obtained by using M Kalman filters matched to the linear models described in (3), and with initial value at k given by the combined estimate $\hat{x}(k)$ and its covariance $P(k)$. Again, in the case of M^2 filters, for example, we have M combined estimates at time k , and each can be subject to a transition based on one of the M macro-states, to yield M^2 estimates. These in turn will be combined again to provide another set of M estimates for propagation to the next stage. Some of these transitions may not be possible due to the structure of the transition probability matrix. In such a case, the number M^2 serves only as an upper bound on the number of filters used. These updated estimates will not be combined until after the measurement updates that are used on each of the individual estimates corresponding to each macro-state.

Measurement Update of Estimates

The estimates after the measurement $y(k+1)$ is available are derived using the models in (3) to yield the standard Kalman filter formulation

$$\hat{x}_i(k+1) = \hat{x}_i(k+1|k) +$$

$$P_i(k+1|k) C' R^{-1} (y(k+1) - C \hat{x}_i(k+1|k)) \quad (8)$$

$$P_i(k+1) = ([P_i(k+1|k)]^{-1} + C' R^{-1} C)^{-1}. \quad (9)$$

The question is now concerned with the measurement update of the macro-state probability estimates. This can be accomplished by using the standard likelihood function for a switched Markov model. It should be noted that such an update is only valid for the true switched Markov model, while it is only an approximation in this case. The expression for the a posteriori probabilities in this case will be proportional to the likelihood function, given by

$$\hat{p}_i(k+1) = \beta \hat{p}_i(k+1|k) A_i(k+1) \quad (10)$$

$$A_i(k) = \exp(-\frac{1}{2} [y(k) - C \hat{x}_i(k)]' R^{-1} [y(k) - C \hat{x}_i(k)]) \quad (11)$$

where β is a normalization coefficient.

The consistency update used earlier to provide the a priori information for the transition probabilities is expected to compensate for the fact that a smaller number of filters is used than warranted by the optimal estimate for the switched Markov approximation. The fact that these macro-states originate in a physical region is used to correct the estimate of the likelihood function representing the a posteriori probabilities of the macro-states.

Combined Estimate

The combined estimate $\hat{x}(k)$ is obtained by using the likelihood weighted probabilities of the macro-states as a weighted sum of the individual estimates as dictated by the optimal scheme for the switched Markov model

$$\hat{x}(k) = \sum \hat{p}_i(k) \hat{x}_i(k). \quad (12)$$

The covariance for the combined estimate can be obtained in a similar fashion by assuming a Gaussian sum approximation, to yield the expression

$$P(k) = \sum \hat{p}_i(k) \{P_i(k) + \hat{x}_i(k)\hat{x}_i'(k)\} - \hat{x}(k)\hat{x}'(k) \quad (13)$$

where the validity of the approximation depends on the validity of the switched Markov model.

The overall updating steps involved in the scheme are illustrated in Figure 1.

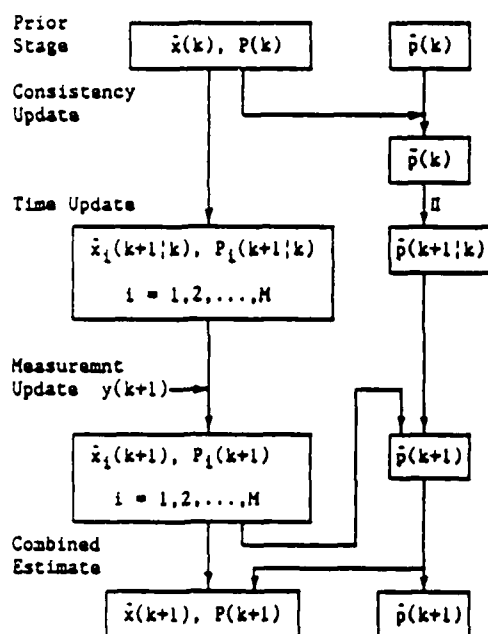


Figure 1. Block Diagram of the Filter Stages

ANALYSIS OF THE FILTER

The complexity of the filter precludes analytical derivation of its performance. One has to rely on simulation and other asymptotic techniques to address the question of performance and convergence. Several observations can be made relative to the behavior of the filter. The filter performance would largely depend on the accuracy of the switched Markov approximation for the piecewise linear system. Hence, the filter is expected to perform well when the process noise covariance is large relative to the expanding regions of the nonlinearity, and small relative to the contracting regions of the nonlinearity. The approximation is such that it can be improved by increasing the fixed number of filters used in the scheme. It is thus possible to improve the performance by taking

more stages of memory in the scheme. Finally, the scheme should perform better than a purely switched Markov approximation even when the approximation itself is not too good, due to the involvement of the consistency updating that relies on the exact model of the system. The consistency updating is, at present, based on an ad hoc formulation. There is room for improvement in selecting an optimal choice for the weighting function $\alpha(P)$. In the next section a scalar case is simulated in order to illustrate the behavior of the filtering scheme.

SCALAR CASE

A special case which is also used for a numerical example to demonstrate some of the properties of the filter is considered here. A scalar system, which is parameterized by four parameters and has three regions is used. The system and observation model are given by

$$\begin{aligned} a_1 x_k + a_0 - a_1 + b u_k, & \text{ for } x_k > 1 \\ x_{k+1} = a_0 x_k + b u_k, & \text{ for } -1 < x_k < 1 \quad (14) \\ a_1 x_k - a_0 + a_1 + b u_k, & \text{ for } x_k < -1 \end{aligned}$$

$$y_k = c x_k + v_k \quad (15)$$

where the noise sequences u and v are white Gaussian with zero means and unit variances. The analysis in this paper is restricted to the case $a_0 > 1$ and $-1 < a_1 < 1$, that yields a stable system with two contracting and one expanding regions. The deterministic system has two stable equilibrium points at $\pm x^*$

$$x^* = (a_0 - a_1)/(1 - a_1) > a_1. \quad (16)$$

Two cases are considered, the first involves the case of small (relatively) process noise, namely, $b \ll (x^* - 1)$. In this case the probability of transitions from the contracting regions is very small, and the steady state probability density function of x_k may be approximated by a Gaussian sum of two densities with means at $\pm x^*$ and variance

$$b^2/(1 - a_1^2).$$

In this case the estimation problem becomes basically a problem in detection. However, the resulting model satisfies the assumptions that render the switched Markov model a valid one for the system. The second case involves the one with $b \gg x^*$, in which case we can rewrite the system equation as

$$x_{k+1} = a_1 x_k + b \left\{ u_k + \frac{a_0 - a_1}{b} \phi(x_k) \right\} \quad (17)$$

where $\phi(x)$ is a nonlinearity with a limiter characteristic. Due to the assumption on the magnitude of b , the additive term to the noise is negligible and the system behaves essentially as a linear system. The range of interest should therefore lie between the two cases discussed above, even though the Markov approximation is better for the first case, the behavior of the system allows simpler approaches.

The symmetry of the problem allows the derivation of the transition matrix of the macro-states that involves only three terms. These can be either derived directly, or in cases of unknown

noise parameters, we may assume values that are compatible with high transition probabilities from the expanding region, and low transition probabilities from the contracting regions. If we use subscripts of +, -, and 0 to denote the three regions, we need only derive the transition probabilities for P_{++} , P_{+-} , P_{0+} . The remaining probabilities are obtained by normalization, and symmetry. The state of the system involves the a posteriori probabilities of being in one of the three macro-states at observation time k , and the estimate of the state and its covariance given any particular sequence of states. The approximation in deriving the filter removes the dependence on an entire sequence, and relies on only a finite number of steps. In order to compensate for the loss of information, the probabilities of being in a given macro-state are updated using the consistency updates described in the previous section. The filter will involve three estimates, with their corresponding covariances and the three macro-states probabilities, which are used to obtain the combined weighted estimate of the system.

Simulation Results

The filter was simulated for a range of values of the parameters b , c , a_0 , and a_1 . The sample variance of the combined filter (C) was compared to the extended Kalman filter (E) and to the linear equivalent filter (L). The results for the values

$$a_0 = 10, a_1 = -0.2,$$

and several values of b and c were obtained for 1000 samples and are shown in Table 1.

$b = 3$			
c	0.05	0.1	1.0
L	79	71	1.9
E	169	152	0.864
C	20	10	0.865

$b = 5$			
c	0.05	0.1	1.0
L	86	67	1.09
E	145	101	0.932
C	77	34	0.929

$b = 10$			
c	0.05	0.1	1.0
L	118	66	0.962
E	145	65.8	0.953
C	121	58.4	0.954

Table 1. The Sample Variance of the Filters

The results indicate that, as expected, the new scheme performs best when the process noise variance, which is determined by b , is high relative to the size of the unstable region, but small relative to the stable regions. The effect of the measurement noise also shows that its performance gets worse relative to the other two filters as the measurement noise is decreased, i.e. c is increased. The results tend to confirm several of the assumptions made when the switched Markov approximation was considered. It is expected that the performance can be improved if additional filters are used, and the sparsity of the transition probability matrix is utilized in determining the viable macro-states as was done in [4], in addition to the consistency updates.

SUMMARY AND CONCLUSIONS

This paper considered a suboptimal filtering scheme for the nonlinear estimation problem in systems with piecewise linear models. The approximations used is based on utilizing the switched Markov model for the system, as well as modifying the resulting filter with the physical constraints of the states of the model. Additional improvements are possible, by incorporating some features that reflect the fact that the transition probability matrix has special characteristics involving fast and slow transitions. Additional properties such as convergence and optimal choice of the consistency updating function need further investigations. Applications to nonlinear tracking and guidance problems are also contemplated.

ACKNOWLEDGEMENT

This research is supported by the U. S. Air Force Armament Laboratory, under Contract F08635-84-C-0273.

REFERENCES

- [1] A. H. Haddad and E. I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 202-206, Princeton University, March 1984.
- [2] A. H. Haddad and E. I. Verriest, "Piecewise Linear Modeling of Multidimensional Stochastic Nonlinear Systems," Proc. Annual Allerton Conference on Communications, Control, and Computing, pp. 777-778, University of Illinois, October, 1985.
- [3] J. K. Tugnait, "Detection and Estimation for Abruptly Changing Systems," *Automatica*, Vol. 18, pp. 607-615, September 1982.
- [4] A. H. Haddad, "On the Modeling and Filtering for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 44-49, Johns Hopkins University, March 1985.
- [5] G. A. Ackerson and K. S. Fu, "On State Estimation in Switching Environment," *IEEE Trans. on Automatic Control*, Vol. AC-15, pp. 10-17, February 1970.

APPENDIX F

APPROXIMATE NONLINEAR FILTERING FOR PIECEWISE LINEAR SYSTEMS

APPROXIMATE NONLINEAR FILTERING FOR PIECEWISE LINEAR SYSTEMS

Abraham H. Haddad, Erik I. Verriest, and Philip D. West
 School of Electrical Engineering
 Georgia Institute of Technology
 Atlanta, GA 30332-0250
 USA

Presented at 44th Symposium of GCP, NATO-AGARD, May 4-8, 1987

SUMMARY

This paper is concerned with an approximate nonlinear filtering scheme for piecewise linear stochastic systems. Such nonlinearities may be useful in modeling the geometric nonlinearities in an air-to-air tracking and guidance problem. The nonlinearity is assumed to be present both in the system model and the observation model. The system is modeled by a switched Markovian transitions among linear models. The optimal estimator for the resulting system requires exponentially increasing number of filters in a combined detection estimation scheme. The approach proposed in this paper reduces the number of such filters by using a consistency test based on the original linear regions of the nonlinear system. In this way an improvement in the accuracy of the scheme using fixed number of filters can be obtained. An illustrative example demonstrate the improvements provided by the scheme.

INTRODUCTION

In an air-to-air engagement scenario, the trajectories of the missile and target are usually nonlinear in nature. Furthermore, the observations used to track the target are typically nonlinear due to the geometry even if the target is flying a steady rectilinear trajectory. This paper is concerned with the problem of tracking the target in such a nonlinear system from noisy nonlinear observations. The solution for such a nonlinear filtering scheme is not implementable exactly even if one ignores the uncertainties in the models of the system. These systems have been approximated by nonlinearities with piecewise linear characteristics. The objective of this paper is to provide a suboptimal scheme for the nonlinear filtering of the states of the systems when these systems are modeled by such piecewise linear nonlinearities. In earlier papers [1-2] these systems have been approximated by multimodel linear subsystems with Markovian jumps among these models. However, the resulting nonlinear models require a nonlinear filtering scheme that uses a detection estimation structure with a number of filters that is exponentially increasing [3]. An earlier approach [4] attempted to reduce the required number of filters in the scheme by utilizing the sparse nature of the Markovian transition probability matrix. However, the scheme based on the approximate model does not take into account the consistency constraints imposed by the original model of the nonlinear system. A scheme that imposes a consistency condition on the resulting filters was shown to be effective in improving the filter performance without increasing the number of filters [5]. The scheme overcame both of these limitations, by employing the constraints of the original model, and requiring a finite number (which can be selected to achieve desired accuracy) of filters for the suboptimal filter. The scheme is a modification of the finite Gaussian sum approximation used in [6], with the addition of the consistency condition imposed by the original piecewise linear model. This paper investigates the scheme further, by generalizing it to the case of nonlinearities in the observation process. It also considers the case of allowing the memory of the suboptimal filter to be larger by keeping a number of filters over several observation samples.

The system under consideration is given by the following discrete time model

$$x_{k+1} = g(x_k) + b w_k \quad (1)$$

where x_k is the n -dimensional state vector of the system at time k , and where the noise sequence (w_k) is assumed to be white and Gaussian with covariance matrix Q . The observation model y_k (m -dimensional vector) is assumed to be nonlinear with additive white Gaussian noise v_k with covariance R

$$y_k = h(x_k) + v_k. \quad (2)$$

The nonlinearities $g(x)$ and $h(x)$ are assumed to have a continuous piecewise linear approximation given by the following model

$$g(x) = G_i x + g_i \quad (3)$$

$$\text{for } x \in Q_{g_i}, i = 1, \dots, M_g$$

$$h(x) = H_i x + h_i \quad (4)$$

$$\text{for } x \in Q_{h_i}, i = 1, \dots, M_h$$

where G_i and H_i are constant matrices, g_i and h_i are constant vectors, and where the regions $\{Q_{g_i}\}$ and $\{Q_{h_i}\}$ form partitions of the entire space. The two partitions may be combined for simplicity of notation such that a new partition containing all nonempty intersections of sets Q_{g_i} and Q_{h_j} for all i and j to provide a partition with sets Q_i , $i = 1, 2, \dots, M$ where M is no larger than $M_g M_h$.

The approximation for (1) discussed in [1-2] provides the basis for the nonlinear filter design. In this approximation the system (1) is assumed to have M_g different linear models as given by (3) where model i is valid when the state $S(k)$ (called macro-state) of an underlying Markov process is equal to S_i where i ranges from 1 to M_g . The transition probability matrix Π of the Markov process is derived from

the original system by considering the transitions from Q_{i1} to Q_{ij} . This may seem like an enormous computational effort, depending on the complexity of the nonlinearity. On the other hand this transition matrix can be precomputed. The steady state probabilities can be found in the same manner. The same assumption may also be made about the observation nonlinearity $h(x)$. Hence the approximation is assumed to hold for the overall combined sets $\{Q_i\}$. The Markov linear approximation yields the following representation for the system

$$x_{k+1} = G_i x_k + G_i + b w_k \quad (5)$$

$$y_k = H_i x_k + h_i + v_k \quad (6)$$

when $S(k) = S_i$, $i = 1, 2, \dots, M$.

where we have combined the two models for g and h and their regions, and where the macro-state S_i corresponds to the regions $x \in Q_i$. The approximation involves two assumptions: The first is that the transition probabilities from state S_i at time k to state S_j at time $k+1$ are not dependent on the actual value of the state x_k . The second involves the approximation that the models given by (5)-(6) are not restricted by the constraints of the nonlinearities.

In general the transition probability from S_i to S_j may be obtained from the following expression

$$\Pi_{ij} = \text{Prob.} \{ x_{k+1} \in Q_j \mid x_k \in Q_i \} \quad (7)$$

where Π is the transition probability matrix for the Markov chain defined by S . The Markov chain also has a corresponding steady-state marginal probabilities p_i of macro-state S_i obtained from

$$p = p \Pi \quad (8)$$

where p is a row vector with components p_i . The validity of the Markov approximation is based on some assumptions on the linear regions $\{Q_i\}$ relative to the noise variances. Two types of regions are allowed for the approximation to be appropriate, depending on a measure, $\mu(\cdot)$, of region size defined relative to the covariance of the white noise processes. The first type is what is termed as a contracting region, satisfying the relation

$$\mu(q(Q_i)) < \mu(Q_i). \quad (9)$$

The second type is called an expanding region, satisfying the relation

$$\mu(q(Q_i)) > \mu(Q_i). \quad (10)$$

Here, q stands for the joint function defined by the intersection of g and h , and the notation $q(Q)$ is used to refer to the image of Q under the mapping of the nonlinearity q . Furthermore, it is assumed that the measures of contracting regions are relatively large, while the measures of expanding regions are relatively small, in order to ensure the validity of the approximation.

The resulting model is a finite state Markov chain with macro-states $\{S_i\}$, having transition probability matrix $\{\Pi_{ij}\}$, and steady state probabilities $\{p_i\}$, where the Π_{ij} are very small for expanding regions and relatively large for contracting regions. When the macro-state $S(k)$ is equal to S_i the system obeys a linear state and observation model (5)-(6).

The optimal filter for such a model (also called switched parameter model) [3] involves a set of Kalman filters matched to all possible sequences of macro-states, and followed by a weighted sum using the generalized likelihood function of each sequence. This filter involves an exponentially increasing number of filters. An earlier approach [4] to reduce this number to polynomial growth utilized the sparseness of the transition probability matrix, Π , and the relative size of the transitions to the different types of regions. In an earlier paper [5] an alternative approach was used that allows a fixed number of filters, and this number may be expanded depending on the need for accuracy. The approach is in some sense a modification of the Gaussian sum approximation in [6] but which also utilizes the structure of the original nonlinear model. The approach which was restricted to state nonlinearities is generalized in this paper to include the observation nonlinearity.

GENERAL FILTERING SCHEME

The scheme is assumed to have M^r possible filters, depending on the number, r , of levels of memory utilized by the scheme. This memory reflects the maximum number of possible sequences of transitions of the macro-states propagated by the scheme. Let $J(k)$ denote a particular possible sequence of macro-state transitions involving r samples ending at time instant k . In other words there are M^r possible such sequences given by

$$\{J(k)\} = \{j_{k-r+1}, \dots, j_{k-1}, j_k\} \quad (11)$$

where

$$\{j_i = 1, 2, \dots, M; i = k-r+1, \dots, k-1, k\}$$

and where each index j_i represent the value of the macro-state at time i within the particular sequence. We also denote by $J(k;i)$ a particular sequence $J(k)$ that ends in macro-state i at time k , i.e.,

$$J(k;i) = \{J(k-1), j_k = i\} \quad (12)$$

In general at time k , the scheme yields a set of M^r estimates $\hat{x}_{J(k)}(k)$, corresponding covariances $P_{J(k)}(k)$, and estimated probabilities of the macro-state sequences obtained as normalized likelihood functions $\Lambda_{J(k)}(k)$ reflecting the a posteriori probability estimate of the sequence of r transitions of the

macro-states corresponding to the sequence of integers defined by $J(k)$. Hence, at each stage the total information state update involves the incorporation of the measurement $y(k+1)$ with the prior information state

$$I(k) = \{\bar{x}_{J(k)}(k), P_{J(k)}(k), \Lambda_{J(k)}(k)\} \quad (13)$$

to the new information state $I(k+1)$. At each stage, the first step is to combine the estimates and their covariances by a weighted sum to arrive at a single estimate $\bar{x}(k)$ and a single covariance $P(k)$ using the macro-state sequence a posteriori probabilities. This is the estimate that is the output of the scheme at stage k and is given in general by

$$\bar{x}(k) = \sum_{J(k)} \bar{x}_{J(k)}(k) \Lambda_{J(k)}(k). \quad (14)$$

where the summation is over all M^F possible sequences $J(k)$. In order to update the information state it is convenient to define the estimates and likelihood functions that correspond to sequences of the form $J(k;i)$. There are M^{F-1} such sequences ending in macro-state i . These estimates and their covariance will be denoted as above with the subscript $J(k;i)$ instead of $J(k)$. In this case we can determine the a posteriori probability of the macro-state at time k equal to i by $\hat{p}_i(k)$ and expressed as

$$\hat{p}_i(k) = \sum_{J(k-1)} \Lambda_{J(k;i)}(k). \quad (15)$$

Similarly we can define the conditional estimate at time k based on the sequence $J(k-1)$ as a weighted sum of the corresponding estimates to $J(k;i)$ after averaging over all possible current macro-states. The resulting estimate is given by

$$\bar{x}_{J(k-1)}(k) = \left\{ \sum_i \bar{x}_{J(k;i)}(k) \Lambda_{J(k;i)}(k) \right\} / \Lambda_{J(k-1)}(k) \quad (16)$$

where

$$\Lambda_{J(k-1)}(k) = \sum_i \Lambda_{J(k;i)}(k) \quad (17)$$

The conditional probability of the current macro-state given the sequence $J(k-1)$ may also be derived in a similar manner

$$\hat{p}_i(k|J(k-1)) = \Lambda_{J(k;i)}(k) / \Lambda_{J(k-1)}(k). \quad (18)$$

This estimate provides another representation for the overall estimate

$$\bar{x}(k) = \sum_{J(k-1)} \bar{x}_{J(k-1)}(k) \Lambda_{J(k-1)}(k) \quad (19)$$

The rationale for such a representation is that it provides an additional way to check the consistency of the estimate by updating the conditional probability of the current estimate of the macro-state by using the estimate of the estimate of the state. This consistency test is made to ensure that the estimate $\bar{x}_{J(k-1)}(k)$ that may be used to propagate to the next stage is consistent with the macro-state probabilities $\hat{p}_i(k|J(k-1))$. This consistency test involves an adjustment of the macro-state probabilities to conform to the state estimate and its conditional covariance $P_{J(k-1)}(k)$ to the region Q_i . The consistency update generates M macro-state conditional probabilities $\hat{p}_i(k|J(k-1))$ to be used in propagating the information state to the next stage. In order to update the information for the next time instant, the M^F estimates need to be aggregated by averaging over the earliest time instant and then updating the filters by using the remaining estimates together with the transition probabilities Π and the M models (5)-(6) to obtain the information state $I(k+1|k)$ prior to the next measurement. These estimates are then updated by incorporating the measurement $y(k+1)$ (corresponding to the appropriate model of the macro-state) via the usual linear Kalman filter matched to the model governed under macro-state S_i , while likelihood functions are used to obtain the measurement update of the a posteriori macro-state probabilities.

This approach can still be combined with other approaches for reducing the filter complexity and the number of filters required. These include the use of the sparseness of the transition probability matrix of the macro-states and the relatively small or large probability of transitions for certain states [4]. Other approaches [7] involve: The aggregation of estimates that are approximately equal in terms of mean and variance. The elimination of sequences that are unlikely based on a posteriori probabilities. The combining of filters whose distance measures are smaller than a certain value. These approaches are expected to further enhance the utility of the proposed approach.

CONSISTENCY UPDATE

Since this step is the major difference between this approach and earlier ones, it will be described first. If the variance of the estimate is small, then the information provided by the estimates of $\hat{p}_i(k|J(k-1))$ can be neglected. In this case, these values are changed based on the position of the estimate $\bar{x}_{J(k-1)}(k)$ in the appropriate region Q_i , to update the a posteriori macro-state probabilities $\hat{p}_i(k|J(k-1))$ to be used for the transition to the next stage for the updating of $\hat{p}_i(k+1|k)$. If the covariance $P_{J(k-1)}(k)$ is large, then the macro-state information is relied on more heavily in determining the macro-state probabilities. One ad hoc way to accomplish this is to use the following weighted update expression

$$\hat{p}_i(k|J(k-1)) = \alpha(P_{J(k-1)}) \hat{p}_i(k|J(k-1)) + \{1 - \alpha(P_{J(k-1)})\} U_i(\bar{x}_{J(k-1)}(k)) \quad (20)$$

where the dependence of P on k is omitted for ease of presentation. Here $\alpha(P)$ is a function of the norm of the covariance of the estimate, that tends to zero as the covariance becomes small, and tends to 1 as the covariance becomes large. The $U_i(x)$ is an indicator function of the region Ω_i that represents the macro-state S_i , i.e. it is equal to unity if $x \in \Omega_i$, and zero otherwise.

TIME UPDATE OF ESTIMATES

We shall address first the question of time updating the macro-state probabilities $\tilde{p}_i(k+1|k)$. These are updated by using the consistency updated values $\tilde{p}_i(k)$ together with the transition probabilities. In this case the conditional probabilities will be propagated (before obtaining the next measurement) by multiplying by the transition probabilities of the macro-states as given via the matrix Π . Explicitly, this can be written as

$$\tilde{\Lambda}_{J(k+1;1)}(k+1|k) = \tilde{p}_j(k|J(k-1)) \Lambda_{J(k-1)}(k) \Pi_{ji} \quad (21)$$

where the left side of the equation is the probability that at time $k+1$ the macro-state ends with state i preceded by the particular sequence $J(k)$, conditional on the data up to and including time k . The updates of the estimates $\tilde{x}_{J(k+1)}(k+1|k)$ and their covariances $P_{J(k+1)}(k+1|k)$ are obtained from the estimates at stage k and the models described by (5)-(6), to yield

$$\tilde{x}_{J(k+1)}(k+1|k) = G_i \tilde{x}_{J(k;1)}(k) + g_i \quad (22)$$

$$P_{J(k+1)}(k+1|k) = G_i P_{J(k;1)}(k) G_i' + b Q b' \quad (23)$$

where a prime is used to denote transposition. This approach assumes in essence that the distribution of the state x satisfies a Gaussian sum approximation. This implies that the update is obtained by using M Kalman filters matched to the linear models described in (5)-(6), and with initial value at k given by the estimates matched to each possible preceding macro-state sequence and its covariance. These in turn will be aggregated again to reduce the total number of filters to M^r for propagation to the next stage. Some of these transitions may not be possible due to the structure of the transition probability matrix. In such a case, the number M^r serves only as an upper bound on the number of filters used. These updated estimates will not be combined until after the measurement updates that are used on each of the individual estimates corresponding to each macro-state. Again, in the case of M filters, i.e., $r = 1$, we have a single combined estimates at time k , and it is propagated based on the transition to any one of the M macro-states, to yield M estimates.

MEASUREMENT UPDATE OF ESTIMATES

The estimates after the measurement $y(k+1)$ is available are derived using the models in (5)-(6) to yield the standard Kalman filter formulation

$$\tilde{x}_{J(k+1;1)}(k+1) = \tilde{x}_{J(k+1;1)}(k+1|k) + P_{J(k+1;1)}(k+1|k) H_i' R^{-1} v_i(k+1) \quad (24)$$

$$P_{J(k+1;1)}(k+1) = \{ [P_{J(k+1;1)}(k+1|k)]^{-1} + H_i' R^{-1} H_i \}^{-1} \quad (25)$$

where the $v_i(k+1)$ is the innovations process based on the macro-state S_i at time $k+1$ defined by

$$v_i(k+1) = (y(k+1) - H_i' \tilde{x}_{J(k+1;1)}(k+1|k) - h_i) \quad (26)$$

The question is now concerned with the measurement update of the macro-state probability estimates. This can be accomplished by using the standard likelihood function for a switched Markov model. It should be noted that such an update is only valid for the true switched Markov model, and it is only an approximation in this case. The expression for the a posteriori probabilities in this case will be proportional to the likelihood functions, $\Lambda_{J(k;1)}(k)$, which for simplicity are defined to include the normalizing constant. The update equations are given in this case by the expression

$$\Lambda_{J(k+1;1)}(k+1) = \beta \tilde{\Lambda}_{J(k+1;1)}(k+1|k) \exp(-\frac{1}{2} v_i'(k+1) R^{-1} v_i(k+1)) \quad (27)$$

where β is a normalization coefficient.

The consistency update used earlier to provide the a priori information for the transition probabilities is expected to compensate for the fact that a smaller number of filters is used than warranted by the optimal estimate for the switched Markov approximation. The fact that these macro-states originate in a physical region is used to correct the estimate of the likelihood function representing the a posteriori probabilities of the macro-states.

COMBINED ESTIMATE

The combined estimate $\hat{x}(k)$ is obtained by using the likelihood weighted probabilities of the macro-states as a weighted sum of the individual estimates as dictated by the optimal scheme for the switched Markov model

$$\hat{x}(k) = \sum_{J(k)} \Lambda_{J(k)}(k) \tilde{x}_{J(k)}(k) \quad (28)$$

The covariance for the combined estimate can be obtained in a similar fashion by assuming a Gaussian sum approximation, to yield the expression

$$P(k) = \sum_{J(k)} \Lambda_{J(k)}(k) \{P_{J(k)}(k) + \tilde{x}_{J(k)}(k) \tilde{x}_{J(k)}'(k)\} - \tilde{x}(k) \tilde{x}'(k) \quad (29)$$

where the validity of the approximation depends on the validity of the switched Markov model. Again, these equations only show the composite estimate. The estimates and their covariances that are matched to a specific macro-state are obtained in exactly the same expression with a summation over the subscript j of $J(k)$ in a similar manner to (15) for the estimates.

The overall updating steps involved in the scheme are illustrated in Figure 1. Similar but simpler algorithms can be drawn for the special cases of $M = 1$, or $M = 2$.

ANALYSIS OF THE FILTER

The complexity of the filter precludes analytical derivation of its performance. One has to rely on simulation and other asymptotic techniques to address the question of performance and convergence. Several observations can be made relative to the behavior of the filter. The filter performance would largely depend on the accuracy of the switched Markov approximation for the piecewise linear system. Hence, the filter is expected to perform well when the process noise covariance is large relative to the expanding regions of the nonlinearity, and small relative to the contracting regions of the nonlinearity. The approximation is such that it can be improved by increasing the fixed number of filters used in the scheme. It is thus possible to improve the performance by taking more stages of memory in the scheme. Finally, the scheme should perform better than a purely switched Markov approximation even when the approximation itself is not too good, due to the involvement of the consistency updating that relies on the exact model of the system. The consistency updating is, at present, based on an ad hoc formulation. There is room for improvement in selecting an optimal choice for the weighting function $\alpha(P)$. In the next section a scalar case is simulated in order to illustrate the behavior of the filtering scheme.

SCALAR CASE

A special case which is also used for a numerical example to demonstrate some of the properties of the filter is considered here. A scalar system, in which the $g(x)$ and $h(x)$ functions have three regions each symmetric (odd symmetry) about the origin is used for demonstration. The nonlinearities are shown in Figure 2 and are seen to be parameterized by five parameters. The system and observation model are given by

$$x_{k+1} = \begin{cases} g_1 x_k + g_0 - g_1 + b w_k, & \text{for } x_k > 1 \\ g_0 x_k + b w_k, & \text{for } -1 < x_k < 1 \\ g_1 x_k - g_0 + g_1 + b w_k, & \text{for } x_k < -1 \end{cases} \quad (30)$$

$$y_k = \begin{cases} h_1 x_k + h_0 - h_1 + v_k, & \text{for } x_k > \alpha \\ h_0 x_k + v_k, & \text{for } -\alpha < x_k < \alpha \\ h_1 x_k - h_0 + h_1 + v_k, & \text{for } x_k < -\alpha \end{cases} \quad (31)$$

where the noise sequences w and v are white Gaussian with zero means and unit variances. The analysis in this paper is restricted to the case $g_0 > 1$ and $-1 < g_1 < 1$, that yields a stable system with two contracting and one expanding regions. The deterministic system has two stable equilibrium points at tx^*

$$x^* = (g_0 - g_1)/(1 - g_1) > g_1. \quad (32)$$

Two cases are considered, the first involves the case of small (relatively) process noise, namely, $b \ll (x^* - 1)$. In this case the probability of transitions from the contracting regions is very small, and the steady state probability density function of x_k may be approximated by a Gaussian sum of two densities with means at tx^* and variance

$$b^2/(1 - g_1^2).$$

In this case the estimation problem becomes basically a problem in detection. However, the resulting model satisfies the assumptions that render the switched Markov model a valid one for the system. The second case involves the one with $b \gg x^*$, in which case we can rewrite the system equation as

$$x_{k+1} = g_1 x_k + b \left\{ w_k + \frac{g_0 - g_1}{b} \phi(x_k) \right\} \quad (33)$$

where $\phi(x)$ is a nonlinearity with a limiter characteristic. Due to the assumption on the magnitude of b , the additive term to the noise is negligible and the system behaves essentially as a linear system. The range of interest should therefore lie between the two cases discussed above, even though the Markov approximation is better for the first case, the behavior of the system allows simpler approaches.

The symmetry of the problem allows the derivation of the transition matrix of the macro-states that involves only five states because of the nonlinearity in the observations. These can be either derived directly, or in cases of unknown noise parameters, we may assume values that are compatible with high transition probabilities from the expanding region, and low transition probabilities from the contracting

regions. For the $g(x)$ we use subscripts of +, -, and 0 to denote the three regions, we need only derive the transition probabilities for Π_{++} , Π_{+-} , Π_{0+} . The remaining probabilities are obtained by normalization, and symmetry. If we include the $h(x)$ we obtain in general five macro-states and their transitions can be derived in a similar fashion. In the simulation the state transition matrix is derived experimentally using about 1000 sample steps. Even though the result is not as accurate as analytical derivation, especially for low probability states, the filter performance was robust relative to the values of the apriori transition probabilities. The state of the system involves the a posteriori probabilities of being in one of the five macro-states at observation time k , and the estimate of the state and its covariance given any particular sequence of states. The approximation in deriving the filter removes the dependence on an entire sequence, and relies on only a finite number of steps. In order to compensate for the loss of information, the probabilities of being in a given macro-state are updated using the consistency updates described in the previous section. The filter will involve five estimates, with their corresponding covariances and the five macro-states probabilities, which are used to obtain the combined weighted estimate of the system.

SIMULATION RESULTS

The filter was simulated for a range of values of the parameters b , h_1 , h_0 , α , while $g_0 = 10$, and $g_- = -0.2$ were held constant. The sample mean and variance of the error of the combined filter (C) were compared to the extended Kalman filter (EKF) for 1000 time steps. Table 1 shows the values of the parameters used in the simulation, as well as the simulation results. The results are also shown in Figures 3 - 6.

The results indicate that, as expected, the new scheme performs best when the process noise variance, which is determined by b , is high relative to the size of the unstable region, but small relative to the stable regions. For the range of values selected the filter always performs better than the Extended Kalman Filter. It would perform worse if the observation is linear and the assumptions of the switched Markov models is not satisfied for the process. The performance improvement is striking when the observation nonlinearity is ambiguous, namely, it has regions with negative slopes. The EKF in this case cannot track the change in the region while the combined filter is able to detect the proper region when the consistency update is used, thus substantially reducing the uncertainty of what the true macro-state is supposed to be. It is expected that for such a simple scalar problem not much improvement can be expected from increasing the filter memory. However, the objective is to consider a multivariable nonlinear system to test the validity of the filter when more than one memory level is used.

SUMMARY AND CONCLUSIONS

This paper considered a suboptimal filtering scheme for the nonlinear estimation problem in systems with piecewise linear models in both the system and observations. The approximations used are based on utilizing the switched Markov model for the system, as well as on modifying the resulting filter with the physical constraints of the states of the model. Additional improvements are possible, by incorporating some features that reflect the fact that the transition probability matrix has special characteristics involving fast and slow transitions. Additional properties such as convergence and optimal choice of the consistency updating function need further investigations. Applications to nonlinear tracking and guidance problems are the motivation for this problem as most such scenarios involve highly nonlinear geometry with a great deal of uncertainty.

REFERENCES

- [1] A. H. Haddad and E. I. Verriest, "Linear Markov Approximations for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 202-206, Princeton University, March 1984.
- [2] A. H. Haddad and E. I. Verriest, "Piece-wise Linear Modeling of Multidimensional Stochastic Nonlinear Systems," Proc. Annual Allerton Conference on Communications, Control, and Computing, pp. 777-778, University of Illinois, October, 1985.
- [3] J. K. Tugnait, "Detection and Estimation for Abruptly Changing Systems," *Automatica*, Vol. 18, pp. 607-615, September 1982.
- [4] A. H. Haddad, "On the Modeling and Filtering for Piecewise Linear Stochastic Systems," Proc. Annual Conference on Information Sciences and Systems, pp. 44-49, Johns Hopkins University, March 1985.
- [5] 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.
- [6] G. A. Ackerson and K. S. Fu, "On State Estimation in Switching Environment," *IEEE Trans. on Automatic Control*, Vol. AC-15, pp. 10-17, February 1970.
- [7] J. K. Tugnait and A. H. Haddad, "A Detection-Estimation Approach to State Estimation in Switching Environment," *Automatica*, Vol. 15, pp. 477-481, July 1979.

ACKNOWLEDGEMENT

This research is supported by the U. S. Air Force Armament Laboratory, under Contract P08635-84-C-0273.

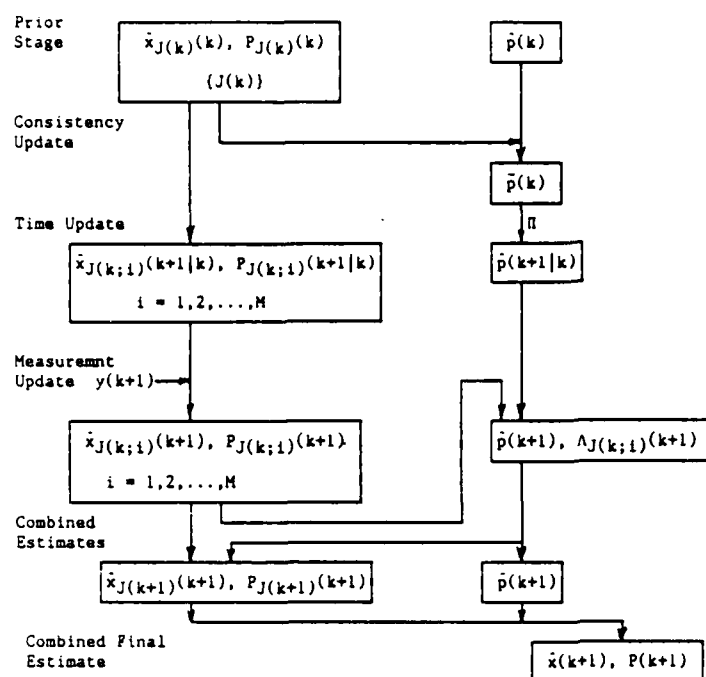


Figure 1. Block Diagram of the Filter Stages

Table 1. Sample Variance and Means for the Combined Filter and the EKF

H_0	H_1	b	a	Combined Filter		EKF	
				Variance	Mean	Variance	Mean
5	-0.1	3	0.5	8.141	-0.035	949.1	19.7
5	-0.1	4	0.5	15.71	-0.024	676.5	34.0
5	-0.1	5	0.5	18.46	-0.216	1106.5	-24.7
5	-0.1	7.5	0.5	40.43	-0.309	1165.3	27.15
5	-0.1	10	0.5	112.5	-1.34	1231.9	-27
5	0.1	3	0.5	8.89	0.015	9.54	0.485
5	0.1	4	0.5	12.9	0.468	33.51	-0.222
5	0.1	5	0.5	15.56	0.434	26.4	0.628
5	0.1	7.5	0.5	23.91	1.31	36.27	0.519
5	0.1	10	0.5	25.4	1.42	67.2	0.463
1	-0.1	3	2	8.67	-0.022	670.0	17.9
1	-0.1	4	2	16.09	-0.611	1036.0	-26.53
1	-0.1	5	2	30.88	-1.42	759.6	21.6
1	-0.1	7.5	2	77.65	-2.59	856.7	-22.85
1	-0.1	10	2	142.3	-0.735	753.5	-20.2
1	0.1	3	2	8.32	-0.067	12.46	0.507
1	0.1	4	2	12.77	0.68	19.23	1.65
1	0.1	5	2	14.5	0.083	29.0	0.337
1	0.1	7.5	2	19.76	1.27	25.51	1.43
1	0.1	10	2	24.4	0.55	35.6	0.36

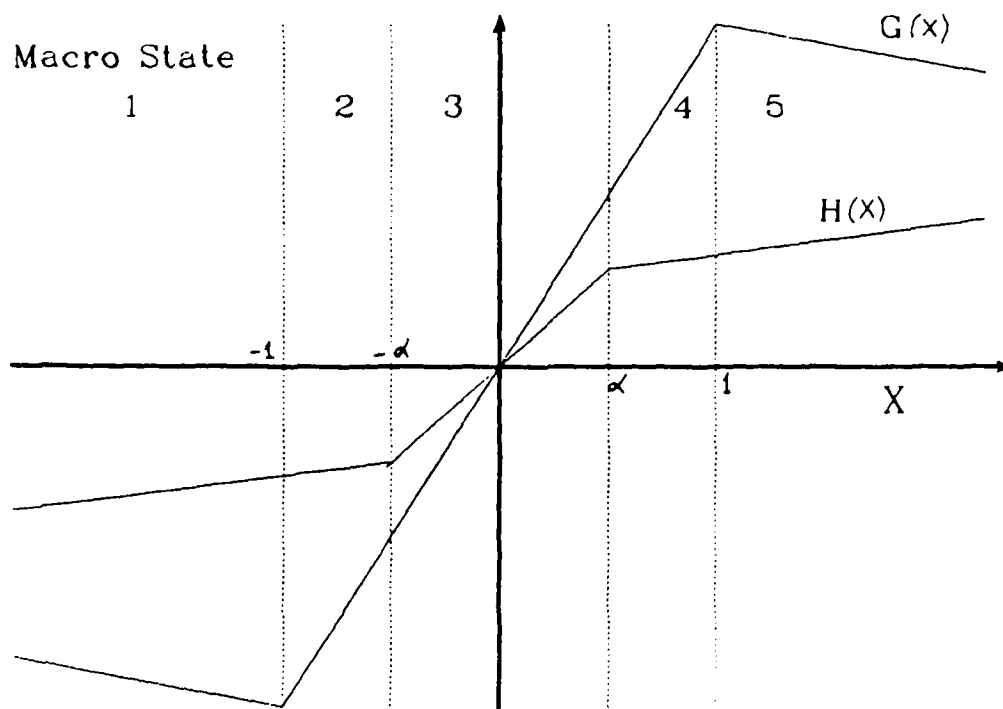
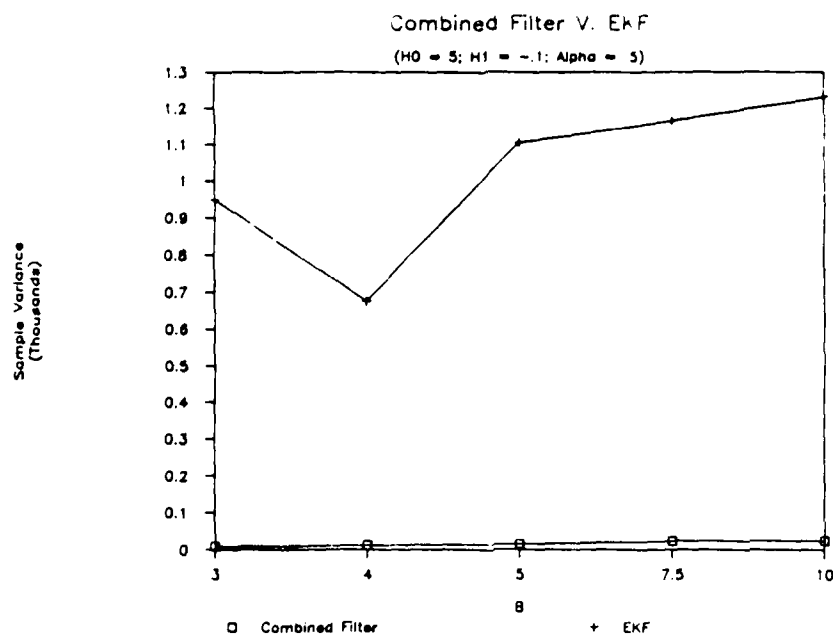


Figure 2. System and Measurement Model for the Scalar Case

Figure 3. Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = .5$, $H_0 = 5$, $H_1 = -0.1$

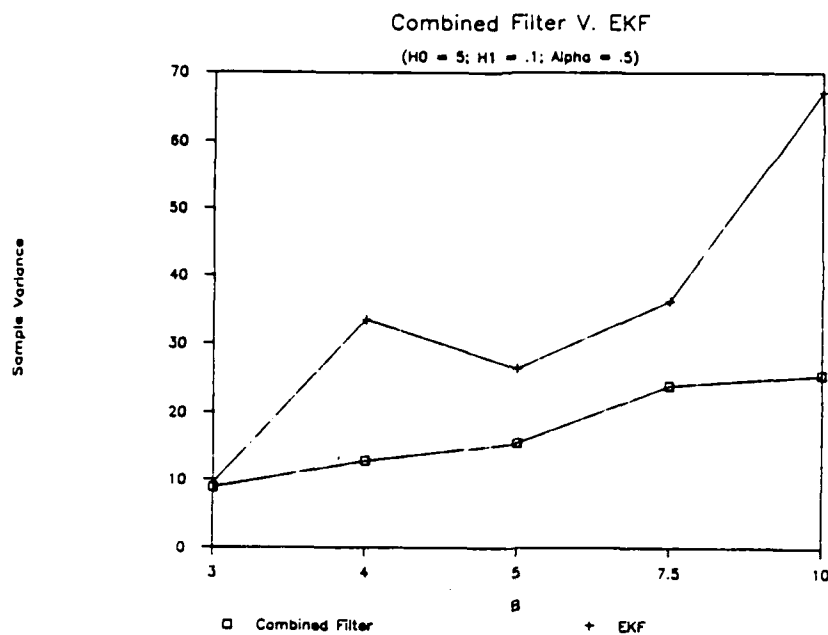


Figure 4. Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = .5$, $H_0 = 5$, $H_1 = 0.1$

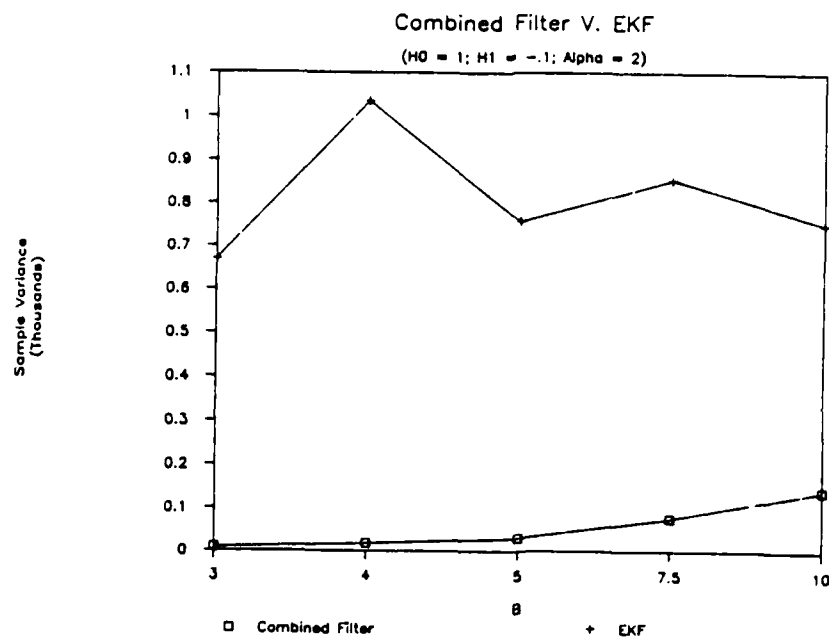


Figure 5. Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = 2$, $H_0 = 1$, $H_1 = -0.1$

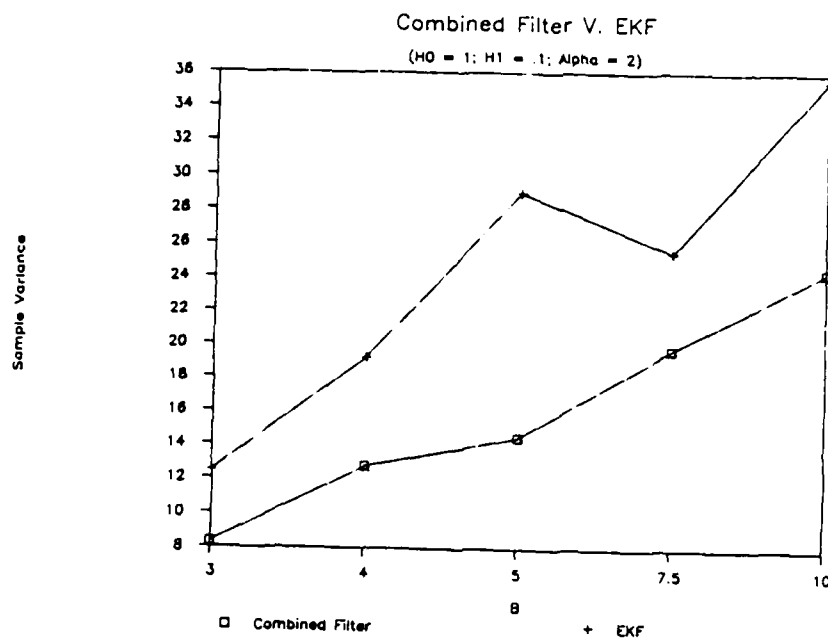


Figure 6. Comparison of the Combined Filter Error Variance and EKF Error Variance for $\alpha = 2$, $H_0 = 1$, $H_1 = 0.1$

APPENDIX G

LINEAR FILTERS FOR LINEAR SYSTEMS WITH MULTIPLICATIVE NOISE
AND NONLINEAR FILTERS FOR LINEAR SYSTEMS WITH NON-GAUSSIAN
ADDITIVE NOISE

LINEAR FILTERS FOR LINEAR SYSTEMS WITH MULTIPLICATIVE NOISE AND NONLINEAR FILTERS FOR LINEAR SYSTEMS WITH NON-GAUSSIAN ADDITIVE NOISE

E. I. Verriest

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
Phone: 894-2949

Abstract

Exact optimal least squares linear filters with precomputable gains are derived for the class of discrete linear systems with state update and output corrupted by white noise multiplying a linear function in the state.

The derived method is then applied to obtain suboptimal nonlinear filters for linear systems with non-Gaussian additive noise.

1. Background and Notation

In this paper an exact linear least squares filter for the general discrete system of the form

$$x_{k+1} = F_k x_k + G_k u_k + \sum_{i=1}^q w_k^{(i)} B_k^{(i)} x_k \quad (1)$$

$$y_k = H_k x_k + v_k + \sum_{i=1}^q w_k^{(i)} D_k^{(i)} x_k \quad (2)$$

is derived. The u_k , v_k and w_k are white second order noise processes with zero mean and covariance

$$\text{Cov}(u_k', v_k', w_k') = \begin{bmatrix} Q_k & \Gamma_k & \Lambda_k \\ \Gamma_k' & R_k & \Phi_k \\ \Lambda_k' & \Phi_k' & \Omega_k \end{bmatrix} \begin{matrix} p \\ p \\ q \end{matrix} \quad (3)$$

Their distribution is assumed to be symmetric about zero. The superscript (i) denotes the ith component of w_k . The random initial condition x_0 is assumed to be uncorrelated with u_k , v_k and w_k and has mean $\bar{x}$ and P .

Because (1) and (2) involve nonlinear operations, Gaussianity of the initial conditions and the noise samples will in general not be conserved in the state and output processes. Hence, we shall omit assumptions on Gaussianity since they will not simplify matters. For this reason the linear estimates obtained will not be the conditional expectations (conditioned on the observed data) and are thus not necessarily the exact least squares estimates. A special case of the model (1)-(3) arises for instance if the measurement devices have a fixed relative error (accuracy) (e.g. through the use of logarithmic sensors). In this case $q=1$ and

$D=H$. This situation also occurs if the equations (1) and (2) model quantities computed in floating point arithmetic [7]. In mechanical systems, the bilinear terms may have their origin in vibrational degradation. Another application area is the filtering of linear systems with additive non-Gaussian noise. This is explored in section 3 of this paper.

Earlier contributions in the area of non-Gaussian filtering are based on a Bayesian approach (Bucy [2]). Alspach and Sorenson [1] approximate the densities by a Gaussian sum. Truncation procedures have also been considered by Buxbaum and Maddad [3]. Other approximations exist, for example point masses, orthogonal functions, etc. [5]. In general these filters are computationally quite involved, and thus difficult to implement. Hence the need for computationally attractive as well as easy to understand and implement filters. Masreliez [6] suggested two "nice" filters which are restricted to either Gaussian state noise or Gaussian observation noise.

Remark that the entities B and D in (1) and (2) are in fact (1,2)-tensors, written here in terms of their component matrices. For notational convenience we shall use the Kronecker product notation [8]. (A tensor notation is too encumberant.) The last terms in (1) and (2) are respectively written as

$$B_k(w_k \otimes x_k) \text{ and } D_k(w_k \otimes x_k)$$

where

$$B_k = [B_k^{(1)} \dots B_k^{(q)}] \text{ and } D_k = [D_k^{(1)} \dots D_k^{(q)}]$$

In section 2 the linear filter for the above problem is derived. The usual form

$$\hat{x}_{k+1} = F_k \hat{x}_k + K_k (y_k - H_k \hat{x}_k) \quad (4)$$

is assumed, and the optimal gain scheduling (K_k) is obtained. The results of this section are then applied in Section 3 to yield suboptimal nonlinear filters for the linear non-Gaussian problem.

2. Design of the Optimal Linear Filter

Defining the estimation error $\tilde{x}$ as $x - \hat{x}$ one obtains from (1) and (4) the recursion,

$$\tilde{x}_{k+1} = F_k \tilde{x}_k + G_k u_k - K_k c_k + B_k(w_k \otimes x_k) \quad (5)$$

where c is the residual

$$\begin{aligned} e_k &= y_k - H_k \hat{x}_k \\ &= H_k \tilde{x}_k + v_k + D_k (w_k \otimes \tilde{x}_k) \end{aligned} \quad (6)$$

The equations (1), (5) and (6) can be combined to

$$\begin{aligned} \begin{bmatrix} \hat{x} \\ \tilde{x} \end{bmatrix}_{k+1} &= \begin{pmatrix} F & KH \\ 0 & F-KH \end{pmatrix}_k \begin{bmatrix} \hat{x} \\ \tilde{x} \end{bmatrix}_k \\ &+ \begin{pmatrix} 0 & K \\ G & -K \end{pmatrix}_k \begin{pmatrix} u \\ v \end{pmatrix}_k + \begin{pmatrix} KD \\ B-KD \end{pmatrix}_k (w_k \otimes \tilde{x}_k) \end{aligned} \quad (7)$$

The covariance equation for (7) follows by "squaring up" and taking expectations together with the fact that

$$E\tilde{x}(\tilde{x}', \tilde{x}') = 0$$

and using standard Kronecker product identities [8]. Denoting the components of the covariance as

$$\text{Cov} \begin{bmatrix} \hat{x} \\ \tilde{x} \end{bmatrix} = \begin{bmatrix} P & C \\ C' & Q \end{bmatrix}$$

the update equations are (suppressing the subscript k on the right-hand sides)

$$\begin{aligned} P_{k+1} &= (P-KH)P(F-KH)' + GQG' - K\Gamma'G' - G\Gamma K' + KRK' \\ &+ (B-KD)[\Omega \otimes (P+C+C'+I)](B-KD)' \\ &+ (B-KD)[(A'G'-\Phi'K') \otimes \tilde{x}] + [(G'-K\Phi) \otimes \tilde{x}'](B-KD)' \end{aligned} \quad (8)$$

$$\begin{aligned} C_{k+1} &= FCP' + KHPP' - FCH'K' - KHPPH'K' + K\Gamma'G' \\ &- KRK' + KD[\Omega \otimes (P+C+C'+I)](B-KD)' \\ &+ KD[(A'G'-\Phi'K') \otimes \tilde{x}] + (K\Phi \otimes \tilde{x}') (B-KD)' \end{aligned} \quad (9)$$

$$\begin{aligned} Q_{k+1} &= F\Gamma F' + KHC'P' + FCH'K' + KHPPH'K' + KRK' \\ &+ KD[\Omega \otimes (P+C+C'+I)]D'K' \\ &+ KD[(\Phi'K' \otimes \tilde{x}) + (K\Phi \otimes \tilde{x}')]D'K' \end{aligned} \quad (10)$$

Direct optimization of the gain sequence is possible by invoking the projection theorem. Indeed the optimization is optimal if and only if the resulting error is uncorrelated (= orthogonal) to the estimate. Hence setting the cross covariance C identical to zero in (9) yields a degenerate equation for the gain. In terms of the generalized innovations

$$\begin{aligned} R^c &= HPH' + R + D[\Omega \otimes (P+I)]D' \\ &+ D(\Phi' \otimes \tilde{x}) + (\Phi \otimes \tilde{x}')D' \end{aligned} \quad (11)$$

the optimal gain is

$$\begin{aligned} K^{opt} &= [FPH' + G\Gamma + B(\Omega \otimes (P+I))]D' \\ &+ (GA \otimes \tilde{x}')D' + B(\Phi' \otimes \tilde{x})R^{-c} \end{aligned} \quad (12)$$

Backsubstitution of this optimal gain in the update equations (8) and (10) yields

$$\Gamma_{k+1} = (F\Gamma F' + KR^c K')_k \quad (13)$$

$$\begin{aligned} P_{k+1} &= [FPF' + GQG' - KR^c K' \\ &+ B(\Omega \otimes (P+I))B' + B(A'G' \otimes \tilde{x}) + (GA \otimes \tilde{x}')B']_k \end{aligned} \quad (14)$$

The initial conditions for (14) and (15) are respectively

$$\Gamma_0 = \bar{x}_0 \bar{x}_0' \text{ and } P_0 = \bar{P} \quad (15)$$

The formulas (12) to (15) yield the most general gain update. The gain dependency on the estimated state $\tilde{x}$ disappears if Φ and Λ are both zero; i.e., if the multiplicative noise w is uncorrelated with the purely additive noises u and v . Note that the formulas obtained are equivalent to the Kalman filter formulas if the noise terms $Gu + B(w \otimes \tilde{x})$ and $v + D(w \otimes \tilde{x})$ are considered as equivalent additive noises $\bar{u}$ and $\bar{v}$.

3. Nonlinear Filter for Linear Systems with Non-Gaussian Noise

Consider again the model (1), (2) but now with B and D zero. It is assumed (without loss of generality) that the noises u and v , although non-Gaussian, have a symmetric distribution. (The asymmetries show up as biases in the odd moments.) The same assumption is made for the initial distribution $P(\tilde{x}_0 - \bar{x})$. It is also assumed that u and v are independent although this can be relaxed at the expense of a higher complexity.

The a priori estimate $\bar{x}$ satisfies a recursion $\bar{x}_{k+1} = F\bar{x}_k + \bar{x}_0$ given. Subtracting this predictable part (trend) the a priori error

$$\tilde{x} = x - \bar{x} \quad (16)$$

satisfies

$$\tilde{x}_{k+1} = F \tilde{x}_k + Gu_k, \quad \tilde{x}_0 = 0 \quad (17)$$

$$\tilde{y}_k = H \tilde{x}_k + v_k, \quad y_k = H\bar{x}_k + \tilde{y}_k \quad (18)$$

Define now the quantity χ

$$\chi = \tilde{x}_{[2]}^{\Delta} \tilde{x} \times \tilde{x} \quad (19)$$

then

$$\begin{aligned} \chi_{k+1} &= F^{[2]} \chi_k \\ &+ (F \otimes G)(\tilde{x}_k \otimes u_k) + (G \otimes F)(u_k \otimes \tilde{x}_k) + G^{[2]} u_{[2]} \end{aligned} \quad (20)$$

where $\Lambda^{[k]}$ is the k-th Kronecker power of Λ . The prior expectation of χ satisfies now

$$\bar{\chi}_{k+1} = F^{[2]} \bar{\chi}_k + G^{[2]} \bar{u} \quad (21)$$

where $\bar{u}$ is the column stacked Q-matrix. $\bar{u} = \text{cat}(Q)$. Subtracting from (20) yields the unbiased form in

$$\tilde{x} = x - \bar{x} = \tilde{x}_{[2]} - E\tilde{x}_{[2]} \quad (22)$$

i.e.

$$\tilde{x}_{k+1} = F^{[2]} \tilde{x}_k$$

$$+ (F \otimes G) (\tilde{x}_k \otimes u_k) + (G \otimes F) (u_k \otimes \tilde{x}_k) + G^{[2]} \quad (23)$$

where $\bar{U}$ and $\bar{U}$ are respectively $E u_{[2]} = \text{cst } Q$ and $u_{[2]} = E u_{[2]}$. The initial condition for (21) is

$$\bar{x}_0 = E(\tilde{x}_0 \otimes \tilde{x}_0) = \text{cst } \bar{P}_0 \quad (\text{cst} = \text{column stack})$$

Similarly we get

$$\tilde{u}_{dk} = H^{[2]} \tilde{x}_k + \tilde{v}_k + H(v \otimes \tilde{x} + \tilde{x} \otimes v) \quad (24)$$

for

$$\tilde{y}_d = \tilde{y}_{[2]} - E\tilde{y}_{[2]} \quad (25)$$

$$\tilde{v} = v_{[2]} - E v_{[2]} \quad (26)$$

The equations (17), (18), (22) and (24) can be combined to yield the system

$$\begin{aligned} \begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_{k+1} &= \begin{bmatrix} F & 0 \\ 0 & F^{[2]} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k + \begin{bmatrix} G & 0 \\ 0 & G^{[2]} \end{bmatrix} \begin{bmatrix} u \\ \tilde{u} \end{bmatrix}_k \\ &+ \begin{bmatrix} 0 \\ G \otimes [F, 0] \end{bmatrix}_k (u_k \otimes \begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k) + \begin{bmatrix} 0 \\ [F, 0] \otimes G \end{bmatrix}_k (\begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k \otimes u_k) \end{aligned} \quad (27)$$

$$\begin{aligned} \begin{bmatrix} \tilde{y} \\ \tilde{u}_d \end{bmatrix}_k &= \begin{bmatrix} H & 0 \\ 0 & H^{[2]} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k + \begin{bmatrix} \tilde{v} \\ \tilde{v} \end{bmatrix}_k \\ &+ \begin{bmatrix} 0 \\ I_p \otimes [H, 0] \end{bmatrix}_k (v_k \otimes \begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k) + \begin{bmatrix} 0 \\ [H, 0] \otimes I_p \end{bmatrix}_k (\begin{bmatrix} \tilde{x} \\ \tilde{x} \end{bmatrix}_k \otimes v_k) \end{aligned} \quad (28)$$

Upon reordering the last terms in (27) and (28) this augmented system is exactly of the form considered in (1) and (2). The top block of the tensors are zero, resulting in a one-way coupling between x and $\tilde{x}$. Here (u', v') takes the role of the multiplicative noise w , hence in (3) we have

$$A = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 0 & R \\ 0 & 0 \end{bmatrix}, \quad \Omega = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$$

$$\text{Cov}(u', \tilde{u}' | v', \tilde{v}' | w') = \begin{bmatrix} Q & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{Q} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & R & 0 & 0 & 0 \\ 0 & 0 & 0 & R & 0 & 0 \\ \hline Q & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & R & 0 \end{bmatrix} \quad (29)$$

with

$$\mathcal{Q} = E u_{[2]} u_{[2]}' - \text{cst}(Q) \text{cst}(Q)' \quad (30)$$

$$\mathcal{R} = E v_{[2]} v_{[2]}' - \text{cst}(R) \text{cst}(R)' \quad (31)$$

The linear filter of section 2 for this augmented system yields least squares estimates for x and $\tilde{x}$ (say $\hat{x}$ and $\hat{\tilde{x}}$) when driven by $\tilde{y}$ and $\tilde{v}$ (which are computable from the data). Thus the proposed nonlinear filter consists of a feedforward model for $\tilde{x}_k$ and $\tilde{x}_k$ which together with the data y_k yields the driving terms $\tilde{y}_k$ for the second order filter. The filter itself follows the recursions of the previous section. Having estimates of x and $\tilde{x}_{[2]}$, one can form a better estimate $\hat{x}^*$ by solving

$$\begin{aligned} (x^* - \hat{x})' [P^{(1)}]^{-1} \\ + (x_{[2]}^* - \hat{\tilde{x}}_{[2]})' [P^{(2)}]^{-1} (x^* \otimes I + I \otimes x^*) = 0 \end{aligned}$$

References

1. D. L. Alspach and H. W. Sorenson, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 6, 1971.
2. R. S. Bucy and D. K. Senne, "Realization of optimum discrete-time nonlinear estimators," *Proc. Symp. Nonlinear Estimation Theory and Its Applications*, San Diego, CA, 1970.
3. P. J. Buxbaum and R. A. Haddad, "Recursive optimal estimation for a class of non-Gaussian processes," in *Proc. Symp. Computer Processing in Commun.*, Polytechnic Inst. Brooklyn, NY, April 1969.
4. Rolf Hendriksen, "The truncated second-order nonlinear filter revisited," *IEEE Trans. Automatic Control*, Vol. AC-27, No. 1, Feb. 1982, pp. 247-251.
5. A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.
6. C.J. Masreliez, "Approximate non-Gaussian filtering with linear state and observation relations," *IEEE Trans. Automatic Control*, Vol. AC-20, Feb. 1975, pp. 107-110.
7. Erik I. Verriest, "Error Analysis of Linear Recursions in Floating Point," *Proc. 1985 Int'l Conf. on Acoustics, Speech, and Signal Processing*, Tampa, FL, March 1985.
8. William J. Vetter, "Matrix calculus operations and Taylor expansions," *SIAM Review*, Vol. 15, No. 2, April 1973, pp. 352-369.

APPENDIX H

STOCHASTIC REDUCED ORDER MODELING OF DETERMINISTIC SYSTEMS

STOCHASTIC REDUCED ORDER MODELING OF DETERMINISTIC SYSTEMS

E. I. Verriest

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

Abstract

A novel approach to reduced order modeling is given. "Artificial" noise is introduced to reflect uncertainties in the reduced model, and a performance measure is associated to validate the approach. Connections with LQG design are discussed.

1. Introduction

The standard reduced order modeling problem is defined as follows. Consider the n^{th} order linear time invariant deterministic system.

$$\begin{aligned}\dot{x} &= Ax + Bu; \quad x(0) = x_0 \\ y &= Cx\end{aligned}\quad (1)$$

where $u(t)$ is a known input (although it will at times be assumed that $u(t)$ is a stochastic process). The problem is to design a linear system of the form

$$\begin{aligned}\dot{\hat{x}} &= F\hat{x} + Gu; \quad \hat{x}(0) = \hat{x}_0 \\ \hat{y} &= H\hat{x}\end{aligned}\quad (2)$$

of dimension $m < n$, with the objective to approximate the output y by $\hat{y}$ in some sense. All existing model reduction methods for deterministic systems, whether based on Markov parameter matching, Hankel matrix reduction, moment matching, Padé methods, balanced realizations, etc., all yield reduced models which can be described by inherently deterministic state space models of the form (2). However, in all these cases a deterministic part of the full model is "deleted" and this results necessarily in a loss of information. Uncertainty is thus intrinsic in the reduced model but this is never taken into account. This paper describes a design method where this uncertainty is "conserved" by artificially introducing observation and process noises. This uncertainty equivalence principle can of course also be brought into the design of reduced order models for stochastic systems. In the latter case the stochastic parameters in the system need to be appropriately augmented.

In this paper some plausibility arguments will be given. A full detailed analysis is deferred to a later paper.

2. A Bilinear Stochastic Model

The approach is based on the ideas of "balancing" for open loop systems and the LQG problem as developed by this author in [2-4], as well as some information theoretic ideas. Some model reduction methods are based on the "projection of dynamics." The reduced model uses for dynamics the projection of the dynamics of the original system to the subspace of the part of the state that is of interest. In balanced realizations this subspace is determined by inspection of the canonical gramian [3]. Within this framework, let a partitioning of A , B , C be given and let (F, G, H) equal (A_{11}, B_1, C_1) . Let the state x also be partitioned as (x_1, x_2) . Finally we assume that A is asymptotically stable. We approach the problem now in several steps.

Step 1: Assume that for the full order model the initial (partial) state x_{20} is unknown. By lack of information of any kind, let us take a probabilistic model, in which x_{20} is gaussian distributed with mean zero and covariance X_2 . This is motivated by the fact that the gaussian distribution is the maximum entropy distribution (i.e., least prejudiced) given the second moments. Most system analysis does indeed not go beyond second moments. Another motivation stems from the ease in dealing with gaussian distributions and this assumption will certainly be an improvement over merely setting $x_{20} = 0$. Of course the problem is now shifted to the determination of X_2 . Here we introduce a plausibility argument, based on considerations of a statistical ensemble of identical systems. The state x_0 is set up by inputs prior to $t = 0$. Again by lack of full knowledge, let these inputs be white gaussian $\sim N(0, q^2 I)$, where the q is now an unknown scalar. Next we assume that the initial condition x_0 is "typical" for a state set up by this white gaussian noise after the system reached a stochastic steady state. Assuming that (A, B, C) is balanced [3], implies that x_0 is zero mean gaussian distributed with covariance $q^2 A$, the canonical gramian. If x_{10} is known the factor q^2 can be estimated by standard statistical methods, e.g.,

$$q^2 = \frac{1}{m-1} \sum_{i=1}^m \frac{x_i^2}{\lambda_i} = q^2 (x_{10}, A_1^{-1} x_{10})_1, \quad m \neq 1 \quad (3)$$

where $x' = (x_1 \dots x_m)$. Then the covariance x_2 is $q^{10}(x_{10}, \Lambda_1) \Lambda_2$.

Step 2: Decouple the x_2 subsystem from the full order model. This means that we substitute the state x_2 by a stochastic variable $\hat{x}_2$ which is gaussian with covariance $q^2(t) \Lambda_2$ at each time, where now $q^2(t) = q^2(x_1(t), \Lambda_1)$ is as discussed in step 1. Clearly, in the original system $x_2(t)$ is not wildly fluctuating, hence it would be somewhat unrealistic to substitute x_2 for white gaussian noise. From the partitioned equation for x_2 , the essential dynamics of x_2 are governed by A_{22} , and the driving terms are both u and x_1 . Here, the approximation is to let x_2 be directly decoupled from x_1 and u , and be modeled as colored gaussian noise (with dynamics corresponding to A_{22}). An "indirect" coupling is retained by letting the covariance of x_2 correspond to x_1 in the above described fashion. To this effect, let

$$\dot{\hat{x}}_2 = A_{22} \hat{x}_2 + \hat{v} \quad (4)$$

where $\hat{v}$ is an assumed innovations process with steady state covariance determined from the Liapunov equation

$$\text{cov}(\hat{v}) = -A_{22}(q^2 \Lambda_2) - (q^2 \Lambda_2) A_{22}' = q^2 B_2 B_2' \quad (5)$$

where the latter equality follows from the balanced realization properties [3]. The system (1) is then approximated stochastically by the bilinear reduced order model driven by the (normalized) colored noise θ

$$\dot{\zeta}_1 = A_{11} \zeta_1 + B_1 u + A_{12} \theta \quad (6)$$

$$n = C_1 \zeta_1 + C_2 \theta \quad (7)$$

where

$$\dot{\theta} = A_{22} \theta + \frac{1}{m-1} B_2 \zeta_1 \Lambda_1^{-\frac{1}{2}} v; \quad v \sim N[0, I] \quad (8)$$

Step 3: The reduction is performed as a simplification of the intervening colored noise. Rather than modeling the unknown dynamics by the process θ , we use now two correlated white gaussian processes, w and μ with covariance $\text{cov}(w) = \begin{pmatrix} W & R \\ R' & M \end{pmatrix}$ (9). The new model as in (6) and (7) but with $A_{12} \theta$ and $C_2 \theta$ substituted by w and μ . Several criteria can now be chosen, based on different design restrictions, to select the covariances. If the output approximation must be smooth, then set $\mu = 0$ and thus also $R = 0$ and $M = 0$. W can then be chosen such that $\text{cov}(n) = \text{cov}(\hat{y})$ (i.e., we retain the same uncertainty in the model). If a wildly fluctuating model output is tolerable, then equality of the output covariances may be combined with the equality of the correlation between process and observation noise. It can for instance be shown that if $2m < n$ and A_{12} nonsingular (which must be true for the SISO balanced realization), then many solutions exist.

Remark 1) The approximation in step 2 will be better if the components $A_{12} x_2$ and $C_2 x_2$ are "small" relative to respectively $A_{11} x_1 + B_1 u$ and $C_1 x_1$. This is the basis for the next section on LQG modeling.

3. Reconciliation with LQG Design

The well known separation property for the solution of the LQG problem breaks down in the LQG modeling. (See [1] and [5])

The approach taken above has motivated (quantitatively) that model reduction should be accompanied by proper specification of (artificial) process and measurement noise covariances. Equivalently, the noise covariances one selects in LQG design should reflect the unmodeled dynamics. In order that this approximation of unmodeled dynamics by noise is tolerable, the variances of these components should not be too large compared to the "main components." From remark 1) above it follows that for the balanced realization, this will be guaranteed if a performance index (P.I.)

$$E\{ ||A_{12} x_2||^2 + \alpha ||C_2 x_2||^2 + \rho ||A_{11} x_1 + B_1 u||^2 + \alpha ||C_1 x_1||^2 \} dt \quad (12)$$

is introduced for some $\alpha > 0$ and $0 < \rho < 1$. But x_1 is not available in the reduced model, so its covariance must be estimated as in step 1. The end result is the P.I. with integrand

$$\begin{pmatrix} x_1' & u' \end{pmatrix} \begin{pmatrix} \Omega & \rho A_{11}' B_1 \\ \rho B_1' A_{11} & \rho B_1' B_1 \end{pmatrix} \begin{pmatrix} x_1 \\ u \end{pmatrix} \quad (13)$$

$$\text{where } \Omega = \frac{1}{m-1} \Lambda_1^{-1} \text{Tr}(x^2 C_2' C_2 + A_{12}' A_{12}) \\ \Lambda_2 + \rho(A_{11}' A_{11} + \alpha^2 C_1' C_1) \quad (14)$$

For systems with slight nonlinearities, standard perturbation methods yield linear models. If again the perturbing nonlinearity is "set to zero" a loss of information results. We can then again match the uncertainty.

In summary then, noise covariances for the LQG model should reflect the degree of accuracy of the assumed model. Qualitatively speaking, in the modeling stage, model uncertainty should be traded for purely stochastic inputs which are simpler to deal with in the analysis and design.

In order to validate the modeling, care must be taken that the expected deviations are not too large by accentuating the cost of such deviations.

REFERENCES

- [1] N. Athans, "The Role and Use of the Stochastic Linear-Quadratic-Gaussian Problem in Control System Design," IEEE Transactions Automatic Control, Vol. AC-16, No. 6, Dec. 1971, pp. 529-532.
- [2] S. I. Verriest, "Low Sensitivity Design and Optimal Order Reduction for the LQG Problem," Proc. 24th Midwest Symp. Circuits Systems, Albuquerque, NM, June 1981, pp. 363-369.
- [3] S. I. Verriest and T. Kailath, "On Generalized Balanced Realizations," IEEE Trans. Automatic Control, Vol. AC-28, No. 8, Aug. 1983, pp. 833-844.
- [4] S. I. Verriest, "Suboptimal LQG Design via Balanced Realizations," IEEE Conf. Dec. Contr., San Diego, CA, Dec. 1981, TA3.

APPENDIX I

UNCERTAINTY EQUIVALENT REDUCED ORDER MODELS FOR DISCRETE
SYSTEMS

ABSTRACT

Model reduction invariably involves a trade off between the available information and the simplicity of the retained model. This information loss leads to an uncertainty about the output of the true system given the output of the reduced model for identical inputs. In this paper a novel approach to the reduction problem is given by incorporating this induced uncertainty in the reduced model. Geometrically, the idea is to construct a tube centered on the model output, to which the actual system output belongs with a high degree of confidence.

1. INTRODUCTION

Given an input-output description of a general deterministic system, a realization can be given in state space form. An essential feature (in fact, the defining property) of the state is that it contains all the information one needs to have about the past inputs to the system in order to predict its output given the future inputs. The concept is obviously equivalent to that of a sufficient statistic in this case.

A deterministic state space model evolves in R^n (or a submanifold of it), for some integer n , the "order" of the system. A reduced model corresponds with a lower dimensional (say $m < n$) state space model, and it is clear that this reduced state cannot contain the full information necessary to produce the exact output. This loss of information entails an uncertainty in the output of the reduced model regarding the true model. To our knowledge, no existing model reduction methods, whether based on Markov parameter matching, Hankel matrix reduction, moment matching, Padé methods, singular perturbation or balanced realizations, etc., incorporate this inherent "uncertainty" due to the "deletion" of a part of the true model.

This paper focuses on the reduction of an n^{th} order linear time invariant system with $\dim u = n_1$ and $\dim y = n_0$

$$x(k+1) = Ax(k) + Bu(k) \quad ; \quad x(0) = x_0 \quad (1)$$

$$y(k) = Cx(k)$$

The classical problem is to design a linear system

$$z(k+1) = Fz(k) + Gu(k) \quad ; \quad z(0) = z_0 \quad (2)$$

$$\hat{y}(k) = Hz(k)$$

of dimension $m < n$, with the objective that $\hat{y}$ approximates y in some sense. As noted earlier, the model (2) necessarily results in a loss of information, but a "user" of the reduced model does not have knowledge of this resulting uncertainty. A design method where this uncertainty is "conserved" by introduction of artificial observation and process noises is outlined in [3]. Uncertainty equivalence is established through equality of certain covariances. The approach is

based on the ideas of "balancing" [2], as well as some information theoretic ideas. Clearly, this uncertainty equivalence principle can be brought into the design of reduced order models for stochastic systems. In the latter case, the stochastic parameters in the system need to be artificially augmented. The next section describes an approach to open loop model reduction based on balanced realizations which leads to a bilinear stochastic reduced order model.

2. A BILINEAR STOCHASTIC MODEL

Some model reduction methods (notably the ones based on a modal decomposition and on balancing techniques) are based on the "projection of dynamics." The reduced model uses for dynamics the projection of the dynamics of the original system to the subspace of the part of the state that is of interest. In balanced realizations this subspace is determined by inspection of the canonical gramian [2]. Within this framework [2], let a partitioning of A, B, C be given and let (F, G, H) equal (A_{11}, B_1, C_1) . Let the state x also be partitioned as (x_1', x_2') . For future reference, the propagation of the first and second order moments for a special bilinear realization is given. Its proof is straightforward.

Theorem: Given the bilinear stochastic realization

$$x_{k+1} = Ax_k + Bu_k + q(x_k)w_k \quad (3)$$

$$y_k = Cx_k + q(x_k)v_k \quad (4)$$

where

$$q(x_k)^2 = x_k' \Omega x_k \quad (5)$$

and u_k is deterministic and w_k is a standard white Gaussian noise sequence then the updates of first and second order moments are

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) \quad (6)$$

$$\Pi(k) = \hat{x}(k)\hat{x}'(k) + P(k) \quad (7)$$

$$P(k+1) = AP(k)A' + WW'Tr(\Omega\Pi(k)) \quad (8)$$

$$\hat{y}(k) = C\hat{x}(k) \quad (9)$$

$$P_y(k) = CP(k)C' + VV'Tr(\Omega\Pi(k)) \quad (10)$$

The uncertainty equivalent modeling problem is now approached in several steps. First, the uncertainty in the full order model, given only a partial initial state, is evaluated. In Step 2, an uncertainty equivalent, one-way decoupled model is set up, leading to a colored noise driven reduced order model. The last step entails the actual simplification. Two methods are suggested to approximate the colored noise model by a white noise model.

AD-A194 685

APPROXIMATIONS AND IMPLEMENTATIONS OF NONLINEAR
FILTERING SCHEMES(U) GEORGIA INST OF TECH ATLANTA
SCHOOL OF ELECTRICAL ENGINEERING A H HADDAD ET AL

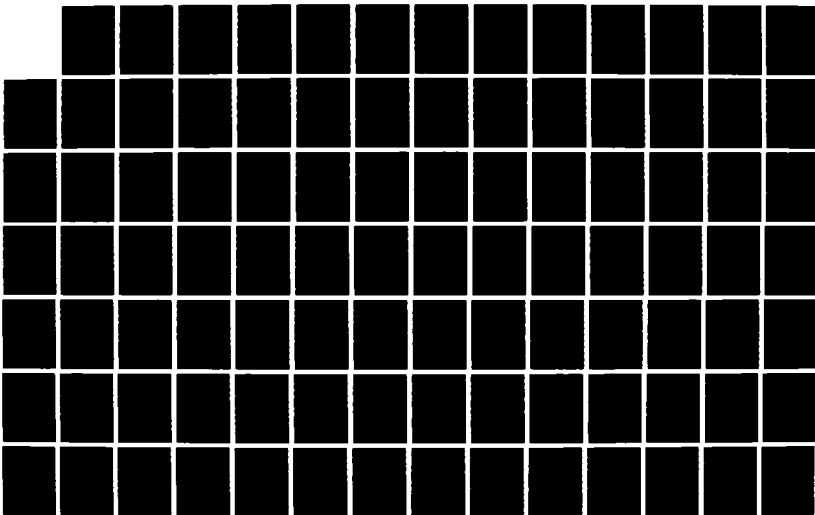
2/4

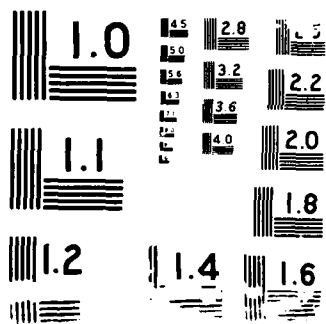
UNCLASSIFIED

FEB 88 AFATL-TR-87-73 F08635-84-C-0273

F/G 12/3

NL





2.1 Equivalence for Incomplete Knowledge of the State

Assume that for the full order model (1) in balanced form, the initial (partial) state x_{20} is unknown. By lack of information of any kind and motivated by the fact that a gaussian distribution is the maximum entropy (i.e., least prejudiced) distribution given the second moments, the uncertainty in x_{20} is modeled by a gaussian random vector with zero mean and covariance X_2 , yet to be specified. This assumption will certainly be an improvement over merely assigning zero to the components of x_{20} . In [3] it was argued, based on statistical considerations, that a proper choice for this covariance is

$$X_2 = \frac{1}{m} x_{10}' A_1^{-1} x_{10} A_2 \\ = q^2(x_{10}, A_1) A_2 \quad (11)$$

2.2 Decoupling of the x_2 Subsystem

The next approximation substitutes the state x_2 by a stochastic variable x_2 which is gaussian, with covariance $q^2(t) A_2$ at each time, where now $q^2(t) = q^2(x_1(t), A_1)$ is as discussed in step 1. From the partitioned equation for x_2 , the essential dynamics (as captured by the correlation function) of x_2 are governed by A_{22} , and the driving terms are both u and x_1 . Here, the approximation is to let x_2 be directly decoupled from x_1 and u , and be modeled as colored gaussian noise (with dynamics corresponding to A_{22}). An "indirect" coupling is retained by letting the covariance of x_2 correspond to x_1 in the above described fashion. To this effect, let

$$\hat{x}_2(k+1) = A_{22} \hat{x}_2(k) + \hat{v}(k) \quad (12)$$

where $\hat{v}$ is an assumed innovation process with steady state covariance determined from the Lyapunov equation

$$\text{cov}(\hat{v}) = q^2 A_2 - A_{22} q^2 A_2 A_{22}' \\ = q^2 (A_{21} A_1 A_{12}' + B_2 B_2') \quad (13)$$

The latter equality follows from the balanced realization properties [2]. The system (1) is then approximated stochastically by the bilinear reduced order model driven by the colored noise θ

$$z(k+1) = A_{11} z(k) + B_1 u(k) + A_{12} \theta(k) \quad (14)$$

$$\eta(k) = C_1 z(k) + C_2 \theta(k) \quad (15)$$

where

$$\theta(k+1) = A_{22} \theta(k) + q(A_{21} A_1^{1/2} \quad B_2) \begin{pmatrix} w(k) \\ v(k) \end{pmatrix}, \quad (16) \\ \begin{pmatrix} w \\ v \end{pmatrix} \sim N[0, I_{m+n_1}]$$

The propagation properties for the model follow then from Theorem 1 using appropriate matrices.

Remark: The approximation will be better if the components $A_{12} x_2$ and $C_2 x_2$ are "small" relative to, respectively, $A_{11} x_1 + B_1 u$ and $C_1 x_1$. This constitutes a basis for the LQG modeling. The ideas are developed in [3].

2.3 White Noise Model

Rather than modeling the unknown dynamics by a colored sequence, further simplification results if a white noise sequence is used instead. There are two ways in which one can proceed: (1) matching the cause and (2) matching the effect. Causal matching is the simplest. White noise sequences of matching covariance are substituted wherever the colored sequence enters. This leads to the innovations model

$$z(k+1) = A_{11} z(k) + B_1 u(k) + q(k) A_{12} A_2^{1/2} v(k) \quad (17)$$

$$\eta(k) = C_1 z(k) + q(k) C_2 A_2^{1/2} v(k) \quad (18)$$

where now $v(k)$ is a standard white Gaussian sequence. By Theorem 1 the stochastic model update equations are then easily obtained.

Note that with this method a deviation of the covariances of the outputs necessarily results between the colored noise and the white noise model. In other words, the "effects" are changed. This suggests then at once the alternative method of matching the effects. The underlying idea is that the optimal (linear) filter for the equivalent white noise model should not result in a state covariance which is less than the covariance for the filter for the colored noise system. The problem is, however, that the optimal filtering problem for the colored model involves the full order state, and may, therefore, be computationally prohibitive. After all we want a reduced model, not an approximation of the same order! Using the results of Jain [1], a suboptimal filter can be designed which yields an error covariance which is guaranteed to be below a certain bound. Only the covariance of the colored noise needs to be known. We suggest the following scheme. Assume a white noise model of the form,

$$z(k+1) = A_{11} z(k) + B_1 u(k) + q(k) w(k) \quad (19)$$

$$\eta(k) = C_1 z(k) + q(k) v(k) \quad (20)$$

The matrices $Q = E(w w')$ and $R = E(v v')$ are determined in such a way that the innovations covariance of the filter for (19)-(20) is the same as for the bound in [1]. Details will be explained in a forthcoming paper.

REFERENCES

- [1] B. N. Jain, "Bounding Estimators for Systems with Colored Noise," IEEE Trans. Auto. Control, Vol. AC-25, pp. 365-368, June 1975.
- [2] E. I. Verriest and T. Kailath, "On Generalized Balanced Realizations," IEEE Trans. Auto. Control, Vol. AC-28, No. 8, pp. 833-844, August 1983.
- [3] E. I. Verriest, "Stochastic Reduced Order Modeling of Deterministic Systems," Proceedings of the 1985 Automatic Control Conference, Boston, MA, June 1985.

APPENDIX J

MODEL REDUCTION VIA BALANCING, AND CONNECTIONS WITH OTHER

METHODS

MODEL REDUCTION VIA BALANCING, AND CONNECTIONS WITH OTHER METHODS

Erik I. Verriest
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

This paper starts with a rather philosophical viewpoint on the concepts of modeling, model reduction, and randomness. The theory of open-loop deterministic balancing is introduced as a particular implementation of a model reduction scheme. The discussion focusses on the choice of the criterion. Thus motivated, it is shown that similar ideas can be employed in the reduction of optimally controlled systems under the presence of noise, leading to the LQG-balanced realizations. This connects to the stochastic balanced realizations. Finally, different stochastic realization algorithms are cast in the common framework of the RV-coefficient, and the deeper geometric significance of this measure is explored.

1 INTRODUCTION:

1.1 Modeling and Model Reduction

Until recently, modeling has been to a large extent a heuristic and unrigorous process where ad-hoc procedures abounded. For this reason, further attention and research to this problem has been more than welcome. In effect, the first half of the eighties has seen a proliferation in modeling and model reduction methods which are firmly based on mathematical rigor. (e.g. [1], [2], [3])

The dichotomy between modeling and model reduction is rather weak and different authors may provide different definitions. Perhaps the most intuitive notion is to let modeling be the process whereby an abstract mathematical model is matched to the physical reality, and model reduction the process whereby a simpler mathematical model is derived

from an existing mathematical model. In this regard modeling is what is usually called "Identification", while model reduction belongs to the realm of Approximation Theory.

Keeping in mind that the physical world and thus all real-life systems are the basic entities, perhaps eluding a description as a whole, one can only abstract some aspects of its behavior and model these properties in a formal theory. In what follows then, it will be assumed that the "physical reality" is that what allows observation. Modeling thus infers a procedure which formalizes in a mathematical abstraction some aspects of the behavior of the physical entity. Clearly such a formalization cannot be unique.

Together with the mathematical abstraction (model) one must give its scope, i.e. which aspects of the physical system it models. Models inherently have their limitations. A linear small signal model of a transistor for instance, no matter how accurate its parametrization, will be unable to predict the switching properties of digital transistor circuits. Clearly then, the scope of the model should be matched to whatever one expects from the model. In the study of the kinematics of machinery, there is no need to apply the theory of relativity, but in the study of particle accelerators, the classical theory no longer suffices.

Once the scope of the model has been laid down, one must determine the accuracy of that model. How well does it describe the domain-aspect of the physical reality? A better model is obviously the one that, given the same domain, predicts the behavior of the physical system more accurately. Biped motion can be crudely modeled with a ball-and-stick model, where for instance each stick is rigid, and perhaps massless. A better model would be the one incorporating the distributed nature of the masses, actuators, etc.

Also, one must be able to explain when a given model is more accurate, or better than another one. More specifically, this consists in finding a measure for the accuracy, or more abstractly a suitable topology, with a meaningful physical interpretation, so that the approximation problem for models, within the same scope, is well defined.

Thirdly, another definitely more practical aspect of a formal theory is its complexity. Roughly speaking, complexity refers to the number of ad-hoc rules (postulates) that the theory requires, as well as the smallest number of parameters that need to be specified a priori in order to obtain uniquely predictable (computable) answers within the model. In a Newtonian mechanistic model the whole universe would be predictable given the initial position, velocity and mass of each particle constituting the universe. In this theory there is one basic postulate: the (Newtonian) universal law of gravity, but the parameter set is ... well, very big indeed. Such a model would clearly be impractical, if not unfeasible, if one were interested in studying the dynamics of the solar system, or the kinetics of a gas in a container.

To summarize, every formal modeling theory should be accompanied by these three quantifiers:

- its domain of validity
- its predictability or accuracy
- its complexity

Hence, there exists only a partial ordering between models, and blank statements as "Model A is better than model B." definitely do not make any sense without any indication of these three quantifiers. Even given the quantifiers, different models may simply not be comparable. Whether one favors a general model of large scope, or several specialized ones of lower scope and complexity is now more a matter of personal taste. Of course the particular purpose or objective of the model should influence such a choice.

Model reduction problems aim at reducing the complexity although there generally is a trade off with the accuracy and the scope. Within the established mathematical framework, this resorts to finding a more attractive subset of the space which is dense (with respect to the topology) in the given space, as for instance in polynomial approximation. Alternatively, it may mean the search for a lower dimensional subspace of some given space, e.g. finite element approximations of distributed systems. At any rate,

the modelers dream is to come up with a mathematical model, which is suitably small in order to allow computable predictions of the reality.

Model reduction can be accomplished in many ways: For example, suppose that one has a large dimensional system, perhaps of weakly interacting subsystems. Existing techniques find a lower dimensional model, e.g. by aggregation. There is no doubt that the result will be a simpler model. On the other hand, one could take the opposite approach, and let the number of weakly interacting subsystems approach infinity, only to realize a statistical or probabilistic description of the system. Such a probabilistic description may result in a fewer number of parameters (e.g. first and second order moments). In fact, this is exactly the approach of statistical dynamics. Again, which approach is favorable will depend on the purpose of the model. If one is only interested in the average behavior of the system, then the statistical description may be preferable. One does not need to know the detailed trajectories of the gas molecules in order to understand the workings of an internal combustion engine.

1.2 Stochastic Models and the Origins of Randomness

In the previous paragraph, we already hinted at building statistical models. The observed data set on which one tries to model some behavior, typically shows fluctuations. These fluctuations arise from two origins:

- i) Some variables (parameters) of the system may be random. In this sense the resulting probabilities are unambiguously defined, i.e. the randomness is imposed from the "outside". (e.g. random boundary conditions).
- ii) Randomness can be introduced in an arbitrary way, to reflect our incomplete knowledge of an exact description of a system. For instance this can be due to uncertainties of a real probabilistic nature (e.g. quantum uncertainty). This uncertainty further arises when the number of variables is so large that a correct description would be practically

impossible. Randomness is then used to replace a knowledge which is too detailed to be useful in practice.

A practical methodology for discarding information can then be organized as follows:

- Retain only a few simple features which seem relevant to the problem. (e.g. based on the different physical consequences that result from the different ways of complexity reduction).
- Give a probabilistic description. (This allows statistical predictions, despite the incomplete information).
- Compute observed quantities from within this model and compare these with experimental results. Here the "scope" and the "accuracy" are tested, thus allowing "feedback" or interaction in the modeling procedure.

A fundamental assumption is the MARKOV assumption, which is justifiable as follows:

The large set of variables, giving an exact complete microscopic description of the system can be divided in two classes, according to their relaxation times. If a first set $\{x\}$ has relaxation times, much greater than all the other variables in the second class, then the timescale of the description (amounting to the scope of the model), is chosen intermediate to the long and short relaxation times. Hence, all memory effects are accounted for by the variables $\{x\}$, and it is adequate to assume that they form a Markov process.

Another frequently made assumption is that of STATIONARITY, implying that

- i) all external influences on the system are time-independent on the chosen timescale.
- ii) the classification of all variables in "fast" and "slow" is preserved during the evolution of the system.

2 OPEN LOOP BALANCING

2.1 Reachability and Observability

A state space model of a continuous time (the theory for discrete time systems is very similar and omitted) linear system with n inputs and p outputs is characterized by a triple of matrices (F, G, H)

$$F \in \mathbb{R}^{n \times n}; \quad G \in \mathbb{R}^{n \times m}; \quad H \in \mathbb{R}^{p \times n}$$

where n is the order of the system. In general, the matrices are indexed by the reals $\mathbb{R}$.

For continuous time systems the relations are

$$\dot{x}(t) = F(t)x(t) + G(t)u(t) \quad (2.1)$$

$$y(t) = H(t)x(t) \quad (2.2)$$

If F , G and H are invariant with time, it is well known [4] that the reachability and observability of the system are determined by the fullrankness of the reachability and observability matrices, respectively $[G, FG, \dots, F^{n-1}G]$ and $[H', F'H', \dots, F^{n-1}H']$. However, the rankdefect of a matrix is very difficult to determine numerically because of the finite precision arithmetic of all computers. Moreover, these criteria do not provide any means to attach a measure of the degree of observability or reachability of the given system. A quantitative measure of the reachability ($\mathcal{R}$) or observability ($\mathcal{Q}$) in some interval (t_0, t_1) is obtained via the (weighted) Gramian matrices, defined as: ($\Phi(\dots)$ is the transition matrix of F)

$$\mathcal{Q}_W[t_0, t] = \int_{t_0}^t \Phi'(\tau, t_0) H'(\tau) W(\tau) H(\tau) \Phi(\tau, t_0) d\tau \quad (2.3)$$

$$\mathbb{R}_M[t_0, t] = \int_{t_0}^t \Phi(t, \tau) G(\tau) M^{-1}(\tau) G'(\tau) \Phi'(t, \tau) d\tau \quad (2.4)$$

Note that these matrices are well defined also in the timevarying case, as long as the integrals converge. The matrices $W(t)$ and $M(t)$ are assumed to be (positive or negative) definite, (usually identity). An interesting interpretation of these Gramians as weighting matrices for energies and uncertainties is given in the following subsection. In fact, this interpretation forms the basis for the model reduction algorithms to be introduced in the next subsection. The last subsection then describes the properties of the so-called balanced realizations, which were first introduced by Moore [5] for the time-invariant case.

2.2 Interpretation of the Gramians.

2.2.1 Deterministic

We start with simple thought experiments. Assume that the relevant input and output signals are in L_2 . Let the system be in the state x_0 initially. The output of the undriven systems is

$$y(t) = H(t) \Phi(t, t_0) x_0 \quad (2.5)$$

In general, the weighted L_2 -norm is a particular measure of the "strength" in the signal, even though there may not be an underlying energy in a physical sense. The cross terms measure the degree of "interference" between the different components. Also, it is always possible to renormalize or take linear combinations of the existing output signals that have a more direct physical interpretation in terms of energy. Equivalently, one can define a weighting matrix W for the outputs, thus effectively measuring the "energy" as a weighted L_2 -norm. With this generalization, the available W -measured output-energy U_W in

the interval t_0 to t_f for a system in state x_0 at time t_0 is given by,

$$\begin{aligned}
 U_W &= \int_{t_0}^{t_f} y(t)' W(t) y(t) dt \\
 &= \int_{t_0}^{t_f} x_0' \Phi'(t, t_0) H'(t) W(t) H(t) \Phi(t, t_0) x_0 dt \\
 &= x_0' \Phi_W [t_0, t_f] x_0
 \end{aligned} \tag{2.6}$$

The generalized Observability Gramian $\Phi_W [t_0, t_f]$ is a weighting matrix for the output L_2 -measure given the initial state. If the system is observable (i.e. Φ_W nonsingular), the state x_0 can be recovered as (assuming that the system is undriven in $[t_0, t_f]$)

$$x_0 = (\Phi_W [t_0, t_f])^{-1} \int_{t_0}^{t_f} \Phi(t, t_0) H(t)' W(t) y(t) dt \tag{2.7}$$

Consider now the dual problem of determining the inputs which drive the system from the zero state at t_0 to any arbitrary state x_f at t_f . If the matrix $\mathfrak{R}_M [t_0, t_f]$ is nonsingular, then a particular input achieving this is

$$u(t) = M(t)^{-1} G(t) \Phi(t_f, t) (\mathfrak{R}_M [t_0, t_f])^{-1} x_f \tag{2.8}$$

The optimality properties of this input are well-known [6]. It is the input with the least amount of "energy", as measured in a M -weighted L_2 -norm.

$$\|u\|_M^2 = \int_{t_0}^{t_f} u(t)' M(t) u(t) dt \tag{2.9}$$

The corresponding minimal energy is

$$U_R = x_f' (R_M [t_0, t_f])^{-1} x_f \quad (2.10)$$

Again we see the role played by the M-weighted Gramian matrix. Its inverse appears as a weighting matrix for the minimal steering effort to the state x_f from the zero state.

2.2.2 Stochastic.

Here also we start with two thought experiments. One characterizing "uncertainties" relating to the inputs to the system, the other one relating the state uncertainty to the outputs.

Let the system be driven by a white gaussian (vector) input signal, of zero mean, and covariance matrix $Q(t)$. Assuming that this input is uncorrelated with the initial state of the system, the state covariance matrix $P(t)$ at time t is given by

$$P(t) = \Phi(t, t_0) P(t_0) \Phi'(t, t_0) + R_{Q^{-1}}[t_0, t] \quad (2.11)$$

The first term equals the covariance $\Pi(t, t_0)$ for the free-running undisturbed system. The second term is the generalized Q^{-1} -weighted Reachability Gramian (2.4) for $M = Q^{-1}$. It is a measure of the uncertainty induced in the state by a maximally random input. The disturbability of the state (as measured by the covariance) in the direction d by a white Gaussian input is given by

$$d'P(t)d = \text{Tr } DP(t) \quad (2.12)$$

where $D = dd'$ and Tr is the trace function. The expected value of the A-weighted state "energy" in the realization is

$$\begin{aligned} E x(t)' A(t) x(t) &= \text{Tr } A(t)P(t) \\ &= \text{Tr } \Pi(t, t_0) + \text{Tr } A(t) R_{Q^{-1}}[t_0, t] \end{aligned} \quad (2.13)$$

Here $\mathbb{R}_{Q^{-1}}[t_0, t]$ appears as a weighting for $D(t)$ and $A(t)$ under the trace-norm. The second term in (2.13) is interpreted as the average energy increase in the states of the given realization due to the process noise with covariance $Q(t)$.

Finally, consider the state estimation problem for a system with observation noise, but no driving terms. If the measurement noise is white with covariancematrix $R(t)$, and, for simplicity, assumed to be uncorrelated with the initial state x_0 , then the (Kalman filter) solution to the problem leads to the classical result ($S_0 = P(t_0|t_0)^{-1}$)

$$P(t_0|t)^{-1} = S_0 + \mathbb{Q}_{R^{-1}}[t_0, t] \quad (2.14)$$

The matrix $\mathbb{Q}_{R^{-1}}[t_0, t]$ is a (matrix valued) measure for the information (or $\mathbb{Q}_{R^{-1}}^{-1}[t_0, t]$ for the uncertainty) conveyed by the observation process in (t_0, t) about the initial state x_0 , and is usually referred to as the "Information matrix" in the estimation literature. In particular, if there is no prior information, $P(t_0|t)^{-1}$ is the zero matrix, and

$$P(t_0|t) = (\mathbb{Q}_{R^{-1}}[t_0, t])^{-1}$$

The above illustrates in a simple way how $\mathbb{R}^{-1}$ and $\mathbb{Q}$ relate to (generalized) energies, while their inverses $\mathbb{R}$ and $\mathbb{Q}^{-1}$ have to do with "uncertainties". The lower the required (minimal) energy to reach a certain state is, the more "reachable" that state is. Similarly, the higher the output energy available from the system, the more information we have about that system, and the smaller the errorcovariance of the filtered initial state. The Gramians provide, therefore, a suitable measure for the degrees of reachability and observability in a system.

With these remarks serving as a motivation, we proceed to the formal definitions.

2.3 Balanced Realizations

2.3.1 The Canonical Gramian

Under a similarity transformation of the state space form of a system, the quantitative reachability and observability properties of a realization are changed. Indeed if $T(t)$ is the (nonsingular) transformation

$$T: (F, G, H) \rightarrow (TFT^{-1}, TG, HT^{-1})$$

then the Gramians for the new realization are

$$T: \mathcal{R}[t_0, t_f] \rightarrow T(t_f) \mathcal{R}[t_0, t_f] T^{-1}(t_f)$$

$$T: \mathcal{Q}[t_0, t_f] \rightarrow T(t_0)^{-T} \mathcal{Q}[t_0, t_f] T^{-1}(t_0)$$

It has been shown [7] that if the matrices F , G , H and the weights W and M are real analytic functions of time, and if the system is completely reachable and observable, then a similarity transformation exists such that both $\mathcal{R}$ and $\mathcal{Q}$ are diagonal and equal. If the diagonal elements are separated, then one can define a (unique) canonical form by inducing some ordering in these elements. In the time invariant case, it is customary to order them according to decreasing magnitude. In the sequel we shall refer to these as the CANONICAL ELEMENTS, and to the Gramians in balanced form as the CANONICAL GRAMIAN. The open loop canonical Gramian will be denoted as Λ . (In the signal processing and digital filtering context, the canonical elements are also known as the second order modes [8]). The resulting realization (TFT^{-1}, TG, HT^{-1}) is then called "balanced with respect to the weights W and M ". (Usually, only balancedness is considered with respect to the weights $W=M=I$). Algorithms for obtaining balanced realizations are based on the singular value decomposition of the Gramians in an arbitrary realization ([5], [7]). Recently more direct methods have been obtained for computing the balanced realizations for time invariant systems ([9], [10], [11], [12], [13]). The timevarying balanced

realizations are an extension of the original balanced realizations for time invariant systems introduced by Moore, and were introduced in [7]. In view of the interpretations of the Gramians developed in the previous section, it is clear that in each coordinate direction of the balanced state space the degree of reachability and observability is the same.

2.3.2 Model Reduction via Balancing

In order to fix the ideas on how this might be used as a criterion for model reduction, consider a nonminimal time-invariant state space realization of a system. It is well known that such a realization is nonreachable and/or unobservable [4]. A minimal realization, having identical input-output properties if the system is initially at rest, can be obtained by removing these unobservable or nonreachable modes from the original description, e.g. via a truncation (projection of dynamics) of the standard decomposition of the nonminimal system, thus effectively deleting the unreachable and noncontrollable parts of the state space. The Gramians give now a quantitative measure for observability and reachability, rather than the binary value assigned by the (Kalman) criterium. If one were now to "delete" a component which has a high cost associated with its reachability, then this component may have very good observability properties, and therefore be very significant in the input-output description of the system. Since in fact the state intervenes only as an interface between input and output, a transformation (for instance just a scaling) can be used to yield a new representation in which this difficult-to-reach state component has become very easy to reach. The opposite would then also be true for its observability properties. Clearly, component reachability and component observability is not an absolute criterion for the importance towards the input-output or external description. What one needs to look for is invariants with respect to arbitrary state space transformations. The product of the reachability and observability Gramians transforms under T as a similarity. Hence the eigenvalues of this product are invariants.

But these eigenvalues are exactly the squares of the canonical elements. Hence it turns out that the relative importance of a (balanced) state space component with respect to the external system behavior is quantitatively determined by the "joint degree of reachability and observability" associated with this system dimension. By virtue of the interpretations, we developed in the previous section, this is exactly described by the elements of the Gramians of the balanced realization. Based on this description, the canonical elements can be used by the system analyst or control designer to decide which components to use in a reduced order model for the original system. Such a reduced order model is then obtained by "projection of dynamics". One partitions the original system (in balanced form) as

$$\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \quad \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \quad [H_1 \quad H_2]$$

where it is assumed that the canonical elements are ordered with respect to their magnitude. The reduced order model is then obtained as (F_{11}, G_1, H_1) . It simply means that the components which were difficult to control and observe are considered as completely uncontrollable and unobservable, and subsequently, the minimal realization is obtained. In reference to section 1, our topology is derived from the trace of the canonical Gramian, under the restriction of "Projection of Dynamics". The above "projection-of-dynamics" method is also applied in the discrete case. However, it leads to some self-inconsistency, since the reduced models of the discrete balanced system are themselves not balanced. Similar interference-effects are also known in realization theory. For instance, the reduction of a Hankel matrix via a singular value decomposition does not yield a matrix with Hankel structure in general. This has been a steady source of critique to the method.

2.3.3 Properties of Balanced Realizations

A basic property of the Gramians introduced in section 2.1 is that they satisfy the following Lyapunov equations (without loss of generality, we take M and W as the identity matrix). (t_0 and t_f fixed):

$$\dot{R}[t_0, t] = F(t)R[t_0, t] + R[t_0, t]F'(t) + G(t)G(t)' \quad (2.15a)$$

$$R[t_0, t_0] = 0 \quad (2.15b)$$

$$\dot{Q}[t, t_f] = -F(t)'Q[t, t_f] - Q[t, t_f]F(t) - H'(t)H(t) \quad (2.16a)$$

$$Q[t_f, t_f] = 0 \quad (2.16b)$$

More general formulas for t_0 and t_f depending on t have been obtained in [7]. They are of interest in "Sliding Interval" Balancing and Model Reduction. It follows that for the balanced realization of a time invariant realization with $t_0 = -\infty$ and $t_f = \infty$, the canonical Gramian satisfies the symmetrical equations.

$$FA + AF' + GG' = 0 \quad (2.17)$$

$$FA + AF + H'H = 0 \quad (2.18)$$

These equations form the basis for the derivations of a whole set of nice properties for balanced realizations (see [9], [11], and [14]). e.g., for some signature matrix E (a diagonal matrix having either +1 or -1 as diagonal elements) one can show that for SISO systems

$$EFE = F' ; EG = H' \quad (2.19)$$

3 Balancing in the LQG-sense

In the open loop case, the balanced realization led to a natural selection of reduced order models through the "projection of dynamics". Adopting this procedure for the design

of a reduced order controller is very dangerous however, due to the feedback around the system. A more direct approach is needed, treating the closed loop as a whole. In fact, the degree of uncertainty (i.e., the noise covariances) and the performance index or cost-functional of the system should be taken into account for the selection of a reduced model. Skelton et al. [15-17] suggested a weighting with respect to the "component-costs". In a stochastic context, this may be undesirable, since it may lead to "bad surprises". Indeed, if the uncertainty associated with a dynamical element, with small expected cost contribution, is high, then the actual cost contribution for a sample trajectory of the stochastic system may be quite different from its (lower) expectation. This motivates the balancing with respect to the optimal deterministic controller, and the stochastic observer via the separation principle. ([18-19], [20])

The basics of the LQG-theory are well established, and can be found in many textbooks. The solution to the optimal control problem for a linear system with a quadratic performance index in the presence of white gaussian noise falls apart into the design of the deterministic controller (i.e. assuming perfect knowledge of the state of a system), and a stochastic observer for the noisy system driven by an external (but assumed known) input. This constitutes the celebrated Certainty-Equivalence Principle [6]. We shall briefly summarize the solution for this stochastic control problem. As was done in the open loop case, here also we shall try to give an interpretation to the solution, and clarify the different components in it. Several different problems are now of interest. In digital control (using fixed point arithmetic) the interest is in minimal sensitivity (with respect to the finite wordlength effects) design of the digital controller. In general control, one might be interested in a suboptimal but reduced order controller (in order to reduce the computational burden). Finally one might just be concerned with the modeling and analysis of the overall feedback system (thus including the plant) for the purpose of assessing the dominant contributions to the performance index, or uncertainty. In this case reduced order models for the combined plant and regulator are of interest.

It will be shown that the ideas of balanced realizations, when properly (re)defined are again very useful. The LQG-balancing for continuous time systems will be motivated from the following analysis.

3.1 The LQG-terminal controller

In order to fix the ideas, consider the stochastic system

$$\dot{x} = Fx + Gu + w \quad ; \quad \dim x = n, \dim u = m \quad (3.1)$$

$$y = Hx + v \quad ; \quad \dim y = p \quad (3.2)$$

with the initial state normally distributed:

$$x(t_0) \sim N(x_0, P_0) \quad (3.3)$$

For simplicity (but without loss of generality) we shall assume that w and v are uncorrelated zero mean white gaussian noise processes, with covariances Q and R respectively. They are further assumed to be independent from the initial conditions. Let the design objective be the minimization of a positive semi-definite quadratic performance index

$$J = E \left(x'(t_f) S_f x(t_f) + \int_{t_0}^{t_f} (x'Ax + u'Bu)dt \right) \quad (3.4)$$

It is assumed that the matrices R and B are positive definite, but otherwise arbitrary. In fact, this amounts to a slight overparametrization, but avoids some preliminary transformation. First, assuming that the states can be perfectly measured, the (deterministic) optimal closed loop system will have dynamics:

$$\dot{x} = (F - GC)x \quad x(t_0) = x \quad (3.5)$$

$$C = B^{-1}G'S \quad (3.6)$$

$$\dot{S} = -S(F-GC) - (F-GC)'S - A-C'BC \quad S(t_f) = S_f \quad (3.7)$$

The solution $S(t; t_f, S_f)$ of (3.7) has an interpretation as a weighting matrix for the minimum "cost-to-go" from the state $x(t)$ at time t . For $S_f = 0$, $S(t)$ is exactly the observability Gramian of the closed loop system sporting the fictitious $(n+m)$ -dimensional output

$$z = Lx \quad (3.8)$$

$$L' = [-CB^{1/2}, A^{1/2}] \quad (3.9)$$

The performance index then is the output energy (as discussed in section 2.2) of this system. The presence of the nonzero S_f can be interpreted as the instantaneous release ("flushing") of the remaining "energy" in the system (due to a nonzero state) at time t_f over a weight matrix S_f . Equivalently, it is also a measure for the amount of information, about an a priori unknown initial condition, that this fictitious output would carry, if corrupted by unit variance white gaussian noise.

Similarly, the filter error dynamics are given by

$$\dot{\tilde{x}} = (F - KH)\tilde{x} + M\omega; \dim \omega = n+p \quad (3.10)$$

$$M = [Q^{1/2}, KR^{1/2}] \quad (3.11)$$

where

$$\tilde{x} = x - \hat{x} \quad (3.12)$$

$$K = PH'R^{-1} \quad (3.13)$$

$$\dot{P} = (F-KH)P + P(F-KH)' + Q + KRK' \quad P(t_0) = P_0 \quad (3.14)$$

and $\omega(t)$ is a white noise of unit variance. Again, $P(t; t_0, P_0)$ is a measure for the uncertainty in the closed loop system (3.10), and characterizes the "disturbability" by the noise. It is in fact the covariance $E(\tilde{x}\tilde{x}')$ of the estimation error at time t , if at time t_0 the error was P_0 .

The Certainty Equivalence principle states that the optimal control for the system (3.1 - 3.2) where the states are not perfectly known, is given by feedback of the

estimates of the states over the optimal control-gains. The overall equations are thus

$$u = -Cx \quad (3.15)$$

$$\dot{\hat{x}} = F\hat{x} + Gu + K(y - H\hat{x}) ; \quad \hat{x}_0 = 0 \quad (3.16)$$

(The variance of the estimate $E(\hat{x}\hat{x}')$ will be denoted by Σ , while Π is used for the state variance $E(xx')$. By the optimality of the estimates, we have then $\Pi = \Sigma + P$. Note that the innovation $e = (y - H\hat{x})$ acts as a white Gaussian noise with covariance R . The following equivalent equations are easily derived.

$$\dot{\hat{x}} = (F - GC)\hat{x} + GC\bar{x} + w \quad (3.17)$$

$$\dot{\bar{x}} = (F - GC)\bar{x} + KH\bar{x} + Kv \quad (3.18)$$

$$\dot{\hat{x}} = (F - GC - KH)\hat{x} + Ky \quad (3.19)$$

The optional performance index (3.4) can now be evaluated in several different forms (using partial integration and the Riccati equations for S and P combined with the Lyapunov equations for Π and Σ in the closed loop)

$$J = \text{Tr} \{ \Pi_f S_f + \int_{t_0}^{t_f} (A\Pi + C'BC\Sigma) dt \} \quad (3.20)$$

$$= \text{Tr} \{ \Pi_0 S_0 + \int_{t_0}^{t_f} (SQ + C'BCP) dt \} \quad (3.21)$$

$$= \text{Tr} \{ P_f S_f + \int_{t_0}^{t_f} (AP + KRK'S) dt \} \quad (3.22)$$

Several (equivalent) interpretations follow from these equations. (3.19) and (3.15) give an open loop representation for the optimal stochastic controller, with input y and output the control u (figure 1). The equations (3.10) and (3.17) lead to a decomposition as a cascade of a system $(F-KH, M, I)$, driven by standard white gaussian noise, connected via a "transmission" matrix GC to the system $(F-GC, I, L)$, (figure 2). Whereas the former

subsystem represents the dynamics of the estimation error $\tilde{x}$ (driven by the fictitious noise ω for which $M\omega = w - Kv$), the latter will represent the plant states x , if and only if an additional gaussian input (w) is summed at its input. This additional noise has covariance Q , and is correlated with the noise ω according to $M E(\omega w') = Q$.

In terms of this decomposition we define a fictitious output $z = z_1 + z_2$ where z_1 is the $(n+m)$ -dimensional output from the second subsystem and $z_2 = \Gamma x$ where Γ is the $(n+m)$ by n matrix $\Gamma' = [C'B^{1/2}, 0]$ and L is as defined in (3.9). Because the outputs z_1 and z_2 are "maximally interfering" (due to their correlation), the variance of their sum z , is actually the difference of their individual variances, which is the integrand in (3.20).

Another interesting representation (figure 3) can be derived starting from the equations (3.10) and (3.18). Again a cascade is formed, beginning with the system $(F-KH, M, I)$ driven by standard white gaussian noise. This time the output (which is the representation of the state estimation error) drives, via the "transmission-matrix" KH , the system $(F-GC, I, L)$. This system will have the variable $\tilde{x}$ as state, if again an additional "correction" input Kv (with variance KRK') is added, having a correlation T with ω satisfying:

$$MTK' = E(M\omega)(Kv)' = -KRK.$$

A fictitious output $z = z_3 + z_4$ is defined which generates the integrand in the performance index. In this case, it is readily verified that this is accomplished by z_3 , the output of the cascade and $z_4 = N\tilde{x}$, where $N' = [0, A^{1/2}]$. Note that in this decomposition the input to the x -subsystem is actually K times the innovations process. This is known to act as a white noise. Here z_3 and z_4 are noninterfering (uncorrelated).

3.2 Interpretation of the Cost Functional.

The two decompositions described in the last paragraph, lead to a "cost-decoupled" interpretation of the various terms. For simplicity, we shall fix the ideas on the LQG-regulator problem. The matrices F , G , H , Q , and R are all supposed to be time-

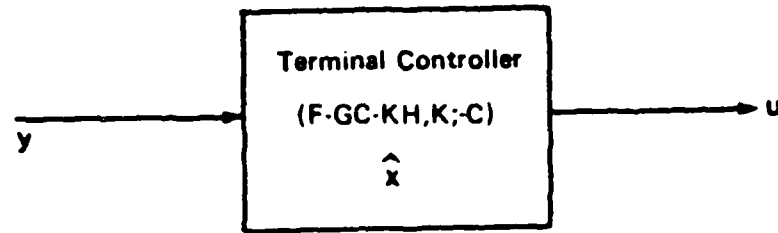


Figure 1: Open Loop representation of the optimal LQG system.

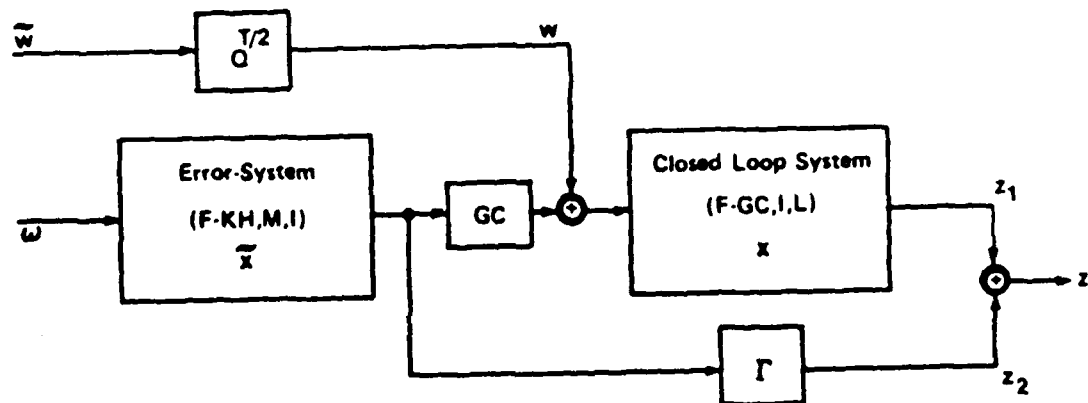


Figure 2: The $(\tilde{x}, x)$ representation of the optimal LQG system.

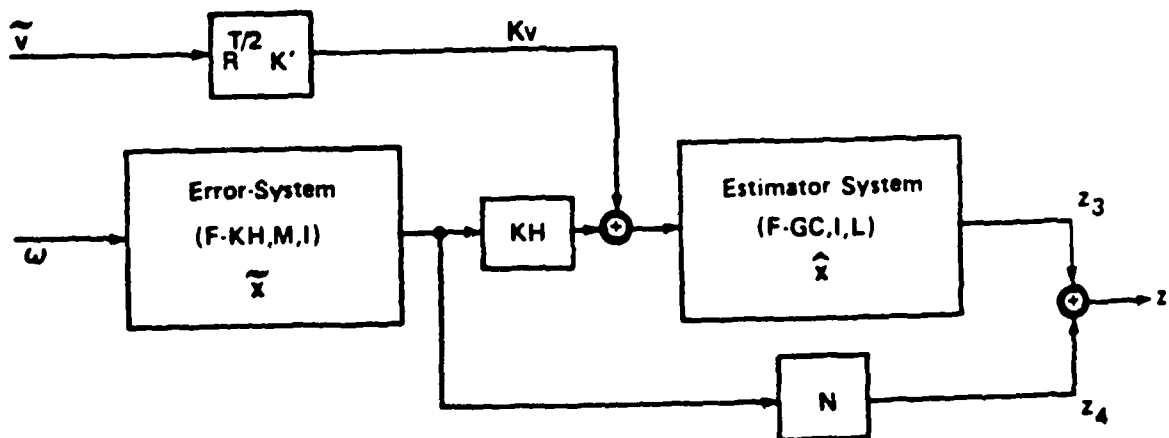


Figure 3: The $(\tilde{x}, \hat{x})$ representation of the optimal LQG system.

invariant, and it will be assumed that the statistical stationary state exists. For the steady-state regulator problem, the cost-rate, rather than the cost (which is infinite) is computed. Let in the LQG terminal controller, t_0 and t_f approach $-\infty$ and $+\infty$ respectively. Then the cost rate is the limit of the expected cost per time unit. It follows then from the equations (3.20 - 3.22) that the cost rate is

$$j = \text{Tr} \{A\Gamma + C'BC\Sigma\} = \text{Tr} \{SQ + C'BCP\} = \text{Tr} \{AP + KRK'S\} \quad (3.23)$$

where now P and S satisfy the algebraic Riccati equations, respectively

$$(F-KH)P + P(F-KH)' + MM' = 0 \quad (3.24)$$

$$(F-GC)'S + S(F-GC) + L'L = 0 \quad (3.25)$$

As already discussed, the following identifications follow, where the "*" entry is irrelevant:

P is the REACHABILITY gramian for the system $(M, F-KH, *)$

S is the OBSERVABILITY gramian of the system $(*, F-GC, L)$

Consider now a fictitious system, consisting of the two decoupled subsystems $(F-GC, Q^{1/2}, L)$ and $(F-KH, M, \Gamma)$. If both are driven by independent standard white gaussian noise, then their outputs (respectively z_1 and z_2) will be uncorrelated as well, and their expected "powers" additive. The contribution of the first is exactly $\text{Tr} \{Q \otimes (F-GC, L)\} = \text{Tr}\{QS\}$, while the contribution of the second is $\text{Tr}\{MM' \otimes (F-KH, \Gamma)\} = \text{Tr}\{\Gamma\Gamma' \otimes (F-KH, M)\} = \text{Tr}\{C'BCP\}$. We used the fact that in the steady state, the A -weighted output power of a system (F, G, H) driven by a zero mean white gaussian input of variance Q , can be expressed as

$$\text{Tr} H'AH \otimes_{Q-1} = \text{Tr} GQG' \otimes_A \quad (3.26)$$

But note that this is exactly the system of figure 2 if one replaced the input noises by uncorrelated ones, and set the transmission matrix GC to zero. The two partial outputs are then indeed noninterfering. This leads to the INTERPRETATION that $\text{Tr}\{QS\}$ is the partial cost rate in the deterministic closed loop system due to the process noise ω only, i.e., assuming full knowledge of the state x . The part $\text{Tr}\{C'BCP\}$ is the cost rate due to the state estimation error. It is as if the role of the transmission matrix GC and the input correlation is to guarantee maximal destructive interference between the outputs, (so that the variance of the output is the difference of the individual variances).

Similarly, with regards to the figure 3, we have the estimator subsystem $(F-GC, K, L)$ driven by the innovations ϵ , and with output z_3 , and the estimation-error system $(F-KH, M, N)$ driven by ω and with output z_4 . The two outputs are uncorrelated, and their sum $z = z_3 + z_4$ has covariance equal to the cost rate of the optimally controlled system. A cost-equivalent decoupled form can be constructed, consisting of the system $(F-GC, K, L)$ (the estimator system) driven by v , considered independent of the noise ω , which drives the error system $(M, F-KH, N)$. Their output "power" contributions are respectively $\text{Tr}\{KRK'S\}$ and $\text{Tr}\{AP\}$. Thus the role of the transmission matrix KH in figure 3 seems to be to effectively uncorrelate the two input noises.

The contribution $\text{Tr}\{KRK'S\}$ can thus be identified as the cost rate for the closed loop system under the assumption that the estimated state is the correct state. However since not $\hat{x}$, but x is the state of the closed loop system, a correction occurs due to the imperfect knowledge of the state (i.e. $\tilde{x}$). This is represented by the (independent) contribution of the error subsystem $(M, F-KH, N)$, driven by the equivalent noise ω , with cost-rate contribution $\text{Tr}\{AP\}$.

3.3 LQG-Balanced Realizations

Since S and P transform under a similarity transformation T as T^{-T} and ST^{-1} and TPT' respectively, it is possible to transform any given realization such that in the new coordinate system,

$$P = S = \Omega$$

where Ω is diagonal, with its elements ordered in magnitude. Ω is called the CANONICAL RICCATIAN. The new realization will then be referred to as the LQG-BALANCED realization. Note that Q as well as A also transform under the similarity. (B and R are of course invariant as they relate to "external" variables).

The cost rate for the optimally regulated system is then in the balanced coordinates,

$$j = \text{Tr } \Omega(Q + C'BC) = \text{Tr } \Omega(A + K RK') \quad (3.27)$$

or, using the fact that Ω is diagonal:

$$j = \sum_{i=1}^n \Omega_i (Q_{ii} + \Omega_{ii}^2 (GB^{-1}G')_{ii}) \quad (3.28)$$

$$j = \sum_{i=1}^n \Omega_i (A_{ii} + \Omega_{ii}^2 (HR^{-1}H')_{ii}) \quad (3.29)$$

It is clear that the cost rates corresponding to the individual state components are not simply determined by the magnitudes of the elements of the canonical Riccatian, but also depend on the relative magnitudes of the diagonal elements of Q , $GB^{-1}G'$, or A and $H'R^{-1}H$.

If the system (in balanced coordinates) is partitioned into two coupled subsystems with $\Omega_1 \geq \Omega_2$, then a sufficient condition for the part corresponding with Ω_1 to have the dominant cost rate contribution, is that either of the following sets of inequalities are satisfied: (the indices refer to the block-entries)

$$\begin{cases} \Omega_1 \geq \Omega_2 \\ Q_{11} \geq Q_{22} \\ (GB^{-1}G')_{11} \geq (GB^{-1}G')_{22} \end{cases} \quad (3.30)$$

$$\begin{cases} \Omega_1 \geq \Omega_2 \\ A_{11} \geq A_{22} \\ (H'R^{-1}H)_{11} \geq (H'R^{-1}H)_{22} \end{cases} \quad (3.31)$$

The various terms can be interpreted as quantifying the following:

- Q : Disturbance (noise) in the plant.
- $GB^{-1}G'$: "Potential" of the system input to decrease the regulation cost.
- A : Cost on the state deviations.
- $H'R^{-1}H$: Information (about the state) gained from the measurements (observations).

The first set of inequalities expresses that the set of variates that are most disturbed by noise and for which at the same time the input-potential is high, are dominant. (The larger the input-potential, the less the cost of control). Alternatively, state variables for which the information contained in the measurements and the state-cost is highest, also contribute to the major parts in the regulation cost-rate.

If one were only interested in obtaining a simple model for the optimally regulated system, for instance with the goal of identifying the dominant contributions to the uncertainties and the costs, then the combined plant and regulator may be reduced by "projection of dynamics". The decision on the order of the reduced model can for instance be based on thresholding the ratios

$$\alpha(r) = \frac{\text{Tr } \Omega_1}{\text{Tr } \Omega} \quad \text{and} \quad \beta(r) = \frac{j(r)}{j(n)}$$

where

$$j(r) = \sum_{i=1}^r \Omega_i (Q_{ii} + \Omega_i^2 (GB^{-1}G')_{ii})$$

If on the other hand one wants to design a reduced order regulator for a fixed plant, the above cannot be taken over directly. It has been shown that the design of a reduced order regulator based on a reduced order model of the plant may be unsatisfactory. Also, a "projection of dynamics" approach on the full order combined plant and regulator, based on the magnitude of the elements of the canonical Riccatian alone may not guarantee the stability of the regulation of the (full order) plant with the obtained reduced regulator ([16-17], [20]). Also in the open loop case, the Λ provides insufficient information ([14], [21]).

In reference to the decompositions in figures 2 and 3, the following property of the transmission matrix is derived.

Theorem The lower (upper) triangular part of the transmission matrix GC (KH) is dominant in the balanced coordinates.

proof: Since $T = GC = GB^{-1}G'\Omega$, it follows that ΩT is symmetric. But then $\Omega_i T_{ij} = T_{ji} \Omega_j$ for all i and j . Hence, $T_{ji} = T_{ij} \Omega_i / \Omega_j$. Since by assumption the elements Ω_i are ordered, we get $T_{ji} \geq T_{ij}$ whenever $j \geq i$. Similarly, $X = KH = \Omega H' R^{-1} H$ and thus $X\Omega$ is symmetric, from which $X_{ji} \leq X_{ij}$ whenever $j \geq i$.

Consider figure 2. Keeping the ordering in mind for the balanced case, it follows that low uncertainty states are more perturbed by the high uncertainty states than vice versa. It further follows from the positive semi-definiteness of the "input-potential" that for $j \geq i$

$$(GB^{-1}G')_{ij}^2 \leq (GB^{-1}G')_{ii} (GB^{-1}G')_{jj}$$

and thus that the elements in the upper left block of $(GB^{-1}G')\Omega$ are larger in magnitude than the elements in the upper right block of T . Hence, x_2 is almost decoupled from the closed loop system $(I, F-GC, L)$. In fact, for the same reason the upper right block of the closed loop system matrix $F-GC$ will be close to that of F itself, so that there is

almost no feedback from the x_2 subsystem. One expects, therefore, that the closed loop dynamics of the plant controlled with the reduced regulator, would have a near optimal behavior. Similar arguments work with figure 3.

4 STOCHASTIC MODELING

4.1 The Stochastic Realization Problem

Desai and Pal [22-23], extended the ideas of balancing in the LQG-sense to the stochastic realization problem. Balancing is here with respect to the state covariance matrices in the forward and the backward innovation representations. These matrices solve dual Riccati equations. The elements of the "canonical Riccatian" are connected to the canonical correlations between the past and the future observations. Arun and Kung [24] contrasted the method based on canonical correlations with a method based on Principal Components. Vaccaro showed its connection with deterministic open loop balancing [25]. Ramos and Verriest [26-27] unified the theory by showing that both the canonical correlation analysis (CCA) and the principal component analysis (PCA) are special cases of a more general optimization problem, using a new tool from multivariate statistics: the RV-coefficient introduced by Escouffier [28]. If two zero mean random vectors X_1 and X_2 (not necessarily of the same dimension) have covariance matrix

$$\text{cov}(X_1, X_2) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad (4.1)$$

then the RV-coefficient is defined as

$$\text{RV}(X, Y) = \frac{T_r(\Sigma_{12}\Sigma_{21})}{\sqrt{T_r(\Sigma_{11}^2) T_r(\Sigma_{22}^2)}} \geq 0 \quad (4.2)$$

This measure shares many of the properties of a correlation coefficient, but is not one itself. (It is the square of the correlation if X and Y are scalar). It also allows the

computation of a "figure of merit" for each algorithm in a consistent way.

This formalism is applied to stochastic realization theory as follows. Given the correlation sequence $\{\Lambda_k\}$ of a discrete time stationary stochastic sequence $\{y_k\}$, the forward and backward predictor subspaces are

$$\begin{aligned} X_k &= \text{Span}(Y_k^+ | Y_k^-) \\ Z_{k-1} &= \text{Span}(Y_k^- | Y_k^+) \end{aligned} \quad (4.3)$$

where Y^+ and Y^- respectively correspond to the "future" and the "past" of the process. Here $(A|B)$ denotes the projection of $\text{span}(A)$ onto $\text{span}(B)$. These two spaces form the information interface between the past and the future. Defining as usually

$$\hat{H}_k = E\{Y_k^+(Y_k^-)'\}, R_k^+ = E\{Y_k^+(Y_k^+)'\}, R_k^- = E\{Y_k^-(Y_k^-)'\} \quad (4.4)$$

then CCA is equivalent to the problem of finding transformations L and M such that $RV(L'Y^+, M'Y^+)$ is maximized, subject to the constraints that $L'(R^+)L$ and $M'(R^-)M$ are diagonal. The PCA is equivalent to the problem of maximizing $RV(Y^+, M'Y^-)$ over M under the constraint $M'(R^-)M$ diagonal. The two methods are also referred to as the one-sided and the two-sided stochastic realization problem.

4.2 Geometrical Interpretation: Correlation between Subspaces

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. The lattice of all closed subspaces of H and the lattice $\text{Proj } H$ of all orthoprojectors on H are isomorphic.

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.,

$$\sum_i \mu(S_i) = \mu(\sum_i S_i) \quad (4.5)$$

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

$$\mu_v(A) = \|P^A(v)\|^2 \quad (4.6)$$

where P^A is the projection operation on A . Clearly, finite convex combinations 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

$$\mu(A) = \text{Tr}(TP^A) \quad (4.7)$$

Gleason [29] 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.

Consider now a tensor product Hilbert space $R^P \otimes H$, and let $\{\Psi_i\}$ be a complete Orthonormal Set (CONS) in H . Any vector x in this tensor product space has then a decomposition

$$x = \sum_i |x_i\rangle \langle \Psi_i| ; x_i \in R^P \text{ for all } i \quad (4.8)$$

The vector x will be referred to as a "prior". Let for all A in $\text{Proj } R^P$: $\mu_i(A) = \|P^A(x_i)\|^2$ and define a "superposition of measures" on $\text{Proj } R^P$ as $\mu_x = \sum_i \mu_i \alpha_i$ for some

square summable positive weights $\{\alpha_i\}$. By Gleason's theorem [29] it follows that there exists an operator $T_x : R^P \rightarrow R^P$ such that

$$\mu_x(A) = \text{Tr } T_x P_A \quad (4.9)$$

This operator is characteristic for the given vector x in $R^P \oplus H$ (in fact, a "sufficient statistic"), and one can think of T (or μ) as "conditioned" by the vector x . Since

$$T_x = \sum_i \alpha_i^2 |x_i\rangle \langle x_i| = xx' \quad (4.10)$$

it can be interpreted as a Gramian or covariance operator.

The measure $\mu_x(A)$ gives a numeric value to the closeness of A to R^P , given the prior x [30]. Define also the extended projectors $P^B \in \text{Proj } (R^P \oplus H)$ by

$$P^B(x) = \sum_i \alpha_i P^B |x_i\rangle \langle x_i| \quad (4.11)$$

They allow now the definition of the "variance" and "covariance" of subspaces in $\text{Proj } R^P$, [31]. The "posterior" variance of $A \in \text{Proj } R^P$ given x is then the operator from R^P to R^P

$$(P^A_x) (P^A_x)' = \sum_i P^A |x_i\rangle \langle x_i| P^A = P^A T_x P^A \quad (4.12)$$

and the covariance

$$(P^B_x) (P^A_x)' = \sum_i P^B |x_i\rangle \langle x_i| P^A = P^B T_x P^A \quad (4.13)$$

This is simply interpreted as the restriction to B of the mapping T_x restricted to the subspace A , and displays the coupling or interface between A and B given x . In order to

REFERENCES

- [1] Caines, P. E., "On the Scientific Method and the Foundation of System Identification," *Int'l Symp. Math. Thy. of Networks and Systems*, Stockholm, Sweden, June 1985, (Proceedings in print).
- [2] Rissanen, J., "A Universal Prior for Integers and Estimation by Minimum Description Length," *Ann. of Stat. (1983)*, Vol. 11, 2, 416-431.
- [3] Larimore, W. E., "Predictive Inference, Sufficiency, Entropy and an Asymptotic Likelihood Principle," *Biometrika (1983)*, Vol. 70, 1, 175-181.
- [4] Kailath, T., *Linear Systems*, Prentice Hall, 1980.
- [5] Moore, B. C., "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Trans. Automatic Control*, AC-26, No. 5 (1) 17-32, January 1981.
- [6] Bryson, A. E., and Ho, Y. C., *Applied Optimal Control*, Ginn and Comp., 1969.
- [7] Verriest, E. I. and Kailath, T., "On Generalized Balanced Realizations," *IEEE Trans. Automatic Control*, AC-28, no. 8, pp. 833-844, 1983.
- [8] Mullis, C. T., and Roberts, R. A., "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits Syst.*, Vol. CAS-23, No. 9, pp. 551-562, 1976.
- [9] Verriest, E. I., "The Structure of Multivariable Balanced Realizations," *Proc. IEEE International Symposium Circuits Syst. 1983*, Newport Beach, CA, pp. 110-113.
- [10] Sveinsson, J. R., and Fairman, F. W., "Minimal Balanced Realizations of Transfer Function Matrices using Markov Parameters," *IEEE Trans. A.C.*, Vol. AC-30, No. 10, pp. 1014-1016, October 1985.
- [11] Fairman, F.W., Mahil, S. S., and De Abreu, J. A., "Balanced Realization Algorithm for Scalar Continuous-Time Systems having Simple Poles," *Int. J. Systems Sci.*, Vol. 15, No. 1, pp. 685-694, 1984.
- [12] Kabamba, P. T., "Balanced Forms: Canonicity and Parametrization", *IEEE Trans. A.C.*, Vol. AC-30, No. 11, pp. 1106-1109, November 1985.
- [13] Young, N. J., "Balanced Realizations via Model Operators," *Int. J. Control*, Vol. 42, No. 2, pp. 369-389, 1985.
- [14] Kabamba, P. T., "Balanced Gains and Their Significance for L^2 Model Reduction," *IEEE Trans. A.C.*, Vol. AC-30, No. 7, pp. 690-693, July 1985.
- [15] Skelton, R. E. and Yousuff, A., "Component Cost Analysis of Large Scale Systems," *Int. J. Control*, Vol. 37, No. 2, pp. 285-304, 1983.
- [16] Yousuff, A., and Skelton, R. E., "A Note on Balanced Controller Reduction," *IEEE Trans. A.C.*, Vol. AC-29, No. 3, pp. 254-257, March 1984.
- [17] Yousuff, A., and Skelton, R.E., "Controller Reduction by Component Cost Analysis," *IEEE Trans. A.C.*, Vol. AC-29, No. 6, pp. 520-530, June 1984.
- [18] Verriest, E. I., "Low Sensitivity Design and Optimal Order Reduction for LQG - Problem," *Proc. 24th Midwest Symp. Circ. Syst.*, Albuquerque, New Mexico, pp. 365-369, (June 1981).
- [19] Verriest, E. I., "Suboptimal LQG-Design via Balanced Realizations," *Proc. 20th IEEE Conf. Dec. Control*, San Diego, CA, pp. 686-687, December 1981.
- [20] Jonckheere, E. A., and Silverman, L. M., "A New Set of Invariants for Linear Systems Applications to Reduced Order Compensation Design," *IEEE Trans. A.C.*, pp. 953-964, October 1983.
- [21] Therapos, C. P., "On the Selection of the Reduced Order via Balanced State Representations," *IEEE Trans. A.C.*, Vol. AC-29, No. 11, pp. 1019-1021, November 1984.
- [22] Desai, U. B. and Pal, D., "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations," *Proc. 21st IEEE Conf. on Decision and Control*, pp. 1105-1111, 1982.
- [23] Desai, U. D., and Pal, D. A., "A Transformation Approach to Stochastic Model Reduction," *IEEE Trans. A.C.*, Vol. AC-29, No. 12, pp. 1097-1100, December 1984.
- [24] Vaccaro, R. J., "Deterministic Balancing and Stochastic Model Reduction," *IEEE Trans. A.C.*, Vol. AC-30, No. 9, pp. 921-923, September 1985.

- [25] Arun, K. S. and Kung, S. Y., "A New SVD Based Algorithm for ARMA Spectral Estimation," *Proc. 1983 IEEE-ASSP Spectral Estimation Workshop*, Tampa, Florida, November 1983.
- [26] Ramos, J. A., and Verriest, E. I., "A Unifying Tool for Comparing Stochastic Realization Algorithms and Model Reduction Techniques," *Proc. 1984 Automatic Control Conf.*, San Diego, CA.
- [27] Ramos, J. A., "A Stochastic Approach to Streamflow Modeling," *Ph.D. Dissertation, School of Civil Engineering*, Georgia Institute of Technology, 1985.
- [28] Escoufier, Y., "Le Traitement des Variables Vectorielles," *Biometrics* 29, 751-760, (1973).
- [29] Gleason, A. M., "Measures on the Closed Subspaces of a Hilbert Space," *J. Math. Mech.* 6, pp. 885-893, 1957.
- [30] Verriest, E. I., "Model Reduction via Projection Methods," *Proc. 1985 Conf. Math. Thy. in Networks Syst.*, (A. Lindquist and C. Byrnes, editors) Amsterdam, North Holland, 1986.
- [31] Verriest, E. I., "A Unified Theory of Model Reduction via Gleason Measures," *Mathematics in Signal Processing*, (T. S. Durani, editor), Oxford University Press, (1985).
- [32] Verriest, E. I., "Reachability, Observability and Discretization," *Proc. 22nd IEEE Conf. Dec. Control*, San Antonio, Texas, pp. 854-855, December 1983.
- [33] Verriest, E. I., "Digital Filter Design Based on a High Fidelity Discretization Procedure," *Proc. Conf. Inf. Sci. and Syst.*, Johns Hopkins, M.D., March 1983.
- [34] Hsu, C. S., Desai, U. B. and Crawley, C. A., "Realization Algorithms and Approximation Methods of Bilinear Systems," *Proc. 22nd IEEE Conf. Dec. Control*, (Dec. 1983), pp. 783-788.
- [35] Verriest, E. I., "Approximations and Order Reduction in Nonlinear Models Using an RKHS Approach," *Proc. Conf. Inf. Sci. and Systems*, Princeton, NJ, pp. 197-201, March 1984.
- [36] Baram, Y., "A Geometric Approach to Stochastic Model Reduction by Canonical Variables," *IEEE Trans. A.C.*, Vol. AC-29, No. 4, pp. 358-359, April 1984.
- [37] Harshavardhana, P., Johckheere, E. A., "Spectral Factor Reduction by Phase Matching: The Continuous-Time Single-Input Single-Output Case," *Int. J. Control*, Vol. 42, No. 1, pp. 43-63, 1985.
- [38] Bacon, B. J., and Frazho, A. E., "A Hankel Matrix Approach to Stochastic Model Reduction," *IEEE Trans. A. C.*, Vol. AC-30, No. 11, pp. 1138-1140, November 1985.

APPENDIX K

REDUCED ORDER LQG DESIGN: CONDITIONS FOR FEASIBILITY

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

Abstract

This paper starts with a discussion on the Projection of Dynamics--principle as a model reduction tool. This is exploited in modal approximations and reductions based on open loop balancing. Similar ideas can be employed for the reduction of the closed loop, terminal controller problem, in the presence of noise, leading to the LQG balanced realizations. An interpretation of the various terms in the optimal performance index leads to the motivation for this reduction technique. It is shown that several "natural" restrictions occur, constraining the faithfulness of this reduction technique. As in the open loop case, it illustrates the insufficiency of the canonical Riccati towards the determination of the reduced order models.

1. Introduction

The theory of balanced realizations, introduced by Moore [5], has proven to be very useful in the model reduction problem for linear systems. The method relies on a transformation of the realization of a given system to the balanced coordinates. In this basis, the reachability and observability gramian of the system are equal and diagonal. These canonical elements (also termed second order modes) can be interpreted in terms of energies or uncertainties [10] and display the relative importance of the (balanced) state components with respect to the external system behavior, as the "joint degree of reachability and observability" associated with this system dimension. The reduced model is then obtained by deleting the system dimensions corresponding to the smaller canonical elements, i.e., Projection of Dynamics. It has been shown by Kambhampati [3] that the above sketched procedure does not yield faithful reduced order models in some cases. The reason hereto is that the canonical gramian only has n "free" parameters, whereas the n -th order model requires $2n$ parameters (we shall restrict our discussion to the Single Input Single Output case). Hence, the decision on the reduced order model is only based on half of the parameter set.

The open loop balancing methods have been extended to the closed loop LQG problem by this author [8,9] and Jonkheere and Silverman [2] for the LQG regulator. The performance of the LQG reduced system was also shown to depend on other variables, not directly related to the elements of the "canonical Riccatian."

This paper will motivate the use of and determine a "complete" set of parameters, which suffice as an information base to determine reduced order models. The more general terminal controller problem will be discussed. The LQG regulator problem will then follow as a special case (infinite final time, and time invariant parameters of system, noise, and performance index). Section 2 describes briefly the method of Projection of Dynamics. In Section 3, the LQG balancing is discussed, and an interpretation of the various

terms in the optimal performance index is given. This then leads to a set of criteria to determine the reduced order models.

2. Projection of Dynamics

By Projection of Dynamics (POD) of a system (A, B, C) , we understand the reduction of the given system (A, B, C) to a smaller system, which has as its dynamics the projection of the original dynamics of the system, confined to the retained state variables. Without loss of generality, one can thus assume that the POD is acting on the $[1,1]$ block system obtained from partitioning the given system into

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

thus yielding the subsystem (A_{11}, B_1, C_1) .

The POD technique arises, for instance, in the model reduction techniques based on modal approximation, the above introduced (open loop) balancing techniques, and the reduction via the Hessenberg form. In each of these cases, the given system is first transformed to a special realization (the diagonal or Gilbert realization in the first case, the balanced realization in the second, and the Hessenberg in the latter). The technique is also the one used to obtain a minimal realization from the Kalman canonical decomposition of a system into its observable and reachable part, its observable and nonreachable part, its unobservable and reachable part, and finally, its nonreachable and unobservable part. Finally, in the design of reduced order observers, the technique surfaces in a restricted sense. Only the subdynamics of the A and B matrices are of interest in the latter case.

By itself, POD, without any further restrictions, is not a viable reduction technique by virtue of the following fact:

Theorem. Given any observable realization of order n , and $r < n$, then any arbitrary r -th order denominator polynomial can be obtained by POD, after a suitable similarity transformation of the full order system.

For instance, in the design of reduced order observers, one method makes use of the above fact by determining a similarity transformation such that the $n-1$ by $n-1$ $[1,1]$ -submatrix of A has arbitrary eigenvalues [4]. More general transformation than the one in [4] are needed to also be able to vary the $[n-2]$ -th order numerator polynomial.

3. Balancing in the LQG Sense

In the open loop case, the balanced realization led to a natural selection of reduced order models through the "projection of dynamics." Adopting this procedure for the design of a reduced order controller

is very dangerous, however, due to the feedback around the system. A more direct approach is needed, treating the closed loop as a whole. In fact, the degree of uncertainty (i.e., the noise covariances) and the performance index or cost-functional of the system should be taken into account for the selection of a reduced model. Skelton [7] suggested a weighting with respect to the "component costs." In a stochastic context, this may be undesirable, since it may lead to "bad surprises." Indeed, if the uncertainty associated with a dynamical element, with small expected cost contribution, is high, then the actual cost contribution for a sample trajectory of the stochastic system may be quite different from its (lower) expectation. This motivates the balancing with respect to the optimal deterministic controller, and the stochastic observer via the separation principle.

The basics of the LQG theory are well established, and can be found in many textbooks. The solution to the optimal control problem for a linear system with a quadratic performance index in the presence of white gaussian noise falls apart into the design of the deterministic controller (i.e., assuming perfect knowledge of the state of a system), and a stochastic observer for the noisy system driven by an external (but assumed known) input. This constitutes the celebrated Certainty-Equivalence Principle [1]. We shall briefly summarize the solution for this stochastic control problem. Also we shall try to give an interpretation to the solution, and clarify the different components in it. Several different problems are now of interest. In digital control (using fixed point arithmetic), the interest is in minimal sensitivity (with respect to the finite wordlength effects) design of the digital controller. In general control, one might be interested in a suboptimal but reduced order controller (in order to reduce the computational burden). Finally, one might just be concerned with the modeling and analysis of the overall feedback system (thus including the plant) for the purpose of assessing the dominant contributions to the performance index, or uncertainty. In this case, reduced order models for the combined plant and regulator are of interest. It will be shown that the ideas of balanced realizations, when properly (re)defined are again very useful. The LQG balancing for continuous time systems will be motivated from the following analysis.

3.1 The LQG Terminal Controller

In order to fix the ideas, consider the stochastic system

$$\dot{x} = Fx + Gu + w; \dim x = n, \dim u = m \quad (1)$$

$$y = Hx + v; \dim y = p \quad (2)$$

For simplicity (but without loss of generality), we shall assume that w and v are uncorrelated zero mean white gaussian noise processes, with covariances Q and R , respectively. They are further assumed to be independent from the initial conditions. Let the design objective be the minimization of a positive semidefinite quadratic performance index.

$$J = E\{x'(t_f)S_f x(t_f) + \int_{t_0}^{t_f} (x'Ax + u'Bu)dt\} \quad (4)$$

It is assumed that the matrices R and B are positive definite, but otherwise arbitrary. In fact, this amounts to a slight overparameterization, but avoids some preliminary transformations. First, assuming that the states can be perfectly measured, the (deterministic) optimal closed loop system will have dynamics:

$$\dot{\tilde{x}} = (F-GC)\tilde{x}; \tilde{x}(t_0) = x \quad (5)$$

$$C = B^{-1}G'S \quad (6)$$

$$\dot{S} = -S(F-GC) - (F-GC)'S - A - C'BC; S(t_f) = S_f \quad (7)$$

The solution $S(t; t_f, S_f)$ of (7) has an interpretation as a weighting matrix for the minimum "cost to go" from the state $x(t)$ at time t . For $S_f = 0$, $S(t)$ is exactly the observability gramian of the closed loop system sporting the fictitious $(n+m)$ -dimensional output

$$z = Lx \quad (8)$$

$$L = [-C'B^{1/2}, A^{1/2}] \quad (9)$$

The performance index then is the output energy of this system. The presence of the nonzero S_f can be interpreted as the instantaneous release ("flushing") of the remaining "energy" in the system (due to a nonzero state) at time t_f over a weight matrix S_f . Equivalently, it is also a measure for the amount of information, about an a priori unknown initial condition, that this fictitious output would carry, if corrupted by unit variance white gaussian noise. Similarly, the filter error dynamics are given by

$$\dot{\tilde{x}} = (F-KH)\tilde{x} + Mw; \dim w = n+p \quad (10)$$

$$M = [Q^{1/2}, KR^{1/2}] \quad (11)$$

where

$$\tilde{x} = x - \hat{x} \quad (12)$$

$$K = PH'R^{-1} \quad (13)$$

$$\dot{P} = (F-KH)P + P(F-KH)' + Q + KRK'; P(t_0) = P_0 \quad (14)$$

and $w(t)$ is a white noise of unit variance. Again, $P(t; t_0, P_0)$ is a measure for the uncertainty in the closed loop system (10), and characterizes the "disturbability" by the noise. It is, in fact, the covariance $E(\tilde{x}\tilde{x}')$ of the estimation error at time t , if at time t_0 the error was P_0 .

The Certainty Equivalence principle states that the optimal control for the system (1)-(2), where the states are now perfectly known, is given by feedback of the estimates of the states over the optimal control gains. The overall equations are thus

$$u = -C\hat{x} \quad (15)$$

$$\dot{\hat{x}} = F\hat{x} + Gu + K(y-H\hat{x}); \hat{x}_0 = 0 \quad (16)$$

(The variance of the estimate $E(\hat{x}\hat{x}')$ will be denoted by Π , while Π is used for the state variance $E(\tilde{x}\tilde{x}')$. By the optimality of the estimates, we have then $\Pi = E + P$. Note that the innovation $\epsilon = (y-H\hat{x})$ acts as a white gaussian noise with covariance R . The following equivalent equations are easily derived.

$$\dot{\tilde{x}} = (F-GC)\tilde{x} + C\tilde{x} + w \quad (17)$$

$$\dot{\hat{x}} = (F-GC)\hat{x} + K\tilde{x} + Kv \quad (18)$$

$$\dot{\tilde{x}} = (F-GC-KH)\tilde{x} + Ky \quad (19)$$

The optimal performance index (4) can now be evaluated in several different forms (using partial integration and the Riccati equations for S and P combined with the Lyapunov equations for Π and I in the closed loop)

$$J = \text{Tr}\{\Pi_f S_f + \int_{t_0}^{t_f} (A\Pi + C'BCI)dt\} \quad (20)$$

$$= \text{Tr} \left[\Pi_0 S_0 + \int_{t_0}^{t_f} (SQ + C'BCP) dt \right] \quad (21)$$

$$= \text{Tr} \left[P_0 S_0 + \int_{t_0}^{t_f} (AP + KKK'S) dt \right] \quad (22)$$

Several (equivalent) interpretations follow from these equations. Equations (19) and (15) give an open loop representation for the optimal stochastic controller, with input y and output the control u (Fig. 1). The Eqs. (10) and (17) lead to a decomposition as a cascade of a system $(P-KH, M, I)$, driven by a standard white gaussian noise, connected via a "transmission" matrix GC to the system $(P-GC, I, L)$ (Fig. 2). Whereas the former subsystem represents the dynamics of the estimation error x (driven by the fictitious noise w for which $Mw = w-Kv$), the latter will represent the plant states x , if and only if an additional gaussian input (w) is summed at its input. This additional noise has covariance Q , and is correlated with the noise w according to $M^T(ww') = Q$. This interpretation will be referred to as the error-driven closed loop decomposition.

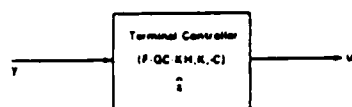


Figure 1 The Closed Loop System.

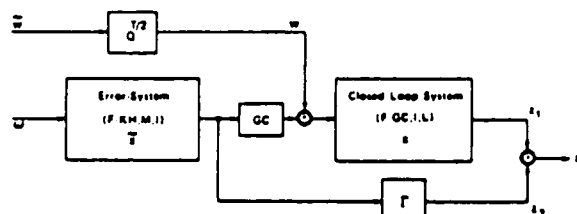


Figure 2 The Error-Driven Closed Loop Decomposition.

In terms of this decomposition, we define a fictitious output $z = z_1 + z_2$ where z_1 is the $(n+m)$ -dimensional output from the second subsystem and $z_2 = \Gamma x$ where Γ is the $(n+m)$ by n matrix $\Gamma' = [C'B^{1/2}, 0]$ and L is as defined in (9). Because the outputs z_1 and z_2 are "maximally interfering" (due to their correlation), the variance of their sum z , is actually the difference of their individual variances, which is the integrand in (20).

Another interesting representation (Fig. 3) can be derived starting from the Eqs. (10) and (18). Again a cascade is formed, beginning with the system $(P-KH, M, I)$ driven by standard white gaussian noise. This time the output (which is the representation of the state estimation error) drives, via the "transmission-matrix" KH , the system $(P-GC, I, L)$. This system will have the variable x as state, if again an additional "correction" input Kv (with variance KKK') is added, having a correlation T with w satisfying:

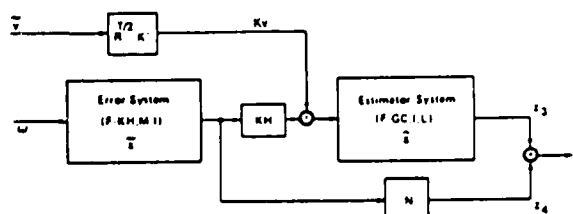


Figure 3 The Error-Driven Estimator Decomposition.

$$MTK' = E(Mw)(Kv)' = -KKK'$$

A fictitious output $z = z_1 + z_2$ is defined which generates the integrand in the performance index. In this case, it is readily verified that this is accomplished by z_1 , the output of the cascade and $z_2 = Nx$, where $N' = [0, A^{1/2}]$. Note that in this decomposition, the input to the x -subsystem is actually K times the innovations process. This is known to act as a white noise. Here z_1 and z_2 are noninterfering (uncorrelated). This decomposition will be called the error-driven estimator decomposition.

3.2 Interpretation of the Cost Functional

The two decompositions described in the last paragraph lead to a "cost decoupled" interpretation of the various terms. For simplicity, we shall fix the ideas on the LQG-regulator problem. The matrices P , G , H , Q , and R are all supposed to be time-invariant, and it will be assumed that the statistical stationary state exists. For the steady state regulator problem, the cost rate, rather than the cost (which is infinite), is computed. Let in the LQG terminal controller, t_0 and t_f approach $-\infty$ and $+\infty$, respectively. Then the cost rate is the limit of the expected cost per time unit. It follows then from the Eqs. (20)-(22) that the cost rate is

$$j = \text{Tr} \{ A\ddot{A} + C'BCI \} = \text{Tr} \{ SQ + C'BCP \} = \text{Tr} \{ AP + KKK'S \} \quad (23)$$

where now P and S satisfy the algebraic Riccati equations, respectively

$$(P-KH)P + P(P-KH)' + MM' = 0 \quad (24)$$

$$(P-GC)'S + S(P-GC) + L'L = 0 \quad (25)$$

As already discussed, the following identifications follow, where the "*" entry is irrelevant:

P is the REACHABILITY gramian for the system $(M, P-KH, *)$

S is the OBSERVABILITY gramian of the system $(*, P-GC, L)$

Consider now a fictitious system, consisting of the two decoupled systems $(P-GC, Q^{1/2}, L)$ and $(P-KH, M, \Gamma)$. If both are driven by independent standard white gaussian noise, then their outputs (respectively z_1 and z_2) will be uncorrelated as well, and their expected "powers" additive. The contribution of the first is exactly $\text{Tr} \{ Q\ddot{A}P-GC, L \} = \text{Tr} \{ QS \}$, while the contribution of the second is $\text{Tr} \{ MM'\ddot{O}(P-KH, \Gamma) \} = \text{Tr} \{ \Gamma\Gamma'(P-KH, M) \} = \text{Tr} \{ C'BCP \}$. We used the fact that in the steady state, the A -weighted output power of a system (P, G, H) driven by a zero mean white gaussian input of variance Q , can be expressed as

$$\text{Tr } H'AH Q_{-1} = \text{Tr } QGQ'G' A \quad (26)$$

where

$$Q_{-1} = \int_0^\infty e^{A\tau} G M^{-1} G' e^{A'\tau} d\tau$$

$$O_w = \int_0^\infty e^{A\tau} H' W H e^{A'\tau} d\tau$$

But note that this is exactly the configuration of the error-driven closed loop system decomposition, Fig. 2, if one replaced the input noises by uncorrelated ones, and set the transmission matrix GC to zero. The two partial outputs are then indeed noninterfering. This leads to the INTERPRETATION that $\text{Tr}\{QS\}$ is the partial cost rate in the deterministic closed loop system due to the process noise w only, i.e., assuming full knowledge of the state x . The part $\text{Tr}\{C'BCP\}$ is the cost rate due to the state estimation error. It is as if the role of the transmission matrix GC and the input correlation is to guarantee maximal destructive interference between the outputs (so that the variance of the output is the difference of the individual variances).

Similarly, with regards to the error-driven estimator decomposition, Fig. 3, we have the estimator subsystem ($F-GC, K, L$) driven by the innovations c , and with output z_3 , and the estimation error system ($F-KH, M, N$) driven by w and with output z_4 . The two outputs are uncorrelated, and their sum $z = z_3 + z_4$ has covariance equal to the cost rate of the optimality controlled system. A cost equivalent decoupled form can be constructed, consisting of the system ($F-GC, K, L$) (the estimator system) driven by v , considered independent of the noise w , which drives the error system ($M, F-KH, N$). Their output "power" contributions are respectively $\text{Tr}\{K'K'S\}$ and $\text{Tr}\{AP\}$. Thus the role of the transmission matrix KH seems to effectively uncorrelate the two input noises.

The contribution $\text{Tr}\{K'K'S\}$ can thus be identified as the cost rate for the closed loop system under the assumption that the estimated state were the correct state. However, since not x , but $\hat{x}$ is the state of the closed loop system, a correction occurs due to the imperfect knowledge of the state (i.e., $\tilde{x}$). This is represented by the (independent) contribution of the error subsystem ($M, F-KH, N$), driven by the equivalent noise w , with cost rate contribution $\text{Tr}\{AP\}$.

3.3 LQG Balanced Realizations

Since S and P transform under a similarity transformation T as $T^{-1}ST^{-1}$ and TPT' , respectively, it is possible to transform any given realization such that in the new coordinate system,

$$P = S = \Omega$$

where Ω is diagonal, with its elements ordered in magnitude. Ω is called the CANONICAL RICCATIAN. The new realization will then be referred to as the LQG BALANCED realization. Note that Q as well as A also transform under the similarity. (B and R are, of course, invariant as they relate to "external" variables.)

The cost rate for the optimally regulated system is then in the balanced coordinates,

$$j = \text{Tr } \Omega(Q+C'BC) = \text{Tr } \Omega(A+K'K') \quad (27)$$

or, using the fact that Ω is diagonal:

$$j = \sum_{i=1}^n \Omega_i (Q_{ii} + \Omega_{ii}^2 (GB^{-1}G')_{ii}) \quad (28)$$

$$j = \sum_{i=1}^n \Omega_i (A_{ii} + \Omega_{ii}^2 (H'R^{-1}H')_{ii}) \quad (29)$$

It is clear that the cost rates corresponding to the individual state components are not simply determined by the magnitudes of the elements of the canonical Riccatian, but also depend on the relative magnitudes of the diagonal elements of Q , $GB^{-1}G'$, or A and $H'R^{-1}H$.

If the system (in balanced coordinates) is partitioned into two coupled subsystems with $\Omega_1 > \Omega_2$, then a sufficient condition for the part corresponding with Ω_1 to have the dominant cost rate contribution, is that either of the following sets of inequalities are satisfied: (the indices refer to the block entries)

$$\left\{ \begin{array}{l} \Omega_1 > \Omega_2 \\ Q_{11} > Q_{22} \\ (GB^{-1}G')_{11} > (GB^{-1}G')_{22} \end{array} \right. \quad (30)$$

$$\left\{ \begin{array}{l} \Omega_1 > \Omega_2 \\ A_{11} > A_{22} \\ (H'R^{-1}H)_{11} > (H'R^{-1}H)_{22} \end{array} \right. \quad (31)$$

The various terms can be interpreted as quantifying the following:

- Q : Disturbance (noise) in the plant.
- $GB^{-1}G'$: "Potential" of the system input to decrease the regulation cost.
- A : Cost on the state deviations.
- $H'R^{-1}H$: Information (about the state) gained from the measurements (observations).

The first set of inequalities expresses that the set of variates that are most disturbed by noise and for which at the same time the input potential is high, are dominant. (The larger the input potential, the less the cost of control.) Alternatively, state variables for which the information contained in the measurements and the state cost is highest, also contribute to the major parts in the regulation cost rate.

If one were only interested in obtaining a simple model for the optimally regulated system, for instance with the goal of identifying the dominant contributions to the uncertainties and the costs, then the combined plant and regulator may be reduced by "projection of dynamics." The decision on the order of the reduced model can, for instance, be based on thresholding the ratios

$$\alpha(r) = \frac{\text{Tr } \Omega_1}{\text{Tr } \Omega}$$

$$\beta(r) = \frac{j(r)}{j(n)}$$

where

$$j(r) = \sum_{i=1}^r \Omega_i (Q_{ii} + \Omega_{ii}^2 (GB^{-1}G')_{ii})$$

If, on the other hand, one wants to design a reduced order regulator for a fixed plant, the above cannot be taken over directly. It has been shown that the design of a reduced order regulator based on a

reduced order model of the plant may be unsatisfactory. Also, a "projection of dynamics" approach on the full order combined plant and regulator, based on the magnitude of the elements of the canonical Riccattian alone may not guarantee the stability of the regulation of the (full order) plant with the obtained reduced regulator [2]. The problem is here, of course, the feedback around the system.

In reference to the decompositions in Figs. 2 and 3, the following property of the transmission matrix is derived.

Theorem. The lower (upper) triangular part of the transmission matrix $GC(KH)$ is dominant in the balanced coordinates.

Proof. Since $T = GC = GB^{-1}G'N$, it follows that NT is symmetric. But then $n_i T_{ij} = T_{ji} n_j$ for all i and j . Hence, $T_{ji} = T_{ij} n_j / n_i$. Since by assumption the elements n_i are ordered, we get $T_{ji} > T_{ij}$ whenever $j > i$.

Similarly, $X = KH = NH'R^{-1}B$ and thus XN is symmetric, from which $x_{ji} < x_{ij}$ whenever $j > i$.

Consider the error-driven closed loop decomposition. Keeping the ordering in mind for the balanced case, it follows that low uncertainty states are more perturbed by the high uncertainty states than vice versa. It further follows from the positive semi-definiteness of the "input-potential" that for $j > i$

$$(GB^{-1}G')_{ij}^2 < (GB^{-1}G')_{ii}(GB^{-1}G')_{jj}$$

and thus that the elements in the upper left block of $(GB^{-1}G')N$ are larger in magnitude than the elements in the upper right block of T . Hence, x_2 is almost decoupled from the closed loop system $(I, F-GC, L)$. In fact, for the same reason the upper right block of the closed loop system matrix $F-GC$ will be close to that of F itself, so that there is almost no feedback from the x_2 subsystem. One expects, therefore, that the closed loop dynamics of the plant controlled with the reduced regulator, would have a near optimal behavior. Similar arguments work with Fig. 3.

4. Conclusions

Through some physical insight in the various terms in the optimal performance index, it was shown that the canonical Riccattian alone does not give sufficient information for a faithful decision towards a reduced order model. Sufficient conditions involving additional parameters (30) or (31) have been given that enable such a faithful reduced model design. Some further generalizations (the terminal controller and follower problem for time varying stochastic systems) are presently under investigation.

References

1. A.E. Bryson, Y.C. Ho, Applied Optimal Control, Ginn and Comp., 1969.
2. E.A. Jonckheere, L.M. Silverman, "A New Set of Invariants for Linear Systems Applications to Reduced Order Compensation Design," IEEE Trans. A.C., pp. 953-964, October 1983.
3. P.T. Kabamba, "Balanced Forms: Canonicity and Parameterization," IEEE Trans. A.C., Vol. AC-30, No. 11, pp. 1106-1109, November 1985.
4. T. Kailath, Linear Systems, Prentice Hall, 1980.
5. B.C. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," IEEE Trans. A. C., Vol. AC-26, No. 1, pp. 17-32, Jan. 1981.
6. C.T. Mullis, R.A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," IEEE Trans. Circuits Syst., Vol. CAS-23, No. 9, pp. 551-562, 1976.
7. R.E. Skelton, A. Yousuff, "Component Cost Analysis of Large Scale Systems," Int. J. Control, Vol. 37, No. 2, pp. 285-304, 1983.
8. E.I. Verriest, "Low Sensitivity Design and Optimal Order Reduction for LQG-Problem," Proc. 24th Midwest Symp. Cir. Syst., Albuquerque, New Mexico, pp. 365-369, June 1981.
9. E.I. Verriest, "Suboptimal LQG-Design via Balanced Realizations," Proc. 20th IEEE Conf. Dec. Control, San Diego, California, pp. 686-687, December 1981.
10. E.I. Verriest, T. Kailath, "On Generalized Balanced Realizations," IEEE Trans. A.C., Vol. AC-28, No. 8, pp. 833-844, 1983.

APPENDIX L

ON REDEFINING THE OPTIMAL LEAST SQUARES FILTER UNDER FLOATING
POINT OPERATIONS

ON REDEFINING THE OPTIMAL LEAST SQUARES FILTER UNDER FLOATING POINT OPERATIONS

E. I. Verriest

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

1. INTRODUCTION

The linear least-squares filtering problem is revisited. The optimal solution is of course the celebrated Kalman Filter. However, the optimality of the solution requires that all computations are performed with infinite precision, which is of course, out of this world. The question we therefore pose is: "Given the wordlength constraints of the digital machine (which results in finite precision) what is the optimal solution to the least squares problem?" Preliminary results indicate that if arithmetic errors (roundoff or chopping) are modeled as additional noise sources, then the original problem chopping) are modeled as additional noise sources, then the original problem can be recast in an equivalent least squares problem, now assuming infinite machine precision, but with additional noise terms.

Under floating point operations, the arithmetic error depends on the magnitude of the operands, and hence the equivalent problem will turn into a nonlinear filtering problem. In order to simplify the mathematics the diffusion approximation is taken. Exact solution of the nonlinear least squares problem is computationally prohibitive since the entire distribution needs to be updated (via the Kolmogorov (Fokker-Planck) partial differential equation). Several approximations are proposed: the truncated second order filter, the Gaussian second order filter, and the method of the quasi-moments (Hermite functions). Since this nonlinear filter has a different structure than the one for which the equivalent noise model was derived, implementation of it on the digital machine will induce a second order type arithmetic error. This error is either assumed to be small and therefore neglected, (preliminary results proved that this was indeed the case under some circumstances) or one can incorporate the error again in an update to a new equivalent nonlinear problem, and keep iterating if needed. A fixed point theorem argument shows convergence to an "exact" nonlinear equivalent model.

This work should complement the recent work by Moroney [1]. There only fixed point arithmetic is considered. Previous work along these lines includes the work on estimation and control

with quantized measurements by Curry [2] (although infinite precision computations are assumed), and the work by Miller and Wrathall [3]. The floating point arithmetic modeling relies on the work of Knuth [4] and Vandergraft [5]. The nonlinear estimation methods used are standard (e.g. [6]).

The next section describes some preliminary results. A full exploration will be deferred to a later paper. The problem is set up as a least squares problem in discrete time. A stochastic model is given and the floating point constraints are to be taken into account. First an equivalent unconstrained model is set up in discrete time. The noise enters nonlinearly in this model. It is shown below that an elegant solution can be obtained by "embedding" this discrete model into a continuous one, in which the noise enters in an additive fashion. This step is what we refer to as the diffusion approximation. This continuous time model is then solved (exact solutions are possible in some cases) or approximated, and finally, the resulting continuous filter is discretized to yield the (sub)optimal discrete filter for the constrained problem. These three steps are discussed more specifically for the linear filter problem (Kalman filter) for the model

$$\begin{aligned}x_{k+1} &= \Phi x_k + \Gamma u_k \\ y_k &= H x_k + v_k\end{aligned}\quad (0)$$

2. EQUIVALENCE OF THE FLOATING POINT CONSTRAINED LINEAR PROBLEM AND THE UNCONSTRAINED NONLINEAR PROBLEM

For these preliminary results it will be assumed that all gains can be precomputed in infinite precision, so that only roundoff involving the estimator state update needs to be considered. If now a certain formula yields a certain vector x , then the actual computation will result in a finite precision approximation $Q[x]$ of it. Hence we can write $x = Q[x] + e$. Applying this to the optimization problem for the measurement update yields the "desired" update x_k from the a priori estimate x_{k-1} and prior covariance P_{k-1} . The actual covariance is $M + \psi_M$ where $M^{-1} = P_{k-1}^{-1} + H'R^{-1}H$ and ψ_M is the truncation error covariance. Similarly, the time update yields $x_{k+1} = \Phi x_k$ with actual covariance

$$P_{k+1} = \Phi(M_k + \psi_M)\Phi' + \Gamma\Gamma' + \psi_T \quad (1)$$

where ψ_M is again the truncation error covariance induced by the x_{k+1} computation. The overall a priori covariance update results in

$$P_{k+1} = \Phi P_k \Phi' + \Gamma\Gamma' + K_k (H P_k H' + R)^{-1} K_k' + \Phi \psi_M \Phi' + \psi_T \quad (2)$$

$$K_k = \Phi P_k H' (H P_k H' + R)^{-1} \quad (3)$$

But this is exactly the covariance one would have had for the unconstrained filter for the system:

$$x_{k+1} = \Phi x_k + \Gamma u_k + e_k \quad (4)$$

$$y_k = H x_k + v_k \quad (5)$$

where the covariance of e_k is $\psi = \Phi \psi_M \Phi' + \psi_T$. One major difference occurs, and that is that for the above model the estimate x and the error x will be conditionally independent. This is not the case in the original problem. Hence the equivalence is only with respect to the error covariance. So far the pdf of the truncation or rounding error has not been modeled. Although this error is obviously uniform in the least significant digit, it will be approximated by a Gaussian distribution (which is even exact for the second order truncated moment filter). The results in [5] for floating point operations can easily be extended to vector and matrix operations. It turns out to that the floating point errors in each component are proportional to that component. Hence the error in the update x_{k+1} can be written as

$$e = \hat{x} - Q[\hat{x}] = \gamma \text{diag}(\hat{x}) c \quad (6)$$

where c is a random vector with mean zero and covariance I and where γ is some predeterminable constant.

Assuming that $\hat{x}$ is sufficiently close to x , the equation (6) gets transformed into

$$e = \gamma \text{diag}(x) c \quad (7)$$

With the previous discussion this results then finally in the equivalent model

$$x_{k+1} = \Phi x_k + \Gamma u_k + \gamma \text{diag}(x_{k+1}) c_k \quad (8)$$

$$y_k = H x_k + v_k \quad (9)$$

The nonlinearity arises in the read-in matrix for the noise vector c_k . Equation (8) is further an implicit equation in x_{k+1} because of the appearance of this term on both sides. Rather than converting this in an explicit form (which destroys additivity of the noise), a diffusion approximation is described in the next section.

3. THE DIFFUSION APPROXIMATION

Using the backward difference (if Φ is non-singular), the discrete model is transformed into a continuous model. If Δ is the time step, then we set $k\Delta = t$. Under the assumption that Δ is sufficiently small, this leads to the Ito-differential equation

$$\begin{aligned} dx(t)dt &= \Phi[x(t)dt - \Delta dx(t)] + \Gamma du(t) \\ &+ \gamma \text{diag}(x(t))dc(t) \end{aligned}$$

or

$$\begin{aligned} dx(t) &= \frac{1}{\Delta} (I - \Phi^{-1})x(t)dt + \frac{\Phi^{-1}\Gamma}{\Delta} du(t) \\ &+ \frac{\gamma}{\Delta} \Phi^{-1} \text{diag}(x(t))dc(t) \end{aligned} \quad (10)$$

which is of the form

$$dx(t) = F(t)x(t)dt + G(x(t))dw(t) \quad (11)$$

where $w(t)$ is a standard Brownian motion with incremental covariance I , and

$$\begin{aligned} G(x(t))G'(x(t)) &= \\ &= \frac{1}{\Delta^2} [\Phi^{-1}\Gamma\Gamma'\Phi^{-1} + \gamma^2 \text{diag}^2(x(t))] \end{aligned} \quad (12)$$

If Φ were singular the forward difference can be approximated to an Ito-equation like (11). Aspects for the "continuization" of a discrete system in the deterministic case are developed in [7].

4. APPROXIMATE SOLUTIONS OF THE NONLINEAR PROBLEM

To demonstrate the feasibility of the proposed method, we generate several approximations for the case of a first order system.

$$dx(t) = f(t)x(t)dt + \sqrt{q_1 + q_2 x^2(t)} dw(t) \quad (13)$$

The truncated second order filter [6] gives the time updates

$$\hat{x}(t|t_{i-1}) = f(t)\hat{x}(t|t_{i-1}) \quad (14)$$

$$\begin{aligned} \hat{P}(t|t_{i-1}) &= 2f(t)P(t|t_{i-1}) \\ &+ (q_1 + q_2 \hat{x}^2(t|t_{i-1}) + q_2 P(t|t_{i-1})) \\ \hat{P}(t|t_{i-1}) &= (2f(t) + q_2)P(t|t_{i-1}) \\ &+ q_1 + q_2 \hat{x}^2(t|t_{i-1}) \end{aligned} \quad (15)$$

it is clear that the covariance decays slower due to the round off noise. In fact, stability of $f(t)$ is no longer sufficient to assure convergence! The measurement update formulas give

$$\hat{x}(t_i^+) = \hat{x}(t_i^-) + K(t_i)(y_i - h\hat{x}(t_i^-)) \quad (16)$$

$$P(t_i^+) = P(t_i^-) - K(t_i)hP(t_i^-) \quad (17)$$

$$R_c(t_i) = P(t_i^-)hR_c^{-1}(t_i) \quad (18)$$

$$R_e(t_i) = h^2 P(\bar{t}_i) + R(t_i) \quad (19)$$

These formulas are identical to the ones in the infinite precision assumption. One can also use the Gaussian second order filter. The measurement updates are again given by (16-19), but now the time update for the covariance has the additional term

$$\frac{3}{4} q_1 q_2 (q_1 + q_2) \bar{x}^2(t|t_{i-1})^{-3/2} P^2(t|t_{i-1})$$

in the right hand side. The estimate update is again as in (14). We finally remark that second order filters provide a performance generally superior to that of first order techniques (such as the extended Kalman filter), especially for small noise strengths [6].

Higher order moment filters can be generated as well, e.g. the cumulants truncation (of which the second order Gaussian is a special case). Since in fact we expect nearly Gaussian distributions, another good alternative is the conditional quasi-moment method. Here the unknown density is expanded in terms of the Hermite functions

$$\frac{H_i(x)}{\sqrt{i!}} f_G(x)$$

where f_G is the nominal Gaussian distribution.

Also the direct approximate solutions for the Fokker-Planck equation can be developed, e.g. discretization and model reduction via balanced realizations. Preliminary work on the feasibility of this scheme is in progress.

In all these cases the remaining step is to discretize the time update equations. The resulting discrete filter has the structure

$$\hat{x}(k+1) = \hat{\phi}x(k) + \hat{K}(k)(y_k - Hx(k))$$

but now $\hat{K}(k)$ differs from the Kalman gain, and is data dependent.

Exact solutions for an important class exist. Namely if there is no (real) process noise ($Q \equiv 0$), then the only noise in the equivalent model is the quantization noise, and the diffusion model leads then to bilinear stochastic differential equations. This will be the case for instance in all deterministic processing, or in systems involving pure Newtonian dynamics as for example in spacecraft.

A theory of filters for bilinear systems has been developed [8-11] and evolves around the theory of Lie-algebras. The important result is that if the underlying Lie algebra is solvable [12], the exact moments (and hence solutions) can be computed. These results are applicable to problems of satellite tracking and rigid body orientation.

Finally, we remark that we can build on existing work [13] for treating the combined LQG problem under floating point arithmetic.

5. CONCLUSION

A novel solution on approximation to the least squares filter problem under floating point arithmetic is presented for a linear stochastic model.

To answer the question where the benefit of this study will be, it is perhaps easiest to state that if the stochastic model has low order and possesses slow time constants, and if a general purpose computer is available with large wordlength, then the finite wordlength effects are going to be negligible and there will be no benefit from this study. On the contrary, if one deals with microprocessor control of large scale systems and/or systems with multiple time scales, (singularly perturbed systems) then potential benefit will be gained from a deeper study of the optimality. The resulting optimal filter may turn out to be nonlinear, but this does not necessarily increase the complexity significantly (e.g. multiplication versus addition).

The more detailed study, under progress, considers the floating point constraints of the gain sequence computation as well. The latter is obviously data dependent. The information filter form is believed to yield the best approach.

V. REFERENCES

1. P. Moroney, Issues in the Implementation of Digital Feedback Compensators, MIT Press, 1983.
2. R. E. Curry, Estimation and Control with Quantized Measurements, MIT Press, 1970.
3. W. Miller and C. Wrathall, Software for Roundoff Analysis of Matrix Algorithms, Academic Press, 1980.
4. D. C. Knuth, The Art of Computer Programming, Vol. 2: Seminumerical Algorithms, Addison-Wesley, 1969.
5. J. S. Vandergraft, Introduction to Numerical Computations, Academic Press, 1983.
6. P. S. Maybeck, Stochastic Models, Estimation and Control, Vol. 2, Academic Press, 1982.
7. E. I. Verriest, "The matrix logarithm and the continuization of a discrete process," submitted for publication.
8. R. W. Brockett, "Nonlinear systems and differential geometry," Proc. IEEE, Vol. 64, no. 1, pp. 61-72, January 1976.

9. G. L. Wise and S. I. Marcus, "Stochastic stability for a class of systems with multiplicative state noise," IEEE Trans. Automatic Control, vol. AC-24, no. 2, pp. 333-337, April 1979.
10. S. I. Marcus, A. S. Willsky and K. Hsu, "The use of harmonic analysis in suboptimal estimator design," IEEE Trans. Automatic Control, pp. 911-916, October 1978.
11. S. D. Chikte and J. T.-H. Lo, "Optimal filters for bilinear systems with nilpotent lie algebras," IEEE Trans. Automatic Control, Vol. 24, no. 6, pp. 948-953, December 1979.
12. R. Hermann, Lie Groups for Physicists, B. Cummings Publishing Co., 1966.
13. A. Bagghi and T. Schilperoort, "Optimal linear stochastic control for systems with multiplicative noise," IEEE Trans. Automatic Control, vol. AC-25, pp. 1005-1007, October 1980.

APPENDIX M

ERROR ANALYSIS OF LINEAR RECURSIONS IN FLOATING POINT

ERROR ANALYSIS OF LINEAR RECURSIONS IN FLOATING POINT

Erik I. Verriest

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

ABSTRACT

A stochastic error model for floating point arithmetic is developed and used to study the effects of finite wordlength on linear recursive formulas. Optimal realizations exist, but they are highly sensitive.

1. INTRODUCTION

The accumulation and roundoff error in long computerized calculations and recursive algorithms is a phenomenon that can destroy an efficient and sound computational procedure based on arithmetic over the real number field. A tell-tale example is the Kalman filter divergence. Analytically this state estimator yields the estimate with minimal covariance, but the actual error covariance may be much larger than the predicted covariance (which solves a certain Riccati equation) or even grow unboundedly whenever a finite precision version of the algorithm is implemented.

It is clear that in order to keep track of the confidence in the computed results, a certain measure of confidence should be computed and "tracked" with each operation or update.

One such measure is the rigorous computation of error bounds via interval analysis. Not only are the additional computations that are required time consuming, but the obtained results may be overly conservative. The individual rounding errors in a compounded expression indeed tend to cancel rather than to reinforce each other if an unbiased rounding rule is used. In this paper another measure of confidence is used. It is of a more probabilistic nature and estimates the propagated covariance due to the finite wordlength errors. Such a study is standard (and straightforward) for fixed point arithmetic, and is well described in several textbooks, culminating in the optimal filter implementations by Mullis and Roberts [8] and the LQG compensator by Moroney [7]. One of the serious shortcomings of the use of fixed point arithmetic is the necessity of scaling in order to provide a higher accuracy.

This work was supported by the U.S. Air Force under Contract F-08635-84-C-0273.

This disadvantage is obliterated when floating point arithmetic is used.

Modern digital technology has rapidly increased the speed of floating point processors. As a result, these modules are increasingly introduced in real time applications of estimation, control, digital filtering, and general signal processing, and the need for a comprehensive analysis of its limitations (due to finite wordlength effects) is obvious. No fully comprehensive model for the wordlength effects in floating point exists, although some very significant contributions have been made in the past. Attempts to give rigorous analysis of a sequence of floating point operations have proven to be so formidable that one has to content one's self with plausibility arguments (e.g., see [4], p 213). A noteworthy contribution is the axiomatization by Kulish [5].

General statistical modeling of floating point errors relies on the work of Wilkinson [16], Forsythe and Moler [2], Stewart [11], Knuth [4], and Vandergraft [12]. Brent [1], Marasa and Matula [6] performed extensive simulations for various finite precision arithmetic systems. An analysis of the effects in digital filtering is due to Kan and Aggarwal [3] and others. Rink and Chong analyzed the performance of a floating point state regulator [9]. Van Wingerden and De Koning [13] recently combined this work with a Monte Carlo identification technique. A dynamical stochastic model was used by Verriest [14] in the computation of a gain correction for (Kalman) filtering applications in floating point. In this paper, the finite wordlength effects on discrete linear recursions of the form ($\dim x = n$, $\dim u = m$, $\dim y = p$)

$$\begin{aligned}x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k\end{aligned}\quad (1)$$

are analyzed. The paper is organized as follows. First a model for the floating point errors is discussed. The third section then uses this error model to obtain representations for various confidence measures.

2. THE ERROR MODEL

We shall work with normalized floating point numbers. For a given base b , they are expressed

by two numbers, e and f . Here e is an integer exponent and f is a signed fraction assumed to be normalized, i.e., $b^{-1} < |f| < 1$. The value of the floating point number is then

$$fb^e$$

Since the coding of the exponent e and the signed fraction must fit into a given wordlength (w), there will be a tradeoff between the precision and the range of the representable numbers in the computer [1]. However, in this work we shall assume an infinite range and only consider the effects of a finite number of digits in the fraction. Besides simplifying the analysis, this assumption can be substantiated by the fact that normally an underflow or overflow would cause program termination, and we are only interested in the finite wordlength effects during a normal program execution. Following Kulish [5], the situation is as follows. We have R , the set of reals and an operation $*$ (which is $+$, $-$, $\times$, or $\div$). On the computer the elements of R , as well as the results of $a * b$ are not exactly representable. Hence the reals must be mapped in a subset P according to a proper (i.e., monotone and symmetric) mapping $Q: R \rightarrow P$. The approximation of the $*$ operation is then $Q[a * b]$

$$Q[a * b] := Q(a * b) \quad (2)$$

Unfortunately, the in general not representable result $a * b$ seems to be necessary for its realization. It can be shown, however, that in all cases where $a * b$ is not exactly representable, it is sufficient to replace it by an appropriate and representable value $a \bar{*} b$ with the property

$$Q(a * b) = Q(a \bar{*} b) \quad (3)$$

Then the proper definition is

$$Q[a * b] := Q(a \bar{*} b) \quad (4)$$

The concrete algorithms for the realization of this formula can then be decomposed into four steps.

1. Identification of the exponent and fraction of a and b .
2. Execution of $a \bar{*} b$.
3. Renormalization.
4. Mapping into P (because accumulation of higher wordlength may be used).

The cause of floating point errors is threefold. First, there are intrinsic errors, due to finite wordlength representation of a given number (parameters, inputs...). Even if two numbers have an exact representation, binary operations (sum, product...) on them may require a longer wordlength for exact representation. Hardware implementation greatly affects this error (presence or absence of guardbits, double register arithmetic, rounding or truncation...). The errors induced by binary operations are referred to as extrinsic errors. Finally all these errors propagate through the recursion and hence accumulate. They

are called the inherent errors since they inherit their properties from the operation sequence and the given recursion.

The intrinsic errors are bounded by the "unit in the least significant digit" times b^e . The error is therefore uniform in an interval with length proportional to the number itself (i.e., b^e). The extrinsic errors are also proportional to the computed result [12], with an exception of subtraction of nearly equal quantities, which may cause a blowup of the relative error. For the inherent error (also called accumulated error) Wilkinson [16] (also Forsythe and Moler [2]) gives errorbounds which are proportional to the computed result. If y is the exact result of a combination of n multiplications and divisions, then the relative error in the computed result is bounded by nY_0 for some Y_0 . Many other bounds on the rounding errors in algebraic processes are also of the form $f(n)$. The linear (in n) bound is rather conservative, for the individual rounding errors in a compounded expression tend to cancel rather than to reinforce each other if an unbiased rounding rule is used. With biased rounding and truncation, the bound may be more realistic.

Because of the above observations, we are led to a stochastic model for the finite wordlength error.

$$y - Q[y] = Y(n)y_c \quad (5)$$

where c is a sample of a standard white Gaussian process and $Y(n)$ is a normalization factor, dependent on the number of operations. The "large" samples c simulate then the occasional blow up due to subtraction of near equal quantities (occurring with empirical frequency .14 [4]). It can be shown (by considering error accumulation in one single batch of n^2 or n batches of n operations) that for consistency of (1) $Y(n)$ must be order $\sqrt{n}$.

Remark: The above approximation (5) will be invalid if the number of binary operations greatly exceeds the number of independent variables occurring as operands.

3. ANALYSIS OF LINEAR RECURSIONS

The formulas (1) are generic state space representations for digital filters or compensators as for instance used in feedback controllers. The signals u_k and y_k are respectively the input and the output vectors. As explained in the previous section, the bilinear error model is assumed, as well as a perfect representation of the parameter matrices A , B , and C . (This entails no loss of generality since the effect of parameter truncation can always be "thrown back" to the data [12]. By equation (5), the recursions in floating point can be modeled by (assuming the use of an unbiased rounding).

$$x_{k+1}^m = Ax_k^m + Bu_k + Y \text{diag}(Ax_k^m + Bu_k) D_1 w_k \quad (6)$$

$$y_k^m = Cx_k^m + \beta \text{diag}(Cx_k^m) D_2 v_k \quad (7)$$

where $w_k^m = v_k^m$ in an $(n+p)$ dimensional standard white gaussian sequence, β and γ are normalization parameters which are purely hardware dependent, while the elements of the matrices D are realization dependent, and reflect dependencies among the computed state or output components. (If two components of the x -vector are updated in identical ways, then their errors must also remain equal.) β and γ are fixed such that the maximal sum of squares of the elements for each row of W is one. If A is a full matrix without any particular structure, then generically one can set $W = I$, and γ corresponds with $(n+m)$ multiplications and $n+m-1$ (signed) additions. A special case occurs for instance if a pair (A, b) is in canonical form, i.e., for

$$A = \begin{bmatrix} -a_1 & \dots & -a_n \\ & & 0 \\ & & \vdots \\ & & 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (8)$$

one has

$$b = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (9)$$

For truncation or biased rounding, the probabilistic model needs to be adjusted to incorporate the bias. For each type of arithmetic, a bilinear state model arises. For this reason we shall refer to the floating point error model in the previous section as the "bilinear error model."

The propagation of the expected value of the above model state is

$$\bar{x}_{k+1}^m = A\bar{x}_k^m + B\bar{u}_k \quad (10)$$

Clearly, if $\bar{u}_k = u_k$ and $\bar{x}_0^m = x_0^m$, then the solution of (10) and (1) are identical. Therefore, the exact recursion (1) can be interpreted as the expectation of the floating point model. Subtracting (10) from (6), the floating point error $x^m = x^m - \bar{x}^m$ satisfies

$$\bar{x}_{k+1}^m = [I + \gamma \text{diag}(Dw_k)] A\bar{x}_k^m + \gamma \text{diag}(Dw_k) \bar{x}_{k+1}^m \quad (11)$$

A criterion for almost sure stability can be established for first order systems based on Grintsevichyus' theorem. For higher order systems (the case of interest), we shall only be concerned with the first and second moments. The following properties were derived:

Theorem 1: For the bilinear model (6)-(11), the error covariances

$$\begin{aligned} P^m &\triangleq E(x^m - \bar{x}^m)(x^m - \bar{x}^m)^T \\ V^m &\triangleq E(y^m - \bar{y}^m)(y^m - \bar{y}^m)^T \end{aligned} \quad (12)$$

solve the recursion

$$P_{k+1}^m = AP_k^m A^T + \gamma^2 D_1 D_1^T \cdot (AP_k^m A^T + \bar{L}_{k+1}) \quad (13)$$

$$\bar{L}_k = E x_k^m (x_k^m)^T \quad (14)$$

$$\bar{x}_{k+1}^m = A\bar{x}_k^m + B\bar{u}_k \quad (15)$$

$$V_{k+1}^m = CP_k^m C^T + \beta^2 D_2 D_2^T \cdot C(P_k^m + \bar{L}_k^m) C^T \quad (16)$$

A proof is straightforward by "squaring up" (11) and taking expectations, noting that w_k is independent from x_k and $\bar{x}_k$. Finally the identity

$$\text{diag}(x) Q \text{diag}(x) = xx^T \cdot Q \quad (17)$$

is used where $\cdot$ is the Schur product (i.e., $(A \cdot B)_{ij} = A_{ij} B_{ij}$).

Remarks:

1. If the input u_k is purely random with zero mean and covariance Q_k , then

$$\bar{L}_{k+1} = A\bar{L}_k A^T + BQ_k B^T \quad (18)$$

is substituted for (16) and (17). The relative error is then the "ratio" $P^m V^{-1}$, where $V_k = E x_k x_k^T$.

2. Defining Z as the unit under $\cdot$, i.e., $E_{ij} = 1$ we can rewrite (15) as

$$P_{k+1}^m = (Z + \gamma^2 DD^T) \cdot AP_k^m A^T + \gamma^2 DD^T \cdot \bar{L}_{k+1} \quad (13')$$

It is clear that even when A is strictly stable, P_k^m may grow unboundedly, due to the presence of the positive Schur factor $Z + \gamma^2 DD^T$.

Theorem 2: The model (6) is second order stochastically stable if the eigenvalues of A are within a circle with radius $(1 + \gamma^2)^{-1/2}$.

Another measure of the similarity of computed result and exact result of the recursion is given by the generalized correlation coefficient, between the outputs of the exact (y) and the finite precision system (y^f). We define this generalized correlation coefficient as

$$\rho^f = \lim_{N \rightarrow \infty} \rho_N(y, y^f) \quad (19)$$

$$\rho_N(y, y^f) = \frac{\text{Tr}(Y(Y^f)^T)}{(\text{Tr}(YY^T) \text{Tr}(Y^f(Y^f)^T))^{1/2}} \quad (20)$$

where Y is the data matrix $(y_1, y_2, \dots, y_N)$, and similarly for Y^f . Note that $\rho_N(y, y^f)$ can be written in terms of the sample correlation functions

$$\rho_N(y, y^f) = \frac{\frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i (y_i^f)')}{\sqrt{\frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i y_i') \frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i^f (y_i^f)')}} \quad (21)$$

assuming a random input with variance Q_k , and invoking ergodicity. For the model (6)-(11), the $\rho_N(y, y^m)$ can be precomputed.

Theorem 3: The steady state correlation between the model output y^m and $y (= y^m \text{ for } \gamma = \beta = 0)$ is

$$\rho^m = \frac{1}{\sqrt{1 + \beta^2}} \sqrt{\frac{\text{Tr}(C \Pi_m C')}{\text{Tr}(C \Pi C')}} \quad (22)$$

where Π_m and Π are the solutions to the (extended) Lyapunov equation

$$X = (A X A' + B Q B') * (E + \sigma D_1 D_1') \quad (23)$$

for $\sigma = \gamma^2$ and $\sigma = 0$, respectively.

Note that the quantity ρ^m can easily be interpreted in terms of a signal to (computation) noise ratio. Based on the bilinear model we can now try to find special realizations for which the error covariance is minimal or the correlation coefficient is maximal. It was found by simulation ($DD' = I$) that many equivalent optima exists. In fact, the error measures fluctuate rapidly between a minimum and maximum value. This high sensitivity may make an optimal realization impractical. We established also the (expected).

Theorem 4: If $DD' = I$, then scaling (i.e., a diagonal similarity transformation) leaves the error properties invariant.

Theorem 5: For a given realization (A,B,C), the error properties are left invariant if Q is multiplied by a positive constant.

Remark: We have not touched upon certain important and interesting issues. If very low precision is used, it can be shown that trapstates may exist. These are vectors of floating point numbers such that $Q[x_{k+1}] = Q[x_k]$ for the undriven system. Obviously, zero will be a trapstate, but many more may exist, depending on the iteration and the precision. Whenever nonzero trapstates exist, the bilinear model will break down. The details are under study.

REFERENCES

- [1] R.P. Brent, "On the Precision Attainable with Various Floating-Point Number Systems," *IEEE Trans. on Computers*, Vol. C-22, No. 6, pp. 601-607, June 1973.
- [2] G.E. Forsythe and C.B. Moler, "Computer Solution of Linear Algebraic Systems," Prentice-Hall, 1967.
- [3] E.P.P. Kan and J.K. Aggarwal, "Error Analysis of Digital Filter Employing Floating-Point Arithmetic," *IEEE Trans. Circuit Theory*, Vol. 18, No. 6, pp. 678-686, November 1971.
- [4] D.E. Knuth, "The Art of Computer Programming," Vol. 2, Addison-Wesley, 1981.
- [5] U. Kulisch, "Mathematical Foundation of Computer Arithmetic," *IEEE Trans. on Computers*, Vol. C-26, No. 7, pp. 610-621, July 1977.
- [6] J.D. Marasa and D.W. Matula, "A Simulative Study of Correlated Error Propagation in Various Finite-Precision Arithmetics," *IEEE Trans. on Computers*, Vol. C-22, No. 6, pp. 587-597, June 1973.
- [7] P. Moroney, "Issues in the Implementation of Digital Feedback Compensators," MIT Press, 1983.
- [8] C.T. Mullis and R.A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits and Systems*, Vol. 23, No. 9, pp. 551-562, September 1976.
- [9] R.E. Rink and H.Y. Chong, "Performance of State Regulator Systems with Floating-Point Computation," *IEEE Trans. Auto. Control*, Vol. 24, No. 3, pp. 411-421, June 1979.
- [10] A.B. Sripad and D.L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," *IEEE Trans. ASSP*, Vol. 25, No. 5, pp. 442-448, October 1977.
- [11] G.W. Stewart, "Introduction to Matrix Computations," Academic Press, 1973.
- [12] J.S. Vandergraft, "Introduction to Numerical Computations," Academic Press, 1983.
- [13] A.J.M. Van Wingerden and W.L. DeKoning, "The Influence of Finite Wordlength on Digital Optimal Control," *IEEE Trans. Auto. Control*, Vol. 29, No. 5, pp. 385-391, May 1984.
- [14] E.I. Verriest, "On Redefining the Optimal Least Squares Filter under Floating-Point Operations," *Proc. ICASSP*, p. 30.9, San Diego, CA, March 1984.
- [15] E.I. Verriest, "Gain Correction in Optimal Filtering using Floating Point Arithmetic," *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, NV, December 1984.
- [16] J.H. Wilkinson, "Rounding Errors in Algebraic Processes," Prentice Hall 1963.

APPENDIX N

A BILINEAR MODEL FOR LINEAR RECURSIVE COMPUTATIONS USING
FLOATING POINT ARITHMETIC

A BILINEAR MODEL FOR LINEAR RECURSIVE COMPUTATIONS USING FLOATING POINT ARITHMETIC

Erik I. Verriest

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

ABSTRACT

A stochastic error model for floating point arithmetic is developed. A characteristic feature of the floating point error in an operation is that its bounds are proportional to the result of the operation. This model is used to study the effects of finite wordlength on linear recursive formulas. Optimal realizations exist, but they are highly sensitive.

1. INTRODUCTION

The accumulation and roundoff error in long computerized calculations and recursive algorithms is a phenomenon that can destroy an efficient and sound computational procedure based on arithmetic over the real number field. A tell-tale example is the Kalman filter divergence. Analytically this state estimator yields the estimate with minimal covariance, but the actual error covariance may be much larger than the predicted covariance (which solves a certain Riccati equation) or even grow unboundedly whenever a finite precision version of the algorithm is implemented.

It is clear that in order to keep track of the confidence in the computed results, a certain measure of confidence should be computed and "tracked" with each operation or update.

One such measure is the rigorous computation of error bounds via interval analysis. Not only are the additional computations that are required time consuming, but the obtained results may be overly conservative. The individual rounding errors in a compounded expression indeed tend to cancel rather than to reinforce each other if an unbiased rounding rule is used. In this paper another measure of confidence is used. It is of a more probabilistic nature and estimates the propagated covariance due to the finite wordlength errors. Such a study is standard (and straightforward) for fixed point arithmetic, and is well described in several textbooks, culminating in the optimal filter implementations by Mullis and Roberts [8] and the LQG compensator by Moroney [7]. One of the serious shortcomings of

the use of fixed point arithmetic is the necessity of scaling in order to provide a higher accuracy. This disadvantage is obliterated when floating point arithmetic is used.

Modern digital technology has rapidly increased the speed of floating point processors. As a result, these modules are increasingly introduced in real time applications of estimation, control, digital filtering, and general signal processing, and the need for a comprehensive analysis of its limitations (due to finite wordlength effects) is obvious. No fully comprehensive model for the wordlength effects in floating point exists, although some very significant contributions have been made in the past. Attempts to give rigorous analysis of a sequence of floating point operations have proven to be so formidable that one has to content one's self with plausibility arguments (e.g., see [4], p. 213). A noteworthy contribution is the axiomatization by Kulish [5].

General statistical modeling of floating point errors relies on the work of Wilkinson [16], Forsythe and Moler [2], Stewart [11], Knuth [4], and Vandergraft [12]. Brent [1], Marasa and Matula [6] performed extensive simulations for various finite precision arithmetic systems. An analysis of the effects in digital filtering is due to Kan and Aggarwal [3] and others. Rink and Chong analyzed the performance of a floating point state regulator [9]. Van Wingerden and De Koning [13] recently combined this work with a Monte Carlo identification technique. A dynamical stochastic model was used by Verriest [14] in the computation of a gain correction for (Kalman) filtering applications in floating point. In this paper, the finite wordlength effects on discrete linear recursions of the form ($\dim x = n, \dim u = m, \dim y = p$)

$$\begin{aligned}x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k\end{aligned}\quad (1)$$

are analyzed. The paper is organized as follows. First a model for the floating point errors is discussed. The third section then uses this error model to obtain representations for various confidence measures.

2. THE ERROR MODEL

We shall work with normalized floating point numbers. For a given base b , they are expressed

This work was supported by the U.S. Air Force under Contract F-08635-84-C-0273.

by two numbers, e and f . Here e is an integer exponent and f is a signed fraction assumed to be normalized, i.e., $b^{-1} < |f| < 1$. The value of the floating point number is then

$$fb^e$$

Since the coding of the exponent e and the signed fraction must fit into a given wordlength (w), there will be a tradeoff between the precision and the range of the representable numbers in the computer [1]. However, in this work we shall assume an infinite range and only consider the effects of a finite number of digits in the fraction. Besides simplifying the analysis, this assumption can be substantiated by the fact that normally an underflow or overflow would cause program termination, and we are only interested in the finite wordlength effects during a normal program execution. Following Kulish [5], the situation is as follows. We have R , the set of reals and an operation $*$ (which is $+$, $-$, $\times$, or $\div$). On the computer the elements of R , as well as the results of $a * b$ are not exactly representable. Hence the reals must be mapped in a subset F according to a proper (i.e., monotone and symmetric) mapping $Q: R \rightarrow F$. The approximation of the $*$ operation is then $Q(a * b)$

$$Q(a * b) := Q(a * b) \quad (2)$$

Unfortunately, the in general not representable result $a * b$ seems to be necessary for its realization. It can be shown, however, that in all cases where $a * b$ is not exactly representable, it is sufficient to replace it by an appropriate and representable value $a * b$ with the property

$$Q(a * b) = Q(a * b) \quad (3)$$

Then the proper definition is

$$Q(a * b) := Q(a * b) \quad (4)$$

The concrete algorithms for the realization of this formula can then be decomposed into four steps.

1. Identification of the exponent and fraction of a and b .
2. Execution of $a * b$.
3. Renormalization.
4. Mapping into F (because accumulation of higher wordlength may be used).

The cause of floating point errors is threefold. First, there are intrinsic errors, due to finite wordlength representation of a given number (parameters, inputs...). Even if two numbers have an exact representation, binary operations (sum, product...) on them may require a longer wordlength for exact representation. Hardware implementation greatly affects this error (presence or absence of guardbits, double register arithmetic, rounding or truncation...). The errors induced by binary operations are referred to as extrinsic errors. Finally all

these errors propagate through the recursion and hence accumulate. They are called the inherent errors since they inherit their properties from the operation sequence and the given recursion.

The intrinsic errors are bounded by the "unit in the least significant digit" times b^e . The error is therefore uniform in an interval with length proportional to the number itself (i.e., b^e). The extrinsic errors are also proportional to the computed result [12], with an exception of subtraction of nearly equal quantities, which may cause a blowup of the relative error. For the inherent error (also called accumulated error) Wilkinson [16] (also Forsythe and Moler [2]) gives errorbounds which are proportional to the computed result. If y is the exact result of a combination of n multiplications and divisions, then the relative error in the computed result is bounded by $n\gamma_0$ for some γ_0 . Many other bounds on the rounding errors in algebraic processes are also of the form $f(n)$. The linear (in n) bound is rather conservative, for the individual rounding errors in a compounded expression tend to cancel rather than to reinforce each other if an unbiased rounding rule is used. With biased rounding and truncation, the bound may be more realistic.

Because of the above observations, we are led to a stochastic model for the finite wordlength error.

$$y - Q[y] = Y(n)\gamma_c \quad (5)$$

where c is a sample of a standard white Gaussian process and $Y(n)$ is a normalization factor, dependent on the number of operations. The "large" samples c simulate then the occasional blow up due to subtraction of near equal quantities (occurring with empirical frequency .14 [4]). It can be shown (by considering error accumulation in one single batch of n^2 or n batches of n operations) that for consistency of (1) $Y(n)$ must be order $\sqrt{n}$.

Remark: The above approximation (5) will be invalid if the number of binary operations greatly exceeds the number of independent variables occurring as operands. In the case of an equal mixture of $+$, $-$, $\times$, and $\div$, Marasa and Matula [6] have shown by extensive simulation on combinations of 10^3 operations, that the relative error grows slightly faster than exponentially.

3. ANALYSIS OF LINEAR RECURSIONS

The formulas (1) are generic state space representations for digital filters or compensators as for instance used in feedback controllers. The signals u_k and y_k are respectively the input and the output vectors. As explained in the previous section, the bilinear error model is assumed, as well as a perfect representation of the parameter matrices A , B , and C . (This entails no loss of generality since the effect of parameter truncation can always be "thrown back" to the data [12]. By equation (5), the

recursions in floating point can be modeled by (assuming the use of an unbiased rounding).

$$x_{k+1}^m = Ax_k^m + Bu_k + \gamma \text{diag} (Ax_k^m + Bu_k) D_1 w_k \quad (6)$$

$$y_k^m = Cx_k^m + \beta \text{diag} (Cx_k^m) D_2 v_k \quad (7)$$

where $w_k' = v_k'$ in an $(n+p)$ dimensional standard white Gaussian sequence, β and γ are normalization parameters which are purely hardware dependent, while the elements of the matrices D are realization dependent, and reflect dependencies among the computed state or output components. (If two components of the x -vector are updated in identical ways, then their errors must also remain equal.) β and γ are fixed such that the maximal sum of squares of the elements for each row of W is one. If A is a full matrix without any particular structure, then generically one can set $W = I$, and γ corresponds with $(n+m)$ multiplications and $n+m-1$ (signed) additions. A special case occurs for instance if a pair (A, b) is in canonical form, i.e., for

$$A = \begin{bmatrix} -a_1 & \dots & -a_n \\ & & Q \\ & & \vdots \\ & & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ Q \\ \vdots \\ 0 \end{bmatrix} \quad (8)$$

one has

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & 0 & \end{bmatrix} \quad (9)$$

For truncation or biased rounding, the probabilistic model needs to be adjusted to incorporate the bias, e.g. for truncation

$$Q_c[x] = x + \gamma_c |x| (w - \frac{1}{2} \text{sgn}(x)) \\ = (1 - \frac{1}{2} \gamma_c) x + \gamma_c x w' \quad (5')$$

where we set $w' = w \text{sgn}(x)$. For each type of arithmetic, a bilinear state model arises. For this reason we shall refer to the floating point error model in the previous section as the "bilinear error model."

The propagation of the expected value of the above model state is

$$\hat{x}_{k+1}^m = A \hat{x}_k^m + B u_k \quad (10)$$

Clearly, if $\hat{u}_k = u_k$ and $\hat{x}_0^m = x_0^m$, then the solution of (10) and (1) are identical. Therefore, the exact recursion (1) can be interpreted as the expectation of the floating point model. Subtracting (10) from (6), the floating point error $x^m = x^m - \hat{x}^m$ satisfies

$$\hat{x}_{k+1}^m = [I + \gamma \text{diag} (Dw_k)] A \hat{x}_k^m + \gamma \text{diag} (Dw_k) \hat{x}_{k+1}^m \quad (11)$$

A criterion for almost sure stability can be established for first order systems based on Grintsevichyus' theorem [17, p.153] for steady-

state conditions (i.e., if $\hat{x}_k + \hat{x}_m = \frac{bu_m}{1-a}$). Namely if

$$-\infty < \mathbb{E} \log |a(1 + \gamma w_k)| < 0$$

then the error x^m is a.s. convergent for all k iff

$$\mathbb{E} \log \sup(\gamma |x_u w_k|, 1) < 0$$

i.e., if $|a| < f(\gamma)$ for some function f of γ . For higher order systems (the case of interest), we shall only be concerned with the first and second moments. The following properties were derived:

Theorem 1: For the bilinear model (6)-(11), the error covariances

$$P_k^m \triangleq \mathbb{E}(x^m - \hat{x}^m)(x^m - \hat{x}^m)', \\ V_k^m \triangleq \mathbb{E}(y^m - \hat{y}^m)(y^m - \hat{y}^m)', \quad (12)$$

solve the recursion

$$P_{k+1}^m = A P_k^m A' + \gamma^2 D_1 D_1' + (A P_k^m A' + L_{k+1}^m) \quad (13)$$

$$L_k^m = \mathbb{E} x_k^m (x_k^m)', \quad (14)$$

$$\hat{x}_{k+1}^m = A \hat{x}_k^m + B u_k \quad (15)$$

$$V_k^m = C P_k^m C' + \beta^2 D_2 D_2' + C(P_k^m + L_k^m)C' \quad (16)$$

A proof is straightforward by "squaring up" (11) and taking expectations, noting that w_k is independent from x_k and $\hat{x}_k$. Finally the identity

$$\text{diag}(x) Q \text{diag}(x) = x x' * Q \quad (17)$$

is used where $*$ is the Schur product (i.e., $(A*B)_{ij} = A_{ij}B_{ij}$).

Remarks:

1. If the input u_k is purely random with zero mean and covariance Q_k , then

$$L_{k+1}^m = A L_k^m A' + B Q_k B' \quad (18)$$

is substituted for (16) and (17). The relative error is then the "ratio" $P^m V^{-1}$, where $V_k = \mathbb{E} x_k x_k'$.

2. Defining E as the unit under $*$, i.e., $E_{ij} = 1$ we can rewrite (13) as

$$P_{k+1}^m = (E + \gamma^2 D D') * A P_k^m A' + \gamma^2 D D' * L_{k+1}^m \quad (13')$$

It is clear that even when A is strictly stable, P_k^m may grow unboundedly, due to the presence of the positive Schur factor $E + \gamma^2 D D'$.

Theorem 2: The model (6) is second order stochastically stable if the eigenvalues of A are within a circle with radius $(1 + \gamma^2)^{-1/2}$.

Another measure of the similarity of computed result and exact result of the recursion is given by the generalized correlation coefficient, between the outputs of the exact (y) and the finite precision system (y^f). We define this generalized correlation coefficient as

$$\rho^f = \lim_{N \rightarrow \infty} \rho_N(y, y^f) \quad (19)$$

$$\rho_N(y, y^f) = \frac{\text{Tr}(Y(Y^f)')}{(\text{Tr}(YY')\text{Tr}(Y^f(Y^f)'))^{1/2}} \quad (20)$$

where Y is the data matrix $(y_1, y_2, \dots, y_N)$, and similarly for Y^f . Note that $\rho_N(y, y^f)$ can be written in terms of the sample correlation functions

$$\rho_N(y, y^f) = \frac{\frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i(y_i^f)')}{\sqrt{\frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i y_i') \frac{1}{N} \text{Tr}(\sum_{i=1}^N y_i^f(y_i^f)')}} \quad (21)$$

assuming a random input with variance Q_k , and invoking ergodicity. For the model (6)-(11), the $\rho_N(y, y^f)$ can be precomputed.

Theorem 3: The steady state correlation between the model output y^m and y ($= y^m$ for $\gamma = \beta = 0$) is

$$\rho^m = \frac{1}{\sqrt{1 + \beta^2}} \sqrt{\frac{\text{Tr}(C \Pi_{\infty} C')}{\text{Tr}(C \Pi_{\infty}^m C')}} \quad (22)$$

where Π_{∞}^m and Π_{∞} are the solutions to the (extended) Lyapunov equation

$$X = (AXA' + BQB') + (E + \sigma D_1 D_1') \quad (23)$$

for $\sigma = \gamma^2$ and $\sigma = 0$, respectively.

Note that the quantity ρ^m can easily be interpreted in terms of a signal to (computation) noise ratio. Based on the bilinear model we can now try to find special realizations for which the error covariance is minimal or the correlation coefficient is maximal. It was found by simulation ($DD' = I$) that many equivalent optima exists. In fact, the error measures fluctuate rapidly between a minimum and maximum value. This high sensitivity may make an optimal realization impractical. We established also the (expected).

Theorem 4: If $DD' = I$, then scaling (i.e., a diagonal similarity transformation) leaves the error properties invariant.

Theorem 5: For a given realization (A,B,C), the error properties are left invariant if Q is multiplied by a positive constant.

Remark: We have not touched upon certain important and interesting issues. If very low precision is used, it can be shown that trapstates may exist. These are vectors of floating point numbers such that $Q(x_{k+1}) = Q(x_k)$ for the undriven system. Obviously, zero will be a trapstate, but many more may exist, depending on the iteration and the precision. Whenever nonzero trapstates exist, the bilinear model will break down. The details are under study.

4. FINAL REMARKS AND CONCLUSIONS

Simulation of the bilinear model yields an output that fluctuates too fast as compared to the output of a simulation of the roundoff effects on the recursions. This leads to the conclusion that the developed error model is only good to provide RMS bounds on the error. The sample paths of the bilinear model (6-7) are in no way a good representation of an actual sample run of a computer with small word length. Our suggestion is to run the covariance updates along with the actual recursion (i.e., one updates all of (13-16)). A one- σ or three- σ confidence ellipse can then be constructed around the computed update y_k^f , based on the matrix v_k^m .

Concluding, we state that a bilinear error model can be used to obtain confidence bounds on computed recursions. These bounds are less conservative than the absolute bounds provided by interval arithmetic. Based on this model, we have shown that optimal realizations exist, but are too sensitive to be of practical value.

REFERENCES

- [1] R.P. Brent, "On the Precision Attainable with Various Floating-Point Number Systems," *IEEE Trans. on Computers*, Vol. C-22, No. 6, pp. 601-607, June 1973.
- [2] G.E. Forsythe and C.B. Moler, "Computer Solution of Linear Algebraic Systems," Prentice-Hall, 1967.
- [3] E.P.F. Kan and J.K. Aggarwal, "Error Analysis of Digital Filter Employing Floating-Point Arithmetic," *IEEE Trans. Circuit Theory*, Vol. 18, No. 6, pp. 678-686, November 1971.
- [4] D.E. Knuth, "The Art of Computer Programming," Vol. 2, Addison-Wesley, 1981.
- [5] U. Kulisch, "Mathematical Foundation of Computer Arithmetic," *IEEE Trans. on Computers*, Vol. C-26, No. 7, pp. 610-621, July 1977.

- [6] J.D. Marasa and D.W. Matula, "A Simulative Study of Correlated Error Propagation in Various Finite-Precision Arithmetics," IEEE Trans. on Computers, Vol. C-22, No. 6, pp. 587-597, June 1973.
- [7] P. Moroney, "Issues in the Implementation of Digital Feedback Compensators," MIT Press, 1983.
- [8] C.T. Mullis and R.A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," IEEE Trans. Circuits and Systems, Vol. 23, No. 9, pp. 551-562, September 1976.
- [9] R.E. Rink and H.Y. Chong, "Performance of State Regulator Systems with Floating-Point Computation," IEEE Trans. Auto. Control, Vol. 24, No. 3, pp. 411-421, June 1979.
- [10] A.B. Sripad and D.L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," IEEE Trans. ASSP, Vol. 25, No. 5, pp. 442-448, October 1977.
- [11] G.W. Stewart, "Introduction to Matrix Computations," Academic Press, 1973.
- [12] J.S. Vandergraft, "Introduction to Numerical Computations," Academic Press, 1983.
- [13] A.J.M. Van Wingerden and W.L. DeKoning, "The Influence of Finite Wordlength on Digital Optimal Control," IEEE Trans. Auto. Control, Vol. 29, No. 5, pp. 385-391, May 1984.
- [14] E.I. Verriest, "On Redefining the Optimal Least Squares Filter under Floating-Point Operations," Proc. ICASSP, p. 30.9, San Diego, CA, March 1984.
- [15] E.I. Verriest, "Gain Correction in Optimal Filtering using Floating Point Arithmetic," Proc. 23rd IEEE Conf. on Decision and Control, Las Vegas, NV, December 1984.
- [16] J.H. Wilkinson, "Rounding Errors in Algebraic Processes," Prentice Hall 1963.
- [17] A. Mukherjea, "Limit Theorems: Stochastic Matrices, Ergodic Markov Chains and Measures on Semigroups," in Probabilistic Analysis and Related Topics, Vol. 2, ed. Bhurucha-Reid, Academic Press, 1979.

APPENDIX O

GAIN CORRECTION IN OPTIMAL FILTERING USING FLOATING POINT

ARITHMETIC

E. I. Verriest

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

Abstract

The effects of the finite word length in a floating point implementation of the least squares filter is discussed. Optimal precomputable gains are given, and a computationally more attractive approximation is given.

1. INTRODUCTION

Optimality of the Kalman filter is only guaranteed if computations can be performed in infinite precision. Therefore the realistically computed estimates are no longer optimal. This paper takes the finite word length constraints of the digital machine into account in the algorithm design.

Finite word length effects in fixed point are now well understood and described in several books, culminating in the optimum filter implementations by Mullis and Roberts [1] and the compensator by Moroney [2].

General statistical modeling of floating point errors relies on the work of Wilkinson [3], Knuth [4] and Vandergraft [5]. The effects on the performance of compensators have been studied [6,7]. A gain adjustment for the filter was given by this author in [8].

2. THE ERROR MODEL

The cause of floating point errors is three-fold. First, there are "intrinsic" errors, due to finite wordlength representation of a given number (parameters, inputs...). Even if two numbers have an exact representation, binary operations (sum, product...) on them may require a longer wordlength for exact representation. Hardware implementation greatly affects this error (presence or absence of guardbits, double register arithmetic, rounding or truncation...). The errors induced by binary operations are referred to as "extrinsic" errors. Finally all these errors propagate through the recursion and hence accumulate. They are called the "inherent" errors since they inherit their properties from the operation sequence and the given recursion.

A characteristic feature of the floating point errors is that they are proportional to the computed results (i.e. one has "multiplicative noise"). If $Q[y_k]$ represents the result of a recursive computation of a vector y_k , then the model of the computation error is

$$y_k - Q[y_k] = \gamma \text{diag}(y_k) w_k \quad (1)$$

where w_k is a sample of an n -dimensional standard white Gaussian sequence, γ is a parameter which is

purely hardware dependent and the elements of W depend on the particular recursion (dependencies of components of y_k).

The rest of the paper deals with the filter for the model

$$x_{k+1} = Fx_k + Gu_k \quad (2)$$

$$y_k = Hx_k + v_k$$

$$(u_k', v_k')' \sim N(0, 0), \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \quad (3)$$

3. THE OPTIMAL FILTER FOR DEGRADED ARITHMETIC

First it will be assumed that all gains can be precomputed and set in infinite precision. This not only simplifies the analysis, but can also be justified. The error due to the difference in the desired and implemented gain can be "thrown" back to the data, a technique known as inverse error analysis. Hence without loss of generality the (actual) computed filter update (with gain K_k) is modeled by (suppressing the "Q"-notation)

$$\hat{x}_{k+1} = F\hat{x}_k + K_k(y_k - H\hat{x}_k) + \phi_k \quad (4)$$

where ϕ_k is the finite wordlength error. The part $F\hat{x}_k + K_k(y_k - H\hat{x}_k)$ which is the desired or theoretical update will be defined as the CORE of the estimate. The estimation error $x = x - \hat{x}$ satisfies then

$$\hat{x}_{k+1} = F\hat{x}_k + Gu_k - K_k H\hat{x}_k - K_k v_k - \phi_k \quad (5)$$

The covariances $P = E\hat{x}\hat{x}'$ and $I = E\hat{x}x'$ satisfy the coupled matrix equations $C = E\hat{x}x'$

$$I_{k+1} = F I_k F' + K_k H C_k' P' + F C_k H' K_k' + K_k R_k K_k' + E(\phi_k \phi_k') \quad (6)$$

$$C_{k+1} = F C_k P' + K_k H P_k P' - F C_k H' K_k' - K_k R_k K_k' - E(\phi_k \phi_k') \quad (7)$$

$$P_{k+1} = (P - K_k H) P_k (P - K_k H)' + G Q G' + K_k R_k K_k' + E(\phi_k \phi_k') \quad (8)$$

where R_k^c is as usual the innovations covariance

$$R_k^c = H P_k H' + R \quad (9)$$

Appropriate initial conditions are

$$I_0 = 0, C_0 = 0, P_0 = \tau_0$$

For fixed point arithmetic, the driving term in (6)-(9) is

$$E(\phi_k \phi_k') = \delta^2 W \quad (10)$$

and thus not only constant, but also independent of C , P and L . The optimal gain can then be found directly by minimization with respect to K_k .

$$K_k^* = PP_k H' R_k^{-c} \quad (11)$$

However, generically, the computed estimate is no longer orthogonal to its error.

The situation is different under floating point arithmetic. In this case the statistics of the computation noise depend on the core estimate. As seen in section 2

$$\phi_k = Y \text{diag}[P_k^* K_k (y_k - Hx_k)] W W_k \quad (12)$$

The resulting driving term in (8) is the expectation ($\circ$ is the Schur product)

$$E(\phi_k \phi_k^*) = Y^2 (L_k \circ W W^*) \quad (13)$$

where L_k is the core update of L , i.e.

$$L_k = P L_k P' + K_k H C_k' P' + P C_k H' K_k' + K_k R_k C_k' \quad (14)$$

This couples (8) to (6) and (7). The optimal filter can be derived via an equivalent constrained minimization problem of the final covariance

$$\phi_N \equiv E|x - x_N|^2 = \text{Tr}(P_N) \quad (15)$$

with respect to the sequence $\{K_k\}$, and subject to (6)-(8), (13) and (14).

Adjoining the constraints, the Hamiltonian for the system is

$$\bar{H}_k = \text{Tr}\{A_{k+1}^P P_{k+1} + A_{k+1}^C C_{k+1} + A_{k+1}^{C'} C_{k+1}' + A_{k+1}^L L_{k+1}\} \quad (16)$$

where the right hand sides of (6)-(8) are actually substituted for the P_{k+1} , C_{k+1} and L_{k+1} . The boundary conditions are $A_N^P = I$, $A_N^C = 0$, $A_N^{C'} = 0$ and the generalized Euler-Lagrange equations lead after some algebra to the (backward) recursions

$$A_k^L = P' A_{k+1}^L P + Y^2 P' A_{k+1}^L P \quad (17)$$

$$A_k^C = H' K_{k+1}^L A_{k+1}^L P + (P - K_{k+1} H) A_{k+1}^C P + Y^2 H' K_{k+1}^L A_{k+1}^L P \quad (18)$$

$$A_k^P = H' K_{k+1}^L A_{k+1}^L K_{k+1} H + (I - K_{k+1} H) A_{k+1}^C K_{k+1} H + H' K_{k+1}^L A_{k+1}^L (P - K_{k+1} H) + (P - K_{k+1} H) A_{k+1}^P (P - K_{k+1} H) + Y^2 H' K_{k+1}^L A_{k+1}^L K_{k+1} H \quad (19)$$

$$\text{where } \bar{A} = (A^L - A^C - A^{C'} + A^P) \circ W W^* \quad (20)$$

The optimality condition $\partial \bar{H}_k / \partial K_k = 0$ yields finally the gain

$$K_k^{\text{opt}} = (A_{k+1}^P U)^{-1} [(A_{k+1}^P - A_{k+1}^{C'}) P P_k - (A_{k+1}^L - A_{k+1}^C + Y^2 \bar{A}) P C_k] H' R_k^{-c} \quad (21)$$

where $U_{ij} = 1 + Y^2 (W W^*)_{ij}$.

It is clear that the potential benefit of an adjustment to the optimal gain is grossly offset by the computational burden in solving the above equa-

tions. Moreover, a two point boundary value problem (TPBVP) needs to be solved since the forward and the backward equations are coupled. The optimal gains K_k^{opt} as computed by (21) will therefore depend on N and should be denoted as K_k^N , $k = 0, \dots, N-1$. Hence in order to obtain the optimal least squares estimate at each time N , a new series of gains for $k=0$ to $k = N-1$ needs to be computed. A filter implementation with precomputed gains would require lots of memory (order N^2).

Rather than implementing the optimal scheme, a suboptimal floating point correction will be derived which is more straightforward to realize. Availability of the optimal solution remains beneficial to analyze the performance of the proposed schemes.

4. SUBOPTIMAL FILTERS

The stepwise optimization formulas of section 3 are used to obtain a minimum error covariance P_1 within the structure. The one-step gain K_0 is computed ($\text{diag } W W^* = I$) from (23) and the subscript 0 simply replaced by k . Thus

$$K_k^{\text{sub}} = \frac{1}{1+Y^2} P P_k H' R_k^{-c} - \frac{Y^2}{1+Y^2} P C_k H' R_k^{-c} \quad (22)$$

where of course R_k^C , P_k and C_k follow the recursions (6)-(8) and (13)-(14).

5. CONCLUSIONS

For the floating point implementation of linear filters the optimal precomputable gains are characterized as the solution to a nontrivial TPBVP. A computationally more tractable approximation is presented. It requires more computational effort than the Kalman filter, but offsets the degradation due to the finite wordlength. This disadvantage disappears since the gains are precomputable as opposed to the method in [8]. The methods and results described can be carried over to the regulator as well.

REFERENCES

1. C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits and Systems*, Vol. 23, No. 9, pp. 551-562, September 1976.
2. P. Moroney, "Issues in the Implementation of Digital Feedback Compensators," MIT Press, 1983.
3. J. H. Wilkinson, "Rounding Errors in Algebraic Processes," Prentice Hall, 1963.
4. D. E. Knuth, "The Art of Computer Programming," Vol. 2, Addison-Wesley, 1981.
5. J. S. Vandergraft, "Introduction to Numerical Computations," Academic Press, 1983.
6. R. E. Rink and H. Y. Chong, "Performance of State Regulator Systems with Floating-Point Computation," *IEEE Trans. AC*, Vol. 24, No. 3, pp. 411-421, June 1979.
7. A. J. M. Van Wingerden and W. L. DeKoning, "The Influence of Finite Wordlength on Digital Optimal Control," *IEEE Trans. AC*, Vol. 29, No. 5, pp. 385-391, May 1984.
8. E. I. Verriest, "On Redefining the Optimal Least Squares Filter under Floating Point Operations," *Proc. ICASSP 1984*, p. 30.9, March 1984, San Diego, CA.

This work was supported by the U.S. Air Force under contract F-08635-84-C-0273.

APPENDIX P

SUBSPACE CORRELATION MEASURES AND APPLICATIONS TO THE

STOCHASTIC REALIZATION PROBLEM

SUBSPACE CORRELATION MEASURES
AND APPLICATION TO
THE STOCHASTIC REALIZATION PROBLEM

Erik I. Verriest

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

(404) 894-2949

ABSTRACT

It has been shown in earlier papers that the canonical correlation analysis and the principal component analysis as applied to the stochastic realization problem can be derived under the unified framework of the RV-coefficient. This RV coefficient also provides a common statistical measure of information that can be used in the evaluation and comparison of the different methods. It is shown here that this RV-coefficient has a very natural interpretation in terms of the information structure (and not just the data) of the problem itself. More specifically, it can be derived from the generalized probability measures (Gleason measures) defined on the propositional system associated with the modeling problem. This problem draws some close analogy to the foundations of modern quantum theory.

Acknowledgement

This research is supported by the U.S. Air Force, under Contract Nos F08635-84-C-0273 and AFOSR-87-0308.

Introduction.

A fundamental problem in modelling, identification, signal processing, digital filtering and cluster analysis is that of finding a finite dimensional Markovian representation of a stochastic sequence from the covariance information. This stochastic realization problem has been widely discussed [A,F,B,DP,AK,RV]. Whenever a finite set of real data is gathered, all processing is performed over finite sets, and in most cases an underlying probabilistic model is absent. As a result, covariances must be estimated from the observed time series. A more direct, data driven approach is favorable. Moreover, for many applications, the Markovian representation or state space model may be too complex, due to a high dimensionality, thus barring efficient computational management. This motivates the quest for lower order models, and a common measure for the evaluation of the performance of the different modeling approaches.

Many methods exist for the determination of a stochastic realization. Two philosophies, deeply rooted in multivariate statistical analysis, are singled out. Akaike [A] and Faurre [F], among others, developed the theory based on the information interface between the past and the future of a time series. This led Desai and Pal [DP] to an algorithm for obtaining a stochastic realization and model reduction scheme, based on the Canonical Correlation Analysis (CCA). Their realizations form the counterpart to the deterministic balanced realizations of Moore [M]. Another method based on the Karhunen-Loeve Method (KLM), also known as the Principal Component Analysis (PCA) has been proposed by Arun and Kung [AK].

Ramos and Verriest [RV] unified the CCA and KLM methods by showing that they are both special cases of a more general optimization problem, using the RV-coefficient introduced in multivariate analysis by Escouffier [E]. Given two zero-mean random vectors x and y , with $\text{Cov} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{xy} & R_{yy} \end{bmatrix}$,

the RV-coefficient is defined by

$$RV(x,y) = \frac{\text{Tr}[R_{xy}R_{yx}]}{(\text{Tr}[R_{xx}^2]\text{Tr}[R_{yy}^2])^{1/2}} \quad ; \text{ where } R_{yx} = R_{xy}' \quad (1)$$

It was shown that this common statistical measure of information provides a rationale for drawing inferences about the performance of the algorithm.

It further unifies the exact covariance and real data case by relating this RV-coefficient to certain operators in a tensor product space $G \otimes H^*$ where G and H are separable Hilbert spaces $[V]$. Here G is the base space, and H is respectively $L_2(\Omega, B, m)$ and R^N .

In this paper, the geometry of the stochastic realization problem, both exact and approximate, is investigated, and it is linked to some notions in the theoretical foundations of quantum mechanics. More precisely, measures on the subspaces of a Hilbert space are introduced, which relate to the density matrices in the quantum mechanical context. It is then shown that the above mentioned RV-coefficient is but one possible measure. Many others can be defined, leading of course to slightly different results. We will not emphasize the algorithmic solution of the realization problem, to which many fine researchers have substantially contributed. A motivation for the use of the RV coefficient in the stochastic realization problem is presented.

Philosophy of Stochastic Realization Theory

As originally formulated, the stochastic realization theory deals with the following problem (for a full mathematical statement, see e.g. [F]): Given a process with covariance sequence R_k , find a Markovian representation or state space model of minimal dimension, that generates an output with the given covariance function. This problem presupposes the knowledge of the exact covariance sequence. A perhaps more realistic formulation is: Given

an observed sample path of a stochastic process, find a realization of minimal dimension which generates an output with the same covariance as the observed sequence. There are thus two parts: on one hand, there is the realization of the model given the exact covariance sequence of the given process. This will be referred to as the exact problem. There also is the problem of determining or estimating the covariance sequence from the observed data sequence. Stochastic realization theory is therefore a statistical theory. The problem with this is that many observed processes occur only once, and that one thus can hardly speak of a statistical model with repeated sample paths. It is our contention to reformulate the realization problem from the point of view of observed data. In order to do so, we rely on some of the same principles that guided physicists to the logical foundations of modern quantum mechanics [P,BV,Gn,Gd,Va], which is founded on the impossibility of certain knowledge. Although we will not discuss deterministic realization theory, it is worth mentioning that a similar approach can be taken here, leading to the usual Boolean logic, also pertinent in the logical foundations of classical mechanics.

The Propositional System.

We will depart slightly from the usual setup of the problem. First of all, only the observed time series is available to the modeler. Hence, any notion of "state" should refer somehow to the observed data sequence $y_1, y_2, \dots, y_k$. As more things are observed with time, the information grows. More possibilities will appear ("splittings" in previously unresolvable information). We shall define the state of the system as the entity representing the maximal information that possibly can be given about the system. In order to build this up axiomatically, from "first principles", we are led to consider the fundamental idea of a question, defined as

"every experiment leading to an alternative of which the terms are only Yes or No." (e.g. "the value obtained by the time series at time i lies in the Borel set B "). All information one has about the system is captured in the totality of all such elementary questions. The problem is that there are very many such questions to ask, and unless we have a good theory about these questions and the relations among them, this framework would be of no value. We say that the question is certain (true) if the truth or falseness of its answer can be stated with absolute certainty. In the situation that whenever the system question Q_1 is true, we have the property that question Q_2 is always true, we say Q_1 implies Q_2 , and write $Q_1 < Q_2$. If we define two questions as equivalent if $Q_1 < Q_2$ and $Q_2 < Q_1$, and denote the resulting equivalence class by $[Q_1]$, then the implication is a partial ordering on the quotient set. The equivalence classes of questions will be called propositions. We can also make new questions from old ones, by introducing the greatest lower bound (GLB) and largest upper bound (LUB) of any two questions. The greatest lower bound of a family of questions Q_i is defined as the question $\bigwedge_i Q_i$ such that $\bigwedge_i Q_i$ true means that in the event of the measurement (verification) of an arbitrary one of the Q_i 's, the result "Yes" is certain. The least upper bound $\bigvee_i Q_i$ of the family of questions is then the greatest lower bound of all questions $S_j > Q_i \quad \forall i$. This induces a greatest lower bound and a least upper bound on the set of propositions. There exist a minimal proposition 0 (always false) and a maximal one, 1, which is always true. It is important to point out that for the stochastic problem exact knowledge of the truth of the proposition $y_1 = 1$, does not allow to say anything with absolute certainty of the output y_2 . One says that the propositions regarding y_1 are incompatible with the propositions regarding y_2 . The greatest lower bound is therefore 0, i.e.

the false statement. Propositions regarding the output at one particular time are compatible (simultaneously verifiable). On the other hand, the propositional system for a deterministic system allows propositions of the form $(y_1=1, y_2=2)$. Furthermore, if Q is a question, then the question that leads to the reverse truth statements is supposed to exist as well, and is called the orthocomplement.

At a more abstract level, a partially ordered set $\mathcal{L}$ is a complete lattice, if each subset of $\mathcal{L}$ admits a GLB and a LUB, which belongs to $\mathcal{L}$. An orthocomplementation in a lattice is a map $\#: \mathcal{L} \rightarrow \mathcal{L} : p \rightarrow p^\#$ such that

$$\begin{aligned} \text{i)} \quad & (p^\#)^\# = p \quad \forall p \in \mathcal{L} \\ \text{ii)} \quad & p \wedge p^\# = 0 \quad ; \quad p \vee p^\# = I \quad \forall p \in \mathcal{L} \\ \text{iii)} \quad & p < q \quad \longrightarrow \quad q^\# < p^\# \end{aligned} \tag{2}$$

Examples of a complete orthocomplemented lattice are the power set of a given set, and (more generally) any σ -algebra $\mathcal{B}$ as for instance used in the Kolmogorovian probability theory. Note that this σ -algebra is isomorphic to the set of propositions that can be made about the probabilistic events. A lattice is weakly modular if $p < q \longrightarrow q \wedge (q^\# \vee p) = p$. If $p < q$, one says that q covers p if $p < x < q \longrightarrow x = p$ or $x = q$. Elements which cover 0 are called atoms. A lattice is atomic if $\forall p \neq 0$, there exists an atom $a < p$. A proposition system or logic is then abstractly defined as a complete orthocomplemented, weakly modular and atomic lattice. More details can be found in the literature (e.g. [BV,P]).

In the above example of a σ -algebra, the distributive property $S_1 \wedge (S_2 \vee S_3) = (S_1 \wedge S_2) \vee (S_1 \wedge S_3)$ holds. A complete orthocomplemented distributive atomic lattice is called a classical (Boolean) proposition system. The usual Aristotelean logic reigns in such a system. The propositional system (propositions about subsets of phase space) of a

subspaces of a Hilbert space?

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 $\text{Proj } H$ of all orthoprojectors on H . Gleason [G] has shown that in a separable Hilbert space of dimension at least three, every measure μ on the closed subspaces can be represented by

$$\mu(A) = \text{Tr } (TP^A) \quad (3)$$

where P^A is the projection operator on the subspace A of H , and T is a positive definite operator of trace class. Geometrically, these measures arise as limits of convex combinations of "elementary" measures of the form

$$\mu_v(A) = \| P^A(v) \|^2 \quad ; \quad v \in H \quad (4)$$

This notion has been extended by Jajte [J]. In particular it has been shown that every (vector-valued) Gleason measure ξ on $\text{Proj } H$ can be extended in a unique way to a continuous operator on $L(H)$, the algebra of all bounded linear operators in H . An important class of Gleason measures taking values in a Hilbert space K are the Orthogonally Scattered Measures (OSG).

Definition: An OSG-measure is a mapping $\xi: \text{Proj } H \rightarrow K$ for which:

- i) For any sequence of pairwise orthogonal projectors $P_1, P_2, \dots$
from $\text{Proj } H$

$$\sum_1 \xi P_i = \xi \left(\sum_1 P_i \right) \quad (5)$$

the series on the left hand side being weakly convergent,

- ii) For any orthogonal projectors P, Q in $\text{Proj } H$

$$P \perp Q \rightarrow \xi P \perp \xi Q \quad (6)$$

in an interval J , then nothing changes this fact in the future (because of causality). So from time 1 on, the statistical subensemble for which $y_1 \notin J$ has been filtered out. The axiomatic foundations of quantum theory led now to a suitable representation of such logics. An arbitrary propositional system can be decomposed as the direct union of irreducible propositional systems, and it is known [Va] that every irreducible propositional system can be realized by the lattice $\text{Proj}(H)$ of all closed linear subspaces of a Hilbert space.

Generalized Probability Measures.

The goal of this section is to calculate the probability of obtaining an answer "yes" for an arbitrary proposition of the system, "prepared in the state" determined from the available data. The terminology "prepared in a certain state" is standard for physicists. In our context, it simply means that we look at an ensemble of systems whose prior information set is identical to the observed time series; i.e. a filtered ensemble.

If the underlying propositional system is a Boolean σ -algebra, then the Kolmogorovian probability theory gives a consistent definition for the probability measures. They are defined on a measurable space $(\Omega, \mathcal{B})$. We illustrated that since the probabilities are actually defined on $\mathcal{B}$, it may not be necessary to invoke Ω at all. $\mathcal{B}$ is the consistent collection of (note that the σ -finiteness axiom cannot be empirically deduced, but is brought in for mathematical convenience) logical statements that can be made about the physical system. It is possible to capture the classical probability theory as a theory of measures on Boolean algebras. Here is the problem: We want to make consistent probabilistic statements, but, we lost the Boolean structure of subsets of a set. How do we generalize a probability so it can be defined on a non-Boolean logic, e.g. the logic of

classical mechanical system has this property. In fact, this is also the logic underlying the relations among the propositions of the probabilistic events in the Kolmogorov sense. By the Loomis representation theorem, there exists then a set Ω such that $\mathcal{L} = \mathcal{B}(\Omega)$, i.e. the propositional system is isomorphic to a set of subsets of some set.

Whereas the Kolmogorovian probability starts from a set of possible outcomes, on which a σ -algebra of events is described, in the above alternate way, the σ -algebra (read: propositional system) comes first. Now this latter viewpoint leads to the right generalizations. However a more general logic $\mathcal{L}$ may fail to be distributive, so that the propositional system can no longer be isomorphic to the lattice of subsets of a set. Yet, even in such a case, one would like to define a reasonable notion of "probability" for the events (propositions). This can be done in a consistent way, generalizing classical probability [Gu].

It is already clear from our remarks regarding the compatibility that the logic of a stochastic system and the logic of a deterministic system are quite different. In the first the distributive property fails to hold in general. To illustrate this, consider the propositions $S_1 = "y_1 > 0"$ and $S_2 = "y_2 > 0"$, and let $S_2^\#$ be its complement. Both propositions are meaningful, the composite proposition $S_2 \vee S_2^\#$, which is the trivial proposition 1 is meaningful as well. Then $S_1 \wedge (S_2 \vee S_2^\#)$ is simply S_1 , but since S_1 and S_2 are incompatible, $(S_1 \wedge S_2) \vee (S_1 \wedge S_2^\#)$ has no meaning. The state of a system is represented by an atom (i.e. a proposition that is only implied by the always false proposition 0) in the propositional system. Conversely if we are given all the propositions true for the system, then the state is defined as the greatest lower bound of these propositions. Conditioning is related to the notion of a filter. This means that if for instance at time 1 the output is observed, and found with absolute certainty

Any OSG defines a positive Gleason measure via

$$\mu P = \|\xi P\|^2 \quad ; \quad P \in \text{Proj } H \quad (7)$$

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

$$\xi P = \text{Tr } TP \quad ; \quad P \in \text{Proj } H \quad (8)$$

The above can be interpreted as a "variance". A "covariance" can be defined by $\text{COV}(P, Q) = \langle \xi P, \xi Q \rangle_K = \text{Tr } TPQ$. In fact, it can be shown that if H and K are real Hilbert spaces, then $\forall P, Q \in \text{Proj } H$

$$\langle \xi P, \xi Q \rangle_K = \text{Tr } TPQ = \text{Tr } TQP \quad (9)$$

where T is given by Gleason's theorem. In quantum mechanics, T is known as the density matrix.

Applications to Realization Theory

We assume now that we have N sample paths of length p of a stationary time series, and organize it in a data matrix $Y \in R^{p \times N}$. This entails no contradiction with our data-approach, as N shifted versions of the same observed data can be used by virtue of the (assumed) stationarity. In the abstract sense, we may also consider the exact realization problem. In this case, we shall work with random variables rather than data matrices. The underlying spaces are $L_2^p(\Omega, B, P)$ for the stochastic realization problem, and $R^{p \times N}$ for the real data case. These spaces are isomorphic respectively with the tensor product spaces

$$L_2^p(\Omega, B, P) = R^p \otimes L_2^p(\Omega, B, P) \quad (10)$$

$$R^{p \times N} = R^p \otimes R^N \quad (11)$$

More generally, if (ψ_i) is a complete orthonormal set in H , then any vector x in the tensor product space $G \otimes H^*$ (a space of operators $H \rightarrow G$) of the form

$$x = \sum |x_i\rangle \langle \psi_i| \quad (12)$$

where (x_i) is a family of "data"-vectors $\in G$, and ψ_i an "evaluation"-vector (e.g. it picks out the i -th sample). The bra-ket notation is used since it seemed to be the most clarifying notation. The vector x will be referred to as the "conditioning". The inner product in $G \otimes H^*$ is

$$\begin{aligned} \langle x, y \rangle &= \langle \sum_i |x_i\rangle \langle \psi_i|, \sum_j |y_j\rangle \langle \psi_j| \rangle \\ &= \sum_{ij} \langle x_i | y_j \rangle_G \langle \psi_i | \psi_j \rangle_H \\ &= \sum_i \langle x_i | y_i \rangle_G \end{aligned} \quad (13)$$

Let μ_i be the Gleason measures corresponding to x_i , i.e. $\mu_i(A) = \|P^A(x_i)\|^2$. Introduce now a superposition of measures on $\text{Proj } G$ induced by this prior.

$$\mu_x = \sum_i \mu_i \quad (14)$$

For all subspaces A of G , it follows that

$$\mu_x(A) = \text{Tr } T_x P^A \quad (15)$$

where $T_x = \sum_i |x_i\rangle \langle x_i| = xx'$ is interpreted as an (unweighted) gramian or covariance operator. The measure $\mu_x(A)$ gives a numeric value to the closeness of A to G , given the prior x . In a quantum mechanical context, it would read as the (unnormalized) expectation value of the observable represented by the projection operator P^A , when the system is prepared in the state x .

The problem of determining 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. Aragon and Couot [AC] also stated several equivalent problems relating to the PCA (KLM). This did not lead to a useful consistent covariance measure for two orthogonal subspaces of G.

The Canonical Correlation Analysis, and an alternative derivation of the Principal Component Analysis are obtained as follows:

For each $B \in \text{Proj } G$, define the operators

$$\hat{P}^B: G \otimes H^* \longrightarrow G \otimes H^* \quad (16)$$

$$\begin{aligned} \hat{P}^B_x &= \sum_i \hat{P}^B |x_i\rangle\langle\psi_i| \\ &= \sum_i (P^B |x_i\rangle) \langle\psi_i| \end{aligned} \quad (17)$$

Note that, $\forall x \in G \otimes H^*$,

$$\begin{aligned} (\hat{P}^B)^2(x) &= \hat{P}^B (\sum P^B |x_i\rangle\langle\psi_i|) \\ &= \sum (P^B)^2 |x_i\rangle\langle\psi_i| \\ &= \sum P^B |x_i\rangle\langle\psi_i| \\ &= \hat{P}^B(x) \end{aligned} \quad (18)$$

Thus $\hat{P}^B$ is a projection operator. Define further

$$\xi_x: \text{Proj } G \longrightarrow G \otimes H^* : \xi_x(B) = \hat{P}^B(x) \quad (19)$$

which is a vector valued (in $G \otimes H^*$) measure (operator valued if you wish) satisfying:

i) For any set (P_i) of pairwise orthogonal subspaces

$$\begin{aligned} \sum_i \xi_x(P_i) &= \sum_{i,j} P_i |x_j\rangle\langle\psi_j| \\ &= \xi_x(\sum_i P_i) \end{aligned} \quad (20)$$

ii) If $P \perp Q$, then $\xi_x(P) \perp \xi_x(Q)$ (21)

Hence, ξ_x is an OSG-measure. Then by (7), there exists a positive measure

μ_x such that

$$\mu_x = \|\xi_x(P)\|_G^2 = \sum_i \|P|x_i\rangle\|^2 = \text{Tr } PT_x P \quad (22)$$

This corresponds to a "coherent" addition of OSG measures, conditioned on x . (i.e. a posterior measure). It stems from the fact that $T_x : G \rightarrow G$ is a characteristic for the given x in $G \otimes H^*$, (in fact, a "sufficient" statistic), and one can think of T_x (or μ_x) as being conditioned by the $x \in G \otimes H^*$. The posterior variance operator for subspace A , and covariance of A and $B \in \text{Proj } G$, given x is then the operator $G \rightarrow G$ given by

$$(\hat{P}_x^B)(\hat{P}_x^A)' = \sum_i P^B|x_i\rangle\langle x_i|P^A = P^B T_x P^A \in G \otimes G^* \quad (23)$$

This is simply interpreted as the restriction to B of the range of the map $T_x|_A$ (restriction of T_x to A), and displays the coupling of the interface between A and B given x . If a norm $\|\cdot\|$ is chosen on the space of operators $G \otimes G^*$, a scalar covariance measure can be associated to this (co)variance operator. It naturally follows that the correlation between the subspaces A and B in $\text{Proj } G$ is

$$\rho(A, B|x) = \|P^A T_x P^B\| / (\|P^A T_x P^A\| \|P^B T_x P^B\|)^{1/2} \quad (24)$$

In particular, the Frobenius norm leads to ρ_F , which is the RV coefficient (1) for the empirical data, used in the realization context in [R,RV] and motivated in [V]. other norms (e.g. spectral) may be taken as well. In the time series analysis, let the observed data be organized in a data matrix $X \in \mathbb{R}^{p \times N}$, and let $\{\phi_i\}$ be the standard orthonormal basis in $\mathbb{R}^p$. Then $T_x = XX'$ is the sample covariance matrix S , and the correlation between the complementary subspaces $K = \text{span}(\phi_1, \dots, \phi_k)$ and $K^\perp = \text{span}(\phi_{k+1}, \dots, \phi_p)$ is (for the obvious partition of S)

$$\rho_F = \text{Tr}(S_{12}S_{21}) / (\text{Tr}S_{11}^2 \text{Tr}S_{22}^2)^{1/2} \quad (25)$$

The PCA (KLM) and CCA are now both optimization problems, but with respect to different constraints. With data transformations M and N on K and $K^\perp$ respectively (considered as past and future in the Markovian modeling problem), corresponding to the global transformation operator

$$L = M P^K + N (P^K)^\perp : G \rightarrow G \quad (26)$$

these problems are formulated as ($O(K)$ is the orthogonal group on K)

$$\text{PCA} : \max_{M \in O(K)} \text{Tr} (M T_{12} T_{21} M') / (\text{Tr}(M T_{11} M')^2 \text{Tr}(T_{22})^2)^{1/2} \quad (27)$$

$$\text{CCA} : \max_{\substack{M \in O(K) \\ N \in O(K^\perp)}} \text{Tr} (M T_{12} N N' T_{21} M') / (\text{Tr}(M T_{11} M')^2 \text{Tr}(N T_{22} N')^2)^{1/2} \quad (28)$$

It is shown in [RV] that the above problems are equivalent to a generalized singular value decomposition. This approach is the "standard" one of [AK,DP]. It also relates to the procrustes problem [GV, p. 426]; i.e. find an orthogonal matrix Q such that the Frobenius norm $\|A - QB\|$ is minimal for given A and B .

Note that T is a covariance in the exact realization theory, while here it corresponds to the sample covariance, respectively for $G \otimes H^* = L_2^P$ or R^{PN} . Discriminant analysis and a rational way for discarding variables in multivariate statistics can also be treated in this way. The use of more general probability measures in pattern recognition has already been explored by Watanabe [W].

Finally, modifications can be made to minimize the "endeffects" due to substitution of zeros where data is missing in the time series, by using a weighted linear superposition of states.

Conclusions and Outlook

The RV-coefficient, used successfully in the unification of various stochastic realization approaches, has been linked to the generalized measures defined on the subspaces of a Hilbert space. The logics defined on these spaces are similar to the representations of the logics that occurring in quantum mechanics and allow for the definition of generalized measures. More importantly, we have indicated that a similar approach can be used in the realization problems. We have only discussed some details for the stochastic realization, which has the same structure as a purely quantal system in physics. The application of this logico-algebraic approach in the deterministic realization theory is pursued elsewhere. We only mention that in this case a usual Boolean (or classical) logic results, and a theory of approximation and modeling can be based on set theoretic measures. The quantum "probability" is different from the Kolmogorov probability. The "logic" is not the one of subsets (Boolean lattice), but the logic of subspaces. The first is a special case: the set of orthogonal subspaces spanned by subsets of a complete orthonormal basis form a distributive logic. How can we further reconcile the analogy with quantum mechanics? Quantum experiments are non-repeatable, i.e. identical realizations are not possible, whereas for identically prepared classical systems, identical propositions follow. Think of the analogy with an ensemble of noninteracting, identically prepared deterministic systems. By definition, each of these gives necessarily the same response, whereas this will not be the case of a stochastic ensemble, because of the different inaccessible noises in each realization. Instead of working with an ensemble of realizations in parallel, we can work with a single realization, but serially in time, if we assume timeinvariant systems and stationary noises. Roughly speaking we then have the following connections:

Kolmogorov measures	-----	Gleason Measures
Classical Mechanics	-----	Quantum Mechanics
Deterministic Realizations	-----	Stochastic Realizations

References

- [A] Akaike, A, Markovian representation of Stochastic Processes by Canonical Variables, SIAM J. Control, 13, No. 4, 1975.
- [AC] Aragon, Y and Couot, J., Une Definition de l'Operateur d'Escoufier, Comptes Rendus, 283, series A, 867-869, 1976.
- [AK] Arun, K.S. and Kung, S.Y., A New Algorithm for Approximate Stochastic Realization, IEEE Intl. Symp. Circuits and Systems, Newport Beach, CA, 1983
- [B] Baram, Y, Realization and Reduction of Markovian Models from Nonstationary Data, IEEE Trans. Auto Ctr., Vol. 26, No.6, 1981.
- [BV] Birkhoff, G. and Von Neumann, J., The Logic of Quantum Mechanics, Annals of Mathematics, Vol. 37, No. 4, 1936, pp. 823-843.
- [DP] Desai, U.B. and Pal, D., A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations, Proc. 21st IEEE Conf. Decision and Control, 1982.
- [E] Escoufier, Y., Le Traitement des Variables Vectorielles, Biometrics 29, 1973.
- [F] Faurre, P.L., Stochastic Realization Algorithms, in System Identification, Mehra and Lainiotis, eds., Academic Press 1976.
- [G] Gleason, A. M., Measures on the Closed Subspaces of a Hilbert Space, J. Math. Mech. 6, 1957.
- [Gn] Guenin, M., Axiomatic Foundations of Quantum Theories, J. Math. Phys., Vol. 7, No.2, pp.271-281, 1966.
- [Gu] Gudder, S.P., An Extension of the Classical Measure Theory", SIAM Review, Vol. 26, No.1, pp. 71-89, 1984.
- [GV] Golubi, G.H. and Van Loan, C.F., Matrix Computations, Johns Hopkins University Press, 1983
- [J] Jajte, R, Gleason Measures, in Probabilistic Analysis and Related Topics, Vol. 2, Academic Press, 1979.
- [M] Moore, B.C., Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, IEEE Trans. Automatic Control, Vol. 26, No. 5, 1981.
- [P] Piron, C., Foundations of Quantum Physias, W.A. Benjamin, Inc, 1976.
- [R] Ramos, J.A., A Stochastic Realization and Model Reduction Approach to Streamflow Modeling, Ph.D. thesis, School of Civil Engineering, Georgia Institute of Technology, Atlanta, Georgia 1985.
- [RV] Ramos, J.A. and Verriest, E.I., A Unifying Tool for Comparing Stochastic Realization Algorithms and Model Reduction Techniques, Proc. 1984 ACC, San Diego.
- [Va] Varadarajan, V.S., Probability, in Physics and a Theorem on Simultaneous Observability, in the Logico-Algebraic Approach to Quantum Mechanics, Hooker, edt. pp. 171-203.
- [V] Verriest E.I., Projection Methods for Linear Model Reduction, in Modelling, Identification and Robust Control, Proc. 1985 MTNS, Stockholm Sweden.
- [W] Watanabe, S., Modified Concepts of Logic, Probability, and Information Based on Generalized Continuous Characteristic Function, Information and Control, Vol. 15, pp. 1-21, 1969.

APPENDIX Q

A UNIFIED RV-COEFFICIENT APPROACH FOR SOLVING THE COVARIANCE
BASED STOCHASTIC REALIZATION PROBLEM

A UNIFIED RV-COEFFICIENT APPROACH FOR SOLVING THE
COVARIANCE BASED STOCHASTIC REALIZATION PROBLEM

José A. Ramos
United Technologies Optical Systems, Inc.
Optics and Applied Technology Laboratory
P.O. Box 109660
West Palm Beach, Florida 33410-9660
(305)-863-4271

Erik I. Verriest
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
(404)-894-2949

Srinivas G. Rao
Seaburn and Robertson, Co.
5510 Gray Suite 118
P.O. Box 23184
Tampa, Florida 33623
(813)-877-9182

ABSTRACT

The multivariate time series identification problem is approached in this paper from a canonical variate analysis point of view. Two different but related problems are extensively studied, namely the generalized symmetric and generalized unsymmetric stochastic realization problems. These are associated with the problem of finding linear transformations (basis vectors) of the forward and/or backward predictor spaces of a second-order stationary vector stochastic process. These basis vectors correspond to the states of a forward and/or backward Kalman filter models, respectively. A new form of solution is presented which provides a unified framework for solving these two related problems and, in addition, motivates algorithmic development. This unified framework, known as the RV-coefficient approach, is used to generalize previously known results in stochastic realization theory and to generate several new others. In particular, it is shown that the canonical realization algorithm and the Karhunen - Loève method, which solve the symmetric and unsymmetric stochastic realization problems, respectively, can be derived under this unified framework. More importantly, the RV-coefficient provides a common statistical measure of information that can be used as a tool for comparing performance between algorithms and for obtaining appropriate reduced - order models. The normalized balanced realizations found in deterministic realization theory are extended to the stochastic case and shown to have some optimality properties in the RV-coefficient sense. Also, the problem of transforming a given pair of forward-backward innovations representations to a certain canonical form (coordinate-free) is shown to be, in the RV-coefficient sense, of the same format as that of the stochastic realization problem.

I. INTRODUCTION

The time invariant stochastic realization problem can be defined as the problem of finding a finite dimensional Markovian representation (state-space model) from knowledge of the autocovariance sequence of a second-order stationary stochastic process. This problem has received a great deal of attention in the recent past due to its fundamental importance in system identification, digital filtering, signal processing, and time series modeling [1] - [11]. In many applications, the Markovian representation or state-space model may be computationally unmanageable due to its high dimension. This may be caused by the introduction of superfluous state components due to noise perturbations in the covariance structure of the stochastic process as a result of roundoff errors, inexact covariance estimates, etc.. The solution then calls for an approximate or reduced-order model, which may be obtained directly from the solution to the stochastic realization problem, provided it yields a coordinate-free representation [7] - [11].

Faurre [4] has clarified the algorithmic aspects of the stochastic realization problem by characterizing the set of all possible Markovian representations in terms of extreme point or canonical representations. This canonical structure was further extended by Akaike [3] [12], [13], who has developed a stochastic realization theory based on the information interface between the past and future of a stochastic process and the concepts of predictor spaces and canonical variables. This theory has been fundamental to modern stochastic realization algorithms, which, in an optimal way, attempt to approximate the information interface between the past and future of the stochastic process.

To fix the ideas, let us mathematically define the stochastic realization problem as follows. Given a zero mean, rational, discrete-time, stationary, vector stochastic process $\{y_k\}$, find a Markovian representation of the form

$$x_{k+1} = Fx_k + w_k \quad (1a)$$

$$\bar{y}_k = Hx_k + v_k \quad (1b)$$

such that it has minimal dimension (n) and the output (1b) generates the same autocovariance function as that of the process $\{y_k\}$. Here x_k is the ($n \times 1$) state vector process, w_k and v_k are respectively ($n \times 1$) and ($m \times 1$) zero mean white Gaussian noise processes, $\bar{y}_k$ is the ($m \times 1$) output vector, and the parameter matrices F and H are of appropriate dimensions. Furthermore, the noise processes w_k and v_k have the following joint covariance structure:

$$E \left\{ \begin{bmatrix} w_k \\ v_k \end{bmatrix} \begin{bmatrix} w_s^T & v_s^T \end{bmatrix} \right\} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{ks} = \Sigma_{wv} \delta_{ks} \quad (2)$$

where δ_{ks} is the Kronecker delta function, E is the expectation operator, A^T denotes the matrix transpose of A , and Q , S , and R are constant matrices of appropriate dimensions such that $Q \geq 0$, $R \geq 0$, and $\Sigma_{wv} \geq 0$, where $A \geq 0$ (≥ 0) refers to a matrix A which is positive definite (positive semi-definite). In addition, (1) is Markov with forward propagation property

$$E\{x_k w_s^T\} = E\{x_k v_s^T\} = 0, \quad s \geq k \quad (3)$$

and output autocovariance function given by

$$\Lambda(k) = [HF^{k-1}G]1_{(k)} + [G^T(F^{-k-1})^T H^T]1_{(-k)} + [\Lambda(0)]\delta_{k0} \quad (4a)$$

where

$$1_{(k)} = \begin{cases} 1 & \text{if } k > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4b)$$

$$G = F\bar{E}H^T + S \quad (4c)$$

$$\Lambda(0) = H\bar{E}H^T + R \quad (4d)$$

$$\bar{E} = F\bar{E}F^T + Q \quad (4e)$$

and $\bar{E} = E\{x_k x_k^T\}$ is the $(n \times n)$ positive definite state covariance matrix.

The stochastic realization problem then amounts to identifying a triple $(F, G, H)_n$ of minimal dimension n and covariance matrices $(\bar{E}, Q, R, S)$ such that (2) - (4) are satisfied. The motivation for solution is due to Akaike [3], [12], [13] and follows.

Let the past and future of $\{y_k\}$ be defined, respectively, by

$$Y_k^- = \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, \quad Y_k^+ = \begin{bmatrix} y_k \\ y_{k+1} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad (5)$$

and let the block Hankel matrix, H , along with the respective past and future covariance matrices, R^- and R^+ , be defined as [8]

$$H = E\{Y_k^+(Y_k^-)^T\} = \begin{bmatrix} \Lambda(1) & \Lambda(2) & \Lambda(3) & . & . \\ \Lambda(2) & \Lambda(3) & \Lambda(4) & . & . \\ \Lambda(3) & \Lambda(4) & \Lambda(5) & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \quad (6)$$

$$R^- = E\{Y_k^-(Y_k^-)^T\} = \begin{bmatrix} \Lambda(0) & \Lambda(1) & \Lambda(2) & . & . \\ \Lambda^T(1) & \Lambda(0) & \Lambda(1) & . & . \\ \Lambda^T(2) & \Lambda^T(1) & \Lambda(0) & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \quad (7)$$

$$R^+ = E\{Y_k^+(Y_k^+)^T\} = \begin{bmatrix} \Lambda(0) & \Lambda^T(1) & \Lambda^T(2) & . & . \\ \Lambda(1) & \Lambda(0) & \Lambda^T(1) & . & . \\ \Lambda(2) & \Lambda(1) & \Lambda(0) & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \quad (8)$$

Now define, respectively, the forward and backward predictor spaces of $\{y_k\}$ as

$$X_k = \text{span} [Y_k^+ | Y_k^-] = \text{span}[H(R^-)^{-1}Y_k^-] \quad (9a)$$

$$Z_{k-1} = \text{span} [Y_k^- | Y_k^+] = \text{span}[H(R^+)^{-1}Y_k^+] \quad (9b)$$

where, for zero mean random vectors a and b , $[a|b] = E\{ab^T\}E\{bb^T\}^{-1}b$ denotes the orthogonal projection of a onto the Hilbert space of random variables spanned by the components of b .

The predictor spaces (9) have infinite components, however, Akaike [3], [13] has shown that the states of a forward and backward Markovian representation are finite dimensional basis vectors of the predictor spaces. Thus, finding a basis respectively for (9a) and (9b) is a matter of finding linear transformation matrices A and B such that the transformed vectors (basis vectors of dimension n), called the state vectors, x_k and z_k , are orthogonal, i.e.,

$$x_k = A^T H(R^-)^{-1} Y_k^- \quad (10a)$$

$$z_{k-1} = B^T H^T(R^+)^{-1} Y_k^+ \quad (10b)$$

with orthogonality property defined by

$$E\{x_k x_k^T\} = \Delta_x = \text{diag} [\delta_{x1}, \delta_{x2}, \dots, \delta_{xn}] \quad (11a)$$

$$E\{z_{k-1} z_{k-1}^T\} = \Delta_z = \text{diag} [\delta_{z1}, \delta_{z2}, \dots, \delta_{zn}] \quad (11b)$$

Notice that z_k is the state of a backward Markov model which is dual to (1) and evolves in the opposite direction of time (see [14] - [16]). If we let $M^T = A^T H(R^-)^{-1}$ and $L^T = B^T H^T(R^+)^{-1}$, then the covariance based stochastic realization problem reduces to that of finding linear transformation matrices L and/or M such that (10a) and (11a) and/or (10b) and (11b) are satisfied.

Fortunately, the solution to this problem is a classical one in multivariate analysis and has led the way to recent stochastic realization algorithms. These algorithms have the added feature of implicitly solving the Riccati equations arising in an earlier algorithm due to Faurre [4], [57].

With this canonical structure, Akaike [13] extended the Ho-Kalman algorithm [17] to the stationary stochastic case, and Baram [11] accounted for the nonstationary stochastic processes generalization. Along the same lines, Desai and Pal [8] introduced the canonical realization algorithm (CRA) for balanced stochastic realizations (defined in the sequel), while Arun and Kung [9] developed the Karhunen-Loeve Method (KLM) for solving a one-sided problem (i.e., x_k or z_{k-1}). In an attempt to unify existing stochastic realization algorithms, Larimore [7] developed the generalized canonical variate method, which successfully breaks into CRA or KLM as specific cases. Other forms or variants of these algorithms have been given in [18] - [20]. These algorithms all have their grounds on some form of multivariate statistical analysis.

Earlier attempts to unify existing stochastic realization algorithms have partly failed due to a lack of a common statistical measure of information that can be used as a rationale for drawing inferences about the performance of the algorithms or for model reduction purposes. Meanwhile, existing measures may lead to results which differ in magnitude depending on the type of solution, and therefore, can create a problem of interpretation.

Escoufier [21] and Robert and Escoufier [31] introduced the RV-coefficient statistic as a tool for solving a large class of problems arising in multivariate statistical analysis. This solution approach, although it has not received much attention in the literature, shows future promise in stochastic realization theory as well as signal processing, pattern recognition, and discriminant analysis, to name only a few [10], [22], [23]. In this paper we

solve the covariance based stochastic realization problem from an RV-coefficient point of view and show that the solutions leading to different algorithms can all be put under this common framework of analysis.

Ramos [10] and Ramos and Verriest [30] showed that the RV-coefficient approach leads naturally to Akaike's stochastic realization theory [13]. In addition, it serves as a tool for algorithmic development and introduces a common statistical measure of information in standardized units (i.e., in the interval $[0,1]$), which can aid significantly in the interpretation of the results. This allows the modeler to compare several algorithms and models as well as model reduction techniques. Recently Verriest [22], [23] extended Escoufier's RV-coefficient to a geometrical framework which, based on certain operator valued measures, defines the correlation between subspaces of a Hilbert space. The approach is motivated by the procrustean problem [32] and leads to several different RV-type measures. Other measures of multivariate association between two sets of random variables are suggested in [33] - [35].

We begin the next section by reviewing the statistical measures used in existing stochastic realization algorithms and discuss their limitations. In Section 3, we solve the general, symmetric stochastic realization problem and its unsymmetric version, both from an RV-coefficient point of view, and further relate the results to the works of Desai and Pal [8] and Arun and Kung [9], respectively. We further introduce several other theoretical extensions and present a unification of existing stochastic realization algorithms. Section 4 treats the problem of transforming a given forward-backward pair of innovations representations into the canonical structure of the symmetric or unsymmetric stochastic realization problems, by taking a direct route via the RV-coefficient approach. Discussions of the present work and conclusions are contained in Section 5.

II. COMPARISON OF EXISTING PERFORMANCE MEASURES AND MOTIVATION FOR UNIFICATION

Since Hotelling's classic paper [24], the theory of canonical variate analysis has been fortified with several measures for quantitatively assessing the degree of association between two sets of random vectors. These measures have been categorized by Cramer and Nicewander [25] into two distinctive classes: redundancy measures and canonical correlation measures. The former type are associated with the predictability of one component with respect to the other, and thus form the basis for the methods of principal components of instrumental variables [26], external single set components analysis [27], and redundancy analysis [28], all of which seem to be equivalent. Canonical correlation measures on the other hand, are measures of multivariate association that attempt to extend the correlation coefficient to two sets of random vectors. These measures form the basis for the methods of canonical correlation analysis, multivariate analysis of variance, and multivariate regression [58]. Canonical correlation based measures are symmetric as opposed to redundancy measures which are asymmetric.

Recently, Larimore [7] introduced a generalized measure of multivariate association which, depending on the type of problem, acts either as a redundancy measure or a canonical correlation measure. For this reason, we classify it as a combined performance measure.

Several performance measures have been developed for use in stochastic realization theory, especially with the model approximation problem. These can also be classified under the above - mentioned categories, however, before we attempt to do so, we first need to look at the symmetry aspects of the

stochastic realization problem and the underlying form of the particular type of solution.

Recall that Y_k^- and Y_k^+ are respectively the past and future of the stochastic process $\{y_k\}$, whose joint covariance matrix is given by

$$E \left\{ \begin{bmatrix} Y_k^+ \\ Y_k^- \end{bmatrix} \begin{bmatrix} Y_k^+, Y_k^- \end{bmatrix}^T \right\} = \begin{bmatrix} R^+ & H \\ H^T & R^- \end{bmatrix} \quad (12)$$

Further recall that the stochastic realization problem is associated with the problem of finding transformation vectors (also canonical or state vectors) $x_k = M^T Y_k^-$ and/or $z_{k-1} = L^T Y_k^+$ such that these are orthogonal. In general, the simultaneous estimation of the transformation matrices M and L will lead to canonical correlation based measures. Here the joint covariance matrix of the state vectors is given by

$$E \left\{ \begin{bmatrix} z_{k-1} \\ x_k \end{bmatrix} \begin{bmatrix} z_{k-1}, x_k \end{bmatrix}^T \right\} = \begin{bmatrix} L^T(R^+)L & L^T H M \\ M^T H^T L & M^T(R^-)M \end{bmatrix} = \begin{bmatrix} \Delta_z & \Gamma \\ \Gamma & \Delta_x \end{bmatrix} \quad (13)$$

where $\Gamma = \text{diag} [\gamma_1, \gamma_2, \dots, \gamma_n]$ is an $(n \times n)$ diagonal matrix of squared canonical correlation coefficients between Y_k^- and Y_k^+ , $n = \text{rank} [H]$ corresponds to the dimension of the state vector, while the other terms are as defined previously. Condition (13) is known in multivariate analysis as the bi-orthogonality condition. In the stochastic realization problem this is equivalent to the problem having a symmetric solution, i.e., the state-space models characterized by z_{k-1} and x_k are dual to each other.

On the contrary, if one is interested in estimating M and L independently, i.e., as two separate problems, then the joint problem becomes unsymmetric and, in general, will lead to a pair of independent redundancy based measures. In this case, we have the following joint state covariance structure

$$E \left\{ \begin{bmatrix} z_{k-1} \\ x_k \end{bmatrix} \begin{bmatrix} z_{k-1} & x_k \end{bmatrix}^T \right\} = \begin{bmatrix} \Delta_z & L^T H M \\ M^T H^T L & \Delta_x \end{bmatrix} \quad (14)$$

As expected (14) lacks bi-orthogonality since the off-diagonal matrices are not diagonal, as a result of the independent orthogonalizations of z_{k-1} and x_k . As we will see later this lack of symmetry will lead to state vectors z_{k-1} and x_k that are not dual to each other as opposed to those obtained satisfying (13).

We now continue our specific discussion with a brief description of the different types of performance measures used in conjunction with the stochastic realization problem.

A. Canonical Correlation Measures

These measures are not, in general, associated with prediction, but rather with maximum correlation or similarity between two sets of random vectors. However, Yohai and Garcia Ben [36] have shown that solving

$$\min_M \rho(Y_k^+, M^T Y_k^-) = | E(Y_k^+ - \hat{Y}_k^+)(Y_k^+ - \hat{Y}_k^+)^T | = | R^+ | \prod_{i=1}^n (1 - \gamma_i) \quad (15)$$

subject to the orthogonality constraint $M^T(R^-)M = \Delta_x$, leads to a measure of prediction accuracy (or confidence) of Y_k^+ based on $x_k = M^T Y_k^-$, i.e.,

$$\hat{Y}_k^+ = E[Y_k^+ | x_k = M^T Y_k^-] = HM[M^T(R^-)M]^{-1}M^T Y_k^- \quad (16)$$

is the best linear predictor of Y_k^+ among all predictors of the form $\hat{Y}_k^+ = D^T x_k$, where D is any $(\infty \times n)$ matrix. Here we have taken $\Delta_x = I_n$ for convenience.

It should be noted, however that (15) involves solving the following generalized eigenvalue - eigenvector problem

$$M^T H^T (R^+)^{-1} H M = \Gamma \quad (17a)$$

$$M^T (R^-) M = \Delta_x = I_n \quad (17b)$$

However, if we minimize the anti-causal function $\rho(L^T Y_k^+, Y_k^-)$ with $\Delta_z = I_n$, then the solution involves a generalized eigenvalue-eigenvector problem dual to (17) whose solution is given by

$$\min_L \rho(L^T Y_k^+, Y_k^-) = |R^-| \prod_{i=1}^n (1-\gamma_i) \quad (18)$$

One can see that (15) and (18) are related to one another by the factor,

$$w = \prod_{i=1}^n (1-\gamma_i), \text{ which is called the alienation coefficient [24]. This}$$

coefficient which carries information from both the forward and backward models, is a measure of independence between Y_k^- and Y_k^+ . Therefore, (15) and (18) being a constant multiple of w , also reflect the independence between Y_k^- and Y_k^+ . Hence, there is practically no gain in information by solving (15) and/or (18) separately as in [36]. It has been shown in [37] that since (15 and (18)

maximize a determinant as opposed to a trace, they cannot be considered prediction measures. Hence, we classify them as canonical correlation measures. Desai and Pal [8] solved both problems simultaneously by performing a singular value decomposition of a weighted Hankel matrix, arriving at the following information based measure due to Gelfand and Yaglom [38].

$$I(L^T Y_k^-, M^T Y_k^+) = -\frac{1}{2} \sum_{i=1}^n \log(1-\gamma_i) \quad (19)$$

Notice that (19) is of the form, $a \log w$, where a is a constant, therefore, it also reflects the independence between Y_k^- and Y_k^+ [7]. Furthermore, when a canonical correlation coefficient is unity (15) and (17) are equal to zero, while (19) degenerates to infinity.

Another measure which has been in use for some time is [32]

$$\inf_{A: \text{rank}[A]=n} ||H - A||_F^2 = \gamma_{n+1} + \gamma_{n+2} + \dots + \gamma_{\infty} \quad (20)$$

where $|| \cdot ||_F$ denotes the Frobenius norm (i.e., $||A||_F = [\text{tr}(A^T A)]^{1/2}$). The matrix A represents the best n th order approximation to the Hankel matrix H . If one divides (20) by $||H||_F^2$, then we obtain

$$\phi(Y_k^+, M^T Y_k^-) = 1 - \frac{\sum_{i=1}^n \delta x_i}{q} \quad (23)$$

where δx_i is the variance of the i th state component x_k^i and corresponds to the i th largest eigenvalue of $HM[M^T(R^-)M]^{-1}M^TH^T$. The measure (23) then represents the proportion of variance unaccounted for by the state-space model with x_k as the n -dimensional state vector. Similarly, for estimating $z_{k-1} = L^T Y_k^+$, one can obtain

$$\phi(L^T Y_k^+, Y_k^-) = 1 - \frac{\sum_{i=1}^n \delta z_i}{p} \quad (24)$$

where $p = \text{tr}[R^-]$ and δz_i is the i th largest eigenvalue of $H^T L[L^T(R^+)L]^{-1}L^TH$. In general, when y_k is not a scalar or when $\Lambda(k) \neq \Lambda(-k)$, $\phi(Y_k^+, M^T Y_k^-) \neq \phi(L^T Y_k^+, Y_k^-)$ due to lack of symmetry in the individual solutions. In this case, the eigenvalues in (23) and (24) are different as opposed to those in (15) and (18) which are the common squared canonical correlation coefficients.

Notice that $(\delta x_i/q)$ in (23) represents the explained fractional variance of Y_k^+ by the i th state component x_k^i . The same holds true for (24) with respect to the variance of Y_k^- in terms of z_k^i . Thus, (23) and (24) are quantitative measures that account for the distribution of the variance of Y_k^+ and Y_k^- by their respective canonical components (state vectors). Unfortunately, this is not the case for canonical correlation based measures. In [10], [30] a redundancy index was derived from the canonical correlation solution in order to determine the individual contribution of the state components to the total variance of Y_k^+ and Y_k^- .

$$\delta(L^T Y_k^+, M^T Y_k^-) = \frac{\sum_{i=n+1}^{\infty} \gamma_i}{\sum_{i=1}^{\infty} \gamma_i} \quad (21)$$

which represents the degree of deterioration in the approximation. We have used here " ∞ " as an upper bound for H , R^- and R^+ , however, in practice one should use a finite upper bound.

B. Redundancy Based Measures

As mentioned earlier, redundancy measures are associated with the prediction efficiency of one of the components (Y_k^- or Y_k^+) with respect to the other. Arun and Kung [9] (see also [39] - [41]) computed $x_k = M^T Y_k^-$ by performing a one-sided Karhunen-Loeve expansion of the forward predictor space $X_k = \text{span}[Y_k^+ | Y_k^-]$. The criterion used was based on minimizing the following prediction efficiency measure

$$\psi(Y_k^+, M^T Y_k^-) = \text{tr}[R^+] - \text{tr}[H M (M^T (R^-) M)^{-1} M^T H^T] \quad (22)$$

For convenience of interpretation, we divide (22) by $q = \text{tr}[R^+]$, the total variance in Y_k^+ , to obtain

C. Combined Measures

The only measure found in this category is due to Larimore [27] and in our notation is defined as

$$\begin{aligned} \max \eta(Y_k^+, M^T Y_k^-)_{\Theta} &= E[(Y_k^+ - \hat{Y}_k^+)^T \Theta^{-1} (Y_k^+ - \hat{Y}_k^+)] \\ &= \alpha_{n+1} + \alpha_{n+2} + \dots + \alpha_{\infty} \end{aligned} \quad (25)$$

where α_i 's are the singular values of a well-defined matrix. When $\Theta = I_{\infty}$, (25) is a redundancy measure and when $\Theta = R^+$, it becomes a canonical correlation measure with α_i 's as the squared canonical correlation coefficients.

One can see that depending on the form of Θ , the magnitude of (25) can change drastically, therefore, presenting a problem of interpretation if one wants to compare both modeling approaches. Similarly, all other performance measures are derived based on one type of problem in mind (i.e., the symmetric or unsymmetric stochastic realization problem), and have no equivalent measure for the converse problem. Therefore, if we want to solve both types of stochastic realization problems by the different methods available, then we cannot compare the results simply because there is no common measure that yields the same units (i.e., variance, correlation, information, similarity, etc.). We are then faced with a problem of interpretation.

To overcome this difficulty, we will next present a unified measure of similarity, known as the RV-coefficient, which will allow us to unify previous algorithms under a common framework of analysis.

III. SOLUTION TO THE COVARIANCE BASED STOCHASTIC REALIZATION PROBLEM:

THE RV-COEFFICIENT APPROACH

A. Mathematical Preliminaries

Consider a particular sample realization of two zero mean stationary vector stochastic processes $\{x_k\}$ and $\{y_k\}$ of dimension $(p \times 1)$, whose squared Euclidean distance is given by

$$\begin{aligned} D^2(x_k, y_k) &= \|x_k - y_k\|^2 = E\{(x_k - y_k)^T (x_k - y_k)\} \quad \forall k \\ &= E\{x_k^T x_k\} - 2E\{x_k^T y_k\} + E\{y_k^T y_k\} \\ &= s_x^2 - 2s_{xy} + s_y^2 \end{aligned} \quad (26)$$

Replacing x_k and y_k by their respective normalized vectors, $\hat{x}_k$ and $\hat{y}_k$, it follows immediately that

$$D^2(\hat{x}_k, \hat{y}_k) = 2(1 - r_{xy}) \quad (27)$$

where r_{xy} is the correlation coefficient between x_k and y_k .

Now, suppose we collect all the information available in x_k and y_k for $k \in [1, N]$ and form the $(p \times N)$ data matrices X and Y (X and Y may have different row dimensions). Then X induces a configuration $C(X)$ of N points in R^p with relative distance matrix given by [31]

$$D^*(X) = \frac{S(X)}{[\text{tr}\{S(X)^2\}]^{1/2}} \quad (28)$$

where $S(X) = X^T X$. A similar expression for $D^*(Y)$ may be obtained from Y . These distance matrices are translation and rotation invariant and also invariant to global changes of scale. Then by making use of the scalar product $[\text{tr}(A^T B)]$ for square matrices A and B , and its induced norm, $||A|| = [\text{tr}(A^T A)]^{1/2}$ (notice that $||D^*(x)|| = ||D^*(Y)|| = 1$), the distance between the two configurations $C(X)$ and $C(Y)$ can be defined in a form similar to the relative distance between two vectors as

$$\begin{aligned} D^2[C(X), C(Y)] &= ||D^*(X) - D^*(Y)||^2 \\ &= 2 \left[1 - \frac{\text{tr}[S(X)S(Y)]}{[\text{tr}[S(X)^2]\text{tr}[S(Y)^2]]^{1/2}} \right] \\ &= 2 [1 - RV(X, Y)] \end{aligned} \quad (29)$$

where,

$$\begin{aligned} RV(X, Y) &= \frac{\text{tr}[S(X)S(Y)]}{[\text{tr}[S(X)^2]\text{tr}[S(Y)^2]]^{1/2}} \\ &= \frac{\text{tr}[S_{xy}S_{yx}]}{[\text{tr}[S_{xx}^2]\text{tr}[S_{yy}^2]]^{1/2}} \end{aligned} \quad (30)$$

Here, $S_{xx} = XX^T$, $S_{yy} = YY^T$, $S_{xy} = XY^T$, and $S_{yx} = YX^T$, are within a multiplicative factor of $1/N$, the covariances and cross-covariances between X and Y . Immediately we can see that (29) is a generalization of the distance between two vectors and $RV(X, Y)$ is a measure for the correlation between the configurations $C(X)$ and $C(Y)$.

The sample RV-coefficient shares some of the properties of a squared correlation coefficient, i.e., has values in $[0,1]$. The closer to 1 it is, the closer are the patterns and the better is Y (or X) as a substitute for $X(Y)$ in characterizing the sample space. Summarized below are other properties of the RV-coefficient [31], [42], [43]:

- i. $D[C(X), C(Y)]$ is a decreasing function of $RV(X,Y)$.
- ii. $RV(X,Y) = RV(Y,X)$
- iii. $RV(A^T X, Y) = RV(X, Y)$ when A is orthogonal
- iv. $RV(X, Y) = 0$ if and only if $XY^T = 0$
- v. $RV(X, Y) = 1$ if and only if $X = kY$ for some nonzero scalar k .
- vi. $RV(\alpha X, Y) = RV(X, Y)$ for some nonzero scalar α .

The RV-coefficient evolved thus from a geometrical interpretation as a measure for the comparison of subspaces. This bears some similarity to the relationship between subspaces of a given space, formalized using the singular value decomposition (SVD).

In particular, the orthogonal Procrustes problem [32] analyses the possibility of rotating a given data matrix X into another given data matrix Y (both of size $p \times N$). The precise problem is

$$\begin{aligned} & \underset{Q}{\text{minimize}} \quad ||X - YQ||_F \\ & \text{subject to } Q^T Q = I_N \end{aligned} \tag{31}$$

which is equivalent to maximizing $\text{tr}[XQ^T Y^T]$ subject to the orthogonality constraint on Q . The solution, in terms of the SVD of $Y^T X$, i.e.,

AD-A194 685

APPROXIMATIONS AND IMPLEMENTATIONS OF NONLINEAR
FILTERING SCHEMES(U) GEORGIA INST OF TECH ATLANTA
SCHOOL OF ELECTRICAL ENGINEERING A H HADDAD ET AL.

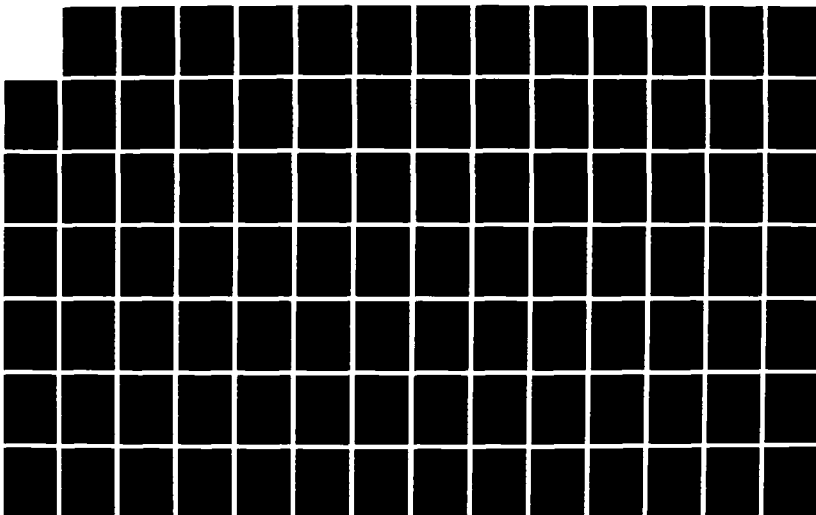
3/4

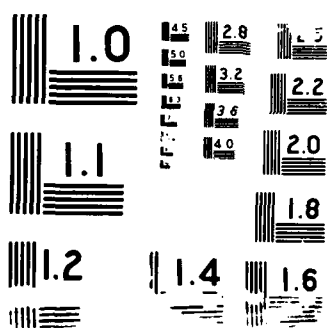
UNCLASSIFIED

FEB 88 AFATL-TR-87-73 F08635-84-C-0273

F/G 12/3

NL





$$Y^T X = U \Sigma V^T \quad (32)$$

is obtained for $Q = UV^T$, yielding

$$\text{maximum } \text{tr}[XQ^T Y^T] = \text{tr}[\Sigma] \quad (33)$$

The minimization of the Frobenius norm in the above problem stems from the fact that if the columns of the data matrices X and Y are displayed in R^p , the a natural measure of the "goodness of fit" between the N point clusters in R^p is

$$\begin{aligned} D^2(X, Y) &= \sum_{k=1}^N \|x_k - y_k\|^2 = \text{tr}[(X-Y)(X-Y)^T] \\ &= \|X-Y\|_F^2 \end{aligned} \quad (34)$$

If the problem is generalized by considering more general transformations T , consisting of translations, rotations, and (uniform) scalings, then it is shown [44] that an asymmetry exists (i.e., $D(X, Y) \neq D(Y, X)$). Gower [45] noted that symmetry can be recovered by centering and normalization of the data matrices. Sibson [46] introduced two related coefficients

$$\beta = \frac{\text{tr}[XX^T] - 2\text{tr}[Y^T XX^T Y]^{1/2} + \text{tr}[YY^T]}{[\text{tr}[XX^T] \text{tr}[YY^T]]^{1/2}} \quad (35)$$

$$\alpha = 1 - \frac{[\text{tr}[Y^T XX^T Y]^{1/2}]}{\text{tr}[X^T X] \text{tr}[Y^T Y]} \quad (36)$$

which also retain the desired symmetry property on $D(X, Y)$.

We will next solve the covariance based stochastic realization problem from an RV-coefficient point of view. We start from given matrices, X and Y , and then search for linear transformation matrices, L and/or M , in such a way that the "images" of the transformed variables are as close as possible. The reason for this stems from Akaike's stochastic realization theory [3], [12], [13], where the transformed variables form a minimal information interface, that is $L^T X$ should contain all the information about Y that is contained in X and vice versa. Also, if we were to reproduce X (or Y) from the transformed variables $M^T Y$ ($L^T X$), then we would like to lose as little information as possible in the approximation.

B. The Generalized Symmetric Stochastic Realization Problem

Let $X = Y_k^+$ and $Y = Y_k^-$, then $S_{xx} = R^+$, $S_{yy} = R^-$, $S_{xy} = H$, and $S_{yx} = H^T$ are as defined previously. Furthermore, if we let the $(n \times 1)$ state vectors of the backward and forward realizations be as defined earlier, i.e.,

$$z_{k-1} = L^T Y_k^+ \quad (37a)$$

$$x_k = M^T Y_k^- \quad (37b)$$

with corresponding diagonal covariance matrices given by

$$E\{z_{k-1} z_{k-1}^T\} = L^T(R^+)L = \Delta_z \quad (38a)$$

$$E\{x_k x_k^T\} = M^T(R^-)M = \Delta_x \quad (38b)$$

where L^T and M^T are the $(n \times \infty)$ transformation matrices, then the generalized symmetric stochastic realization problem can now be formulated as the following constrained optimization problem:

$$\begin{aligned} (P1): \quad \underset{L, M}{\text{Maximize}} \quad RV(L^T Y_k^+, M^T Y_k^-) &= \frac{\text{tr}[L^T H M M^T H L]}{[\text{tr}[L^T(R^+)L]^2 \text{tr}[M^T(R^-)M]^2]^{1/2}} \\ \text{Subject to:} \quad L^T(R^+)L &= \Delta_z \\ M^T(R^-)M &= \Delta_x \end{aligned} \quad (39)$$

By introducing Lagrange multipliers $(\lambda_1, \lambda_2, \dots, \lambda_n)$ and $(\psi_1, \psi_2, \dots, \psi_n)$, we can transform (P1) into the following unconstrained optimization problem

$$\begin{aligned} (P2): \quad \underset{L, M}{\text{Maximize}} \quad \phi(L, M) &= \text{tr}[L^T H M M^T H L] - \sum_{i=1}^n \lambda_i [L^T(R^+)L]_{ii} \\ &\quad - \sum_{i=1}^n \psi_i [M^T(R^-)M]_{ii} \end{aligned} \quad (40)$$

Then upon taking the derivative of $\phi(L,M)$ with respect to L and M and equating them to zero, followed by some algebraic manipulations, we get the following optimality conditions [10]

$$H(R^-)^{-1}H^T L - (R^+)L\Gamma = 0 \quad (41a)$$

$$H^T(R^+)^{-1}HM - (R^-)M\Gamma = 0 \quad (41b)$$

where Γ is the $(n \times n)$ diagonal matrix of squared canonical correlation coefficients. In addition to (41), we have the following identities [10], [31]

$$\Delta_z \Lambda = \Delta_x \Psi \quad (42a)$$

$$\Lambda = \Delta_x \Gamma \quad (42b)$$

$$\Psi = \Delta_z \Gamma \quad (42c)$$

where $\Lambda = \text{diag} [\lambda_1, \lambda_2, \dots, \lambda_n]$ and $\Psi = \text{diag} [\psi_1, \psi_2, \dots, \psi_n]$. Notice how (42) displays the symmetry properties of $RV(L^T Y_k^+, M^T Y_k^-)$. As a result of this symmetry, we can formulate the solution as given by (41), into the format of a

weighted singular value decomposition [47], [48], which in turn leads to the following canonical decomposition for H

$$\begin{aligned} H &= (R^+) L \Delta_z^{-1/2} \Gamma^{1/2} \Delta_x^{-1/2} M^T (R^-) \\ &= [(R^+) L \Delta_z^{-1/2} \Gamma^{1/4}] [\Gamma^{1/4} \Delta_x^{-1/2} M^T (R^-)] \\ &= OC \end{aligned} \quad (43)$$

where O and C are respectively the observability and controllability matrices. Upon solving for L and M from the respective expressions for O and C in (43), and substituting these into the state equations, we finally get the desired solution, i.e.,

$$z_{k-1} = \Delta_z^{1/2} \Gamma^{-1/4} O^T (R^+)^{-1} Y_k^+ \quad (44a)$$

$$x_k = \Delta_x^{1/2} \Gamma^{-1/4} C (R^-)^{-1} Y_k^- \quad (44b)$$

Then for selected values of Δ_z and Δ_x , the optimal value of $RV(L^T Y_k^+, M^T Y_k^-)$ is given by

$$RV(L^T Y_k^+, M^T Y_k^-) = \frac{\sum_{i=1}^n (\delta_{z_i} \gamma_i \delta_{x_i})}{\left[\left(\sum_{i=1}^n \delta_{z_i}^2 \right) \cdot \left(\sum_{i=1}^n \delta_{x_i}^2 \right) \right]^{1/2}} \quad (45)$$

It is worth mentioning at this point that the same problem for fixed Δ_z and Δ_x , was originally solved by Akaike [13] from a canonical correlation analysis point of view, and was later algorithmically extended by Desai and Pal [8]. We will show next that our approach, although different, is general in the sense that it covers these previous results as specific cases. This is due to the flexibility in the choice for Δ_z and Δ_x .

NORMALIZED AND BALANCED STOCHASTIC REALIZATIONS

Here we will make use of the flexibility in the choice for Δ_z and Δ_x in order to characterize different Markovian representations. It is a known fact in realization theory that equivalent minimal Markovian realizations are related by a similarity transformation. Thus, if we restrict Δ_z and Δ_x to be either the identity matrix or some function of Γ , then we can define a set of coordinate systems which give essentially a unique system representation under sign changes of the basis vectors. The following Lemma characterizes the existence conditions for these coordinate systems.

Lemma 1

Given a canonical factorization for H as in (43) with Δ_z and Δ_x satisfying (38), if we let $\Delta_x = I_n$ (or $\Delta_z = I_n$), then the optimal level for $\Delta_z(\Delta_x)$ in the RV-coefficient sense is reached at $\Delta_z = \Gamma$ ($\Delta_x = \Gamma$).

proof: From the definition of the RV-coefficient, we know that

$$\frac{\sum_{i=1}^n (\delta_{z_i} \gamma_i \delta_{x_i})}{\left[\left(\sum_{i=1}^n \delta_{z_i}^2 \right) \left(\sum_{i=1}^n \delta_{x_i}^2 \right) \right]^{1/2}} \leq 1 \quad (46)$$

Multiplying (46) by its denominator and collecting terms, we get

$$\Delta = \sum_{i=1}^n (\delta_{z_i} \gamma_i \delta_{x_i}) - \left[\left(\sum_{i=1}^n \delta_{z_i}^2 \right) \left(\sum_{i=1}^n \delta_{x_i}^2 \right) \right]^{1/2} \quad (47)$$

Then, as we increase Δ , the RV-coefficient approaches its upper bound. Hence, we need to maximize Δ . Let us fix Δ_x , then taking the derivative of Δ with respect to Δ_z and equating it to zero, we get the following optimality conditions

$$\frac{\partial \Delta}{\partial \Delta_z} = \sum_{i=1}^n \gamma_i \delta_{x_i} - \left(\sum_{i=1}^n \delta_{z_i} \right) \left[\frac{\sum_{i=1}^n \delta_{x_i}^2}{\sum_{i=1}^n \delta_{z_i}^2} \right]^{1/2} = 0 \quad (48a)$$

Similarly, fixing Δ_z and operating on $\frac{\partial \Delta}{\partial \Delta_x} = 0$, we get

$$\frac{\partial \Delta}{\partial \Delta_x} = \sum_{i=1}^n \gamma_i \delta_{z_i} - \left(\sum_{i=1}^n \delta_{x_i} \right) \left[\frac{\sum_{i=1}^n \delta_{z_i}^2}{\sum_{i=1}^n \delta_{x_i}^2} \right]^{1/2} = 0 \quad (48b)$$

Suppose further that $\Delta_x = I_n$ in (48a), then one can verify that the optimal choice for Δ_z is to make it equal to Γ . Similarly, let $\Delta_z = I_n$ in (48b), then $\Delta_x = \Gamma$ corresponds to the optimal choice, and the lemma follows.

The following definition extends the normalized realizations of Moore [49] to the stochastic case. This is possible from Lemma 1 which shows the optimality of these in the RV-coefficient sense.

Definition 1

A stochastic realization for $\Lambda(k)$ is input - normal if and only if $\Delta_z = I_n$ and $\Delta_x = \Gamma$; output-normal if and only if $\Delta_x = I_n$ and $\Delta_z = \Gamma$; and balanced if and only if $\Delta_z = \Delta_x = \Gamma^{1/2}$.

The results of Lemma 1 and Definition 1 lead us to the following Theorem for characterizing normalized and balanced stochastic realizations from the canonical decomposition of H .

Theorem 1

There exist nonsingular transformation matrices T_i , for $i=1,2,3$ such that the state vectors

$$z_{k-1}^i = T_1^{-T} O^T (R^+)^{-1} Y_k^+ \quad (49a)$$

$$x_k^i = T_1 C (R^-)^{-1} Y_k^- \quad (49b)$$

are transformed under the rule

$$T_i = \Gamma^{\left(\frac{j-1}{4}\right)} \quad \text{for } j = 0, 1, 2 \text{ and } i = j+1 \quad (50)$$

into the following realizations:

- a) $i=1$; input-normal
- b) $i=2$; balanced
- c) $i=3$; output-normal

Furthermore, the basis vectors for these coordinate systems are aligned and differ only by scale factors:

$$T_1 = \Gamma^{-1/4} T_2 \quad (51a)$$

$$T_3 = \Gamma^{1/4} T_2 \quad (51b)$$

proof: Let $P_z^i = \Delta_z^{1/2} \Gamma^{-1/4}$ and $P_x^i = \Delta_x^{1/2} \Gamma^{-1/4}$ for $i = 1, 2, 3$. Then from Lemma 1 and Definition 1, one can verify that $P_z^1 = \Gamma^{-1/4}$, $P_x^1 = \Gamma^{1/4}$, $P_z^3 = \Gamma^{1/4}$, and $P_x^3 = \Gamma^{-1/4}$. If we now let $\Delta_z = \Delta_x = \Gamma^{1/2}$, then $P_z^2 = P_x^2 = I_n$. This shows that $T_i^{-T} = P_z^i$ and $T_i = P_x^i$ both satisfy (50). Furthermore, T_i is a diagonal matrix with positive elements only, thus invertible. Finally, (51) follows easily from the fact that $T_2 = I_n$. This completes the proof.

Interestingly, the normalized stochastic realizations not only have proved to be optimal in the RV-coefficient sense, but also, along with the balanced stochastic realizations [8], they share the same generic property found in deterministic realization theory. This property was first introduced by Moore [49] for deterministic systems, and since it exactly carries over for stochastic systems, we will omit the details here. In addition, these stochastic realizations are unique modulo a sign matrix if the canonical correlation

coefficients are distinct, whereas, for scalar processes, a sign symmetry property is observed. Ramos [10] has recently shown the existence of a cross-Ricatti equation for scalar or symmetric stochastic realizations (this type of symmetry is related to the condition $\Lambda(k) = \Lambda(-k)$ and should not be confused with symmetry in the sense of (42)). This cross-Ricatti equation is the stochastic counterpart to a deterministic cross-Gramian equation introduced by Fernando and Nicholson [50] and further studied in [51], [52]. Further properties of balanced stochastic realizations, hence, normalized stochastic realizations because of their generic property, can be found in [8], [48], [53], [54].

As a final remark, it is worth mentioning that when $RV(L^T Y_k^+, M^T Y_k^-) = 1$, the state vectors are related to one another by a rotation and a scale factor, i.e.,

$$z_{k-1} = \alpha P x_k \quad (52)$$

where α is a nonzero scalar and P is an orthogonal matrix. This follows from properties 3, 5, and 6 in the definition of the RV-coefficient and implies that z_{k-1} has a reversibility property. Interestingly, Anderson and Kailath [14] introduced this property for self-dual forward-backward Markovian pairs. A condition for this self-duality is that the autocovariance function of $\{y_k\}$, $\Lambda(k)$, is symmetric. This is precisely the condition for the existence of the cross-Ricatti equation found by Ramos [10].

C. The Generalized Unsymmetric Stochastic Realization Problem

Canonical correlation based measures are, in general, invariant under nonsingular transformations of Y_k^+ and Y_k^- . This property, although indirectly used by Desai and Pal [8] to prove uniqueness of the canonical correlations and that realizations obtained from CRA are unique modulo a sign matrix if the canonical correlations are distinct, is not always a desirable one. This stems from the fact that the total variance of a set of variables is not invariant under all nonsingular transformations of the variables, but only under orthogonal transformations. For instance if L is an orthogonal matrix, then $RV(L^T Y_k^+, M^T Y_k^-) = RV(Y_k^+, M^T Y_k^-)$ can be interpreted as the ability of Y_k^- to predict a linear combination of Y_k^+ which accounts for a large proportion of the total variance of Y_k^+ . This situation has appeared recently in the work of Arun and Kung [9], who introduced the Karhunen-Loève method (KLM) (see also [39] - [41], where the same method is called the unweighted principal component algorithm (UPC)) for obtaining a forward Markovian representation. The idea here is to optimally approximate the information interface between Y_k^- and Y_k^+ via a one-sided Karhunen-Loève expansion of the forward predictor space. We will next show that this corresponds to a more general problem in the RV-coefficient framework.

In the case of obtaining only a forward Markovian representation such as $x_k = M^T Y_k^-$, we solve the following optimization problem:

$$\begin{aligned}
 \text{(P3):} \quad & \text{Maximize} \quad RV(L^T Y_k^+, M^T Y_k^-) = \frac{\text{tr}[L^T H M M^T H^T L]}{[\text{tr}[L^T (R^+) L] \text{tr}[M^T (R^-) M]]^{1/2}} \\
 & \text{Subject to:} \quad L^T L = I_n \\
 & \quad \quad \quad M^T (R^-) M = \Delta_x
 \end{aligned} \tag{53}$$

Notice that L does not alter the form of the solution since it is an orthogonal matrix, however, we will retain it to better illustrate the properties of the solution and its connection with that of the symmetric problem. Again, if we introduce Lagrange multiplier matrices $\Lambda_x = \text{diag} [\lambda_{x_1}, \lambda_{x_2}, \dots, \lambda_{x_n}]$ and $\Psi = \text{diag} [\psi_1, \psi_2, \dots, \psi_n]$, then we are led to the maximization of

$$\phi(M, L) = \text{tr}[HMM^T H^T] - \sum_{i=1}^n \lambda_{x_i} [M^T (R^-) M]_{ii} - \sum_{i=1}^n \psi_i [L^T L]_{ii} \quad (54)$$

whose optimality conditions are given by

$$\frac{\partial \phi(L, M)}{\partial L} = HMM^T H^T L - L\Psi \quad (55a)$$

$$\frac{\partial \phi(L, M)}{\partial M} = H^T L L^T H M - (R^-) M \Lambda_x = 0 \quad (55b)$$

After some algebra, one can show that L is given by the solution to the following eigenvalue - eigenvector problem

$$H(R^-)^{-1} H^T L = L \Lambda_x \quad (56)$$

where Λ_x is the diagonal matrix of eigenvalues in descending order of magnitude. From (55) and (56), the solution for M can be obtained as

$$M = (R^-)^{-1} H^T L \Lambda_x^{-1/2} \Delta_x^{1/2} \quad (57)$$

and the optimum RV-coefficient is given by

$$RV(L^T Y_k^+, M^T Y_k^-) = RV(Y_k^+, M^T Y_k^-) = \frac{\sum_{i=1}^n (\lambda_{x_i} \delta_{x_i})}{[tr[R^+]^2 (\sum_{i=1}^n \delta_{x_i}^2)]^{1/2}} \quad (58)$$

If we now let $\pi = \Delta_x^{-1/2} \Lambda_x^{1/2}$, then the following decomposition for H is immediate

$$\begin{aligned} H &= L \pi M^T (R^-) \\ &= [L \pi^{1/2}] [\pi^{1/2} M^T (R^-)] \\ &= OC \end{aligned} \quad (59)$$

Substituting for M^T as a function of C into the state equation, we get

$$x_k = \Delta_x^{1/2} \Lambda_x^{-1/2} C(R^-)^{-1} Y_k^- \quad (60)$$

which has the same format as its counterpart from the symmetric problem. However, these are not equal because of the difference in the canonical decomposition for H from both approaches. This is easily seen from the following canonical covariance structure

$$E \left\{ \begin{bmatrix} L^T Y_k^+ \\ M^T Y_k^- \end{bmatrix} \begin{bmatrix} L^T Y_k^+, & M^T Y_k^- \end{bmatrix}^T \right\} = \begin{bmatrix} L^T(R^+)L & L^T H M \\ M^T H^T L & M^T(R^-)M \end{bmatrix} \quad (61)$$

$$= \begin{bmatrix} L^T(R^+)L & \Lambda_x^{1/2} \Delta_x^{1/2} \\ \Lambda_x^{1/2} \Delta_x^{1/2} & \Delta_x \end{bmatrix}$$

where $L^T(R^+)L$ is a general $(n \times n)$ matrix, therefore since $L^T Y_k^+$ is not a basis for $Z_{k-1} = \text{span}[Y_k^-, Y_k^+]$, it cannot be considered a state vector. On the other hand, if we maximize $RV(L^T Y_k^+, M^T Y_k^-)$ with $M^T M = I_n$ and (38a) as constraints, then $z_{k-1} = L^T Y_k^+$ corresponds to a backward state vector, but $M^T Y_k^-$ loses the state vector property.

Finally, from (60) one can see that different choices for Δ_x leads to different Markovian realizations. In particular, if we let $\Delta_x = \Lambda_x$, then (60) is equivalent to the state equation obtained by Arun and Kung [9] via KLM.

D. Unification of Coordinate-Free Stochastic Realization Algorithms

Most of the stochastic realization algorithms that yield canonical Markovian models (also known as coordinate-free representations) are special cases of the two general problems solved earlier in this section. The main differences rest upon the type of constraints used and the particular choice for Δ_x and Δ_z . In Figure 1, we present a hierarchy of solutions to the stochastic realization problem from an RV-coefficient point of view. Here, we start with a

general RV-type maximization problem and depending on the constraints, one can move towards a symmetric or unsymmetric solution. These are then classified by algorithms or type of solutions according to the choice for the state covariance matrices.

It should be noted that in all cases the parameters $(F, G, H)_n$ and (L, Q, R, S) are obtained from the canonical decomposition of the Hankel matrix H (i.e., (43) for the symmetric stochastic realization problem and (59) for the unsymmetric problem) as follows [8] - [10]

$$n = \text{rank } [H] \quad (62a)$$

$$L = \Delta_x \quad (62b)$$

$$(F, G, H)_n = ([\tilde{C}C^\# = O^\#O^\uparrow], [CE], [E^TO])_n \quad (62c)$$

$$(Q, R, S) = ([L - FEF^T], [\Lambda(O) - HEH^T], [G - FEH^T]) \quad (62d)$$

where " $\uparrow$ " and " $\leftarrow$ " are respectively the shift-up and shift-left operations on block matrices, " $\#$ " denotes the pseudo inverse for non square matrices, and $E = [I_m, 0, 0, \dots, 0]^T$.

To show how different choices for the state covariance matrices lead to different solutions, consider the symmetric stochastic realization problem with $\Delta_x = I_n$ and let $\Delta_x^* = \Gamma$ be the maximum state covariance matrix of a forward anti-filter model of the form

$$x_{k+1}^* = Fx_k^* + w_k^* \quad (63a)$$

$$\bar{y}_k = Hx_k^* + v_k^* \quad (63b)$$

where w_k^* and v_k^* are white Gaussian noise processes. Then since $x_k^* = \Delta_z^{-1} z_k$ and $\Delta_x^* = \Delta_z^{-1}$ [4], $\Delta_z = \Gamma^{-1}$ must be satisfied, thus leading to the solution by Akaike [13]. By similar arguments one can verify that $\Delta_z = \Delta_x = I_n$ correspond to the state covariance matrices in Baram's algorithm [11]. The solutions by Desai and Pal [8] and Arun and Kung [9] follow easily from our previous results. Since the algorithms are mainly related by scale factors, the generality of the RV-coefficient approach is immediately noted.

The use of different RV-coefficient statistics (see Figure 1) as a measure for comparison between symmetric and unsymmetric stochastic realization algorithms, as well as a tool for model reduction, is illustrated by Ramos [10] for a set of streamflow data from the Nile River. His results indicate that the unsymmetric stochastic realization problem is more stable in terms of the rank of the Hankel matrix, resulting in smaller reduced-order models as well as better RV-coefficient performance.

However, for prediction purposes, both approaches led to similar results. These results will be published in a separate paper on stochastic model reduction.

IV. A DIRECT APPROACH FOR OBTAINING COORDINATE-FREE MARKOVIAN REPRESENTATIONS

A. Symmetric Case

Here we assume that a forward-backward pair of Markovian models of the innovations representation type are given and characterized by x_{*k} and z_{*k} , with respective state covariance matrices P and N (not diagonal) satisfying a pair of algebraic Riccati Equations (see [4], [48], [54] for such construction). We are then interested in finding a similarity transformation of the form $x_k = Tx_{*k}$ and $z_k = T^{-T}z_{*k}$ such that in the new coordinate system, the state covariance matrices satisfy the properties of Definition 1.

Let $x_k = M^T x_{*k}$ and $z_k = L^T z_{*k}$ be the transformed state vectors. Then, we wish to find L and M such that x_k and z_{k-1} are as close as possible in the RV-sense. This amounts to solving the following RV-optimization problem.

$$(P4): \underset{L, M}{\text{Maximize}} \quad RV(L^T z_{*k-1}, M^T x_{*k}) = \frac{\text{tr}[L^T N P M M^T P N M]}{[\text{tr}[L^T N L]^2 \text{tr}[M^T P M]^2]^{1/2}}$$

$$\text{Subject to: } L^T N L = \Delta_z \quad (64)$$

$$M^T P M = \Delta_x$$

where $E\{z_{*k-1} x_{*k}^T\} = NP$ has been shown in [55]. If we let $\Xi = NP$, then the solution corresponds to a singular value decomposition of $\Xi = N^{-1/2} \bar{\Xi} P^{-T/2}$, which leads to

$$\Xi = N L \Delta_z^{-1/2} \Gamma^{1/2} \Delta_x^{-1/2} M^T P \quad (65)$$

where $\Gamma^{1/2}$ is again the diagonal matrix of canonical correlations [48]. The correspondence between L and M is shown to be

$$L^T = \Delta_z^{1/2} \Gamma^{-1/2} \Delta_x^{-1/2} M^T \Xi^T N^{-1} \quad (66a)$$

$$M^T = \Delta_x^{1/2} \Gamma^{-1/2} \Delta_z^{-1/2} L^T \Xi P^{-1} \quad (66b)$$

from which the following identity is immediate

$$L^T \Xi M = \Delta_z^{1/2} \Gamma^{1/2} \Delta_x^{1/2} \quad (67a)$$

Now, since $\Xi = NP$, from (65) one can see that

$$L \Delta_z^{-1/2} \Gamma^{1/2} \Delta_x^{-1/2} M^T = I_n \quad (67b)$$

must be satisfied. Furthermore, if we recall that $M^T = T$, then for any choice of Δ_x and Δ_z satisfying the properties of Definition 1, the identity $L = M^{-T}$ holds, which tells us that applying the transformation L and M simultaneously is equivalent to applying a similarity transformation T to both the forward and backward models. We now have the following theorem which establishes this fact.

Theorem 2

Given a canonical decomposition for Ξ as in (65) with L and M satisfying the constraints in (64), if we apply a transformation of the form

$$T_i = \Gamma^{\binom{j-1}{4}} T, \quad j = 0, 1, 2 \text{ and } i = j+1 \quad (68)$$

$$T = \Gamma^{1/4} \Delta_x^{-1/2} M^T = \Gamma^{-1/4} \Delta_z^{1/2} L^{-1}$$

to the pair of Markovian models, then the normalized and balanced stochastic realizations are characterized by

$$x_k^i = T_i x_{*k} \quad (69a)$$

$$z_k^i = T_i^{-T} z_{*k} \quad (69b)$$

where

- a) $i = 1$: input-normal
- b) $i = 2$: balanced
- c) $i = 3$: output-normal

Furthermore, the transformations T_i differ only by scale factors, i.e.,

$$T_1 = \Gamma^{-1/4} T_2 \quad (70a)$$

$$T_3 = \Gamma^{1/4} T_2 \quad (70b)$$

proof: Follows easily from (67), Definition 1, and Lemma 1.

Remark 1: This direct approach is equivalent to artificially symmetrizing a system of the form $\Xi = T \Lambda^2 T^{-1}$ via a singular value decomposition [48], [54].

Remark 2: For deterministic systems, the equivalent approach would be to maximize $RV(L^T O^T, M^T C)$ subject to $L^T W_o L = \Delta_o$ and $M^T W_c M = \Delta_c$, where W_o and W_c are respectively the observability and controllability gramians [56]. This establishes a tri-equivalence between deterministic balancing, stochastic balancing, and the symmetric stochastic realization problem, in terms of solution strategy.

B. Unsymmetric Case

A direct approach is also possible for the unsymmetric problem by simply solving the following RV optimization problem

$$(P5): \underset{L, M}{\text{Maximize}} \text{RV}(L^T Y_k^+, M^T x_{*k}) = \frac{\text{tr}[L^T \Sigma_{yx} M M^T \Sigma_{xy} L]}{[\text{tr}[L^T (R^+) L] \text{tr}[M^T P M]]^{1/2}}$$

$$\begin{aligned} \text{Subject to: } L^T L &= I_n \\ M^T P M &= \Delta_x \end{aligned} \quad (71)$$

where $\Sigma_{yx} = E\{Y_k^+ x_{*k}^T\}$.

Following the same Lagrangian optimization procedure as in the previous problems, one can easily show that the solution is given by solving the following eigen-system

$$\Sigma_{yx} M M^T \Sigma_{xy} L - L \Psi = 0 \quad (72a)$$

$$\Sigma_{xy} L L^T \Sigma_{yx} M - P M \Delta_x = 0 \quad (72b)$$

from which the following identity is observed

$$\Psi = \Delta_x \Delta_x \quad (73)$$

The solution for M is then given by

$$M^T = \Delta_x^{-1/2} \Delta_x^{-1/2} L^T \Sigma_{yx} P^{-1} \quad (74)$$

Now, notice that $x_{*k} = E(x_k | Y_k^-) = C(R^-)^{-1} Y_k^-$, where x_k is any state vector satisfying (1), then $E_{yx} = H(R^-)^{-1} C^T = 0$ and (74) simplifies to

$$H^T = \Delta_x^{1/2} \Lambda_x^{-1/2} L^T O \quad (75)$$

which transforms the state vector into

$$\begin{aligned} x_k &= \Delta_x^{1/2} \Lambda_x^{-1/2} L^T O x_{*k} \\ &= \Delta_x^{1/2} \Lambda_x^{-1/2} L^T H(R^-)^{-1} Y_k^- \end{aligned} \quad (76)$$

One can see that (71) is the same state equation obtained for the generalized unsymmetric stochastic realization problem (see, e.g., $x_k = M^T Y_k^-$ using M from (57)).

This last result completes the equivalence between the stochastic realization problem and a direct approach for obtaining coordinate-free Markovian representations, initiated by Desai and Pal [8].

V. CONCLUSIONS

The need for introducing a common statistical measure of information as a tool for comparison between different stochastic realization algorithms has led to a unification of Akaike's stochastic realization theory to handle other forms of multivariate analysis. This has been possible via Robert and Ecofier's novel RV-coefficient approach, which leads to a natural interpretation for the stochastic realization problem and can be used to advantage in the following context: 1) for algorithmic development, 2) for performance evaluation, and 3) for model approximation. Two types of problems have been identified and solved

from an RV-coefficient point of view, namely, the generalized symmetric stochastic realization problem and its unsymmetric version. Previous algorithms, which include among others, the canonical realization algorithm (CRA) and the Karhunen-Loève method (KLM), are shown to be particular solutions to these general problems. In each case, an RV-coefficient statistic has been presented based on the different choices for the state covariance matrices.

The normalized realizations found in deterministic realization theory have been carried over to the stochastic case, showing optimality in the RV-coefficient sense. Alternatively, the problem of finding a similarity transformation that brings a pair of innovations representations to a certain canonical form (coordinate-free) has also been tackled from an RV-coefficient point of view. This again can be formulated as two general problems which bear resemblance with the generalized symmetric and unsymmetric stochastic realization problems. Furthermore, a parallelism between deterministic balancing, stochastic balancing, and the generalized symmetric stochastic realization problem, from a solution standpoint, has been established.

Finally, we remark that the RV-coefficient approach provides us with a rich theory for solving a large class of multivariate problems. In particular, areas such as signal processing, random fields, discriminant analysis, and pattern recognition, to name only a few, should benefit from this analytical tool. Preliminary results addressing this issue have already been presented in [22], [23], where an abstract representation of the RV-coefficient has been introduced.

REFERENCES

- [1] T. Kailath, "A view of three decades of linear filtering theory," IEEE Trans. Inf. Theory, vol. IT-20, pp. 145-181, Feb. 1974.
- [2] A. S. Willsky, Digital Signal Processing and Control and Estimation Theory, MIT Press, Massachusetts, 1979.
- [3] H. Akaike, "Stochastic theory of minimal realization," IEEE Trans. Automat. Contr., vol. AC-19, pp. 667-674, June 1974.
- [4] P. Faurre, "Stochastic realization algorithms," in: System Identification: Advances and Case Studies, Ed. R. K. Mehra and D. G. Lainiotis, Academic Press, New York, 1976.
- [5] M. Gevers and W. R. E. Wauters, "An innovations approach to discrete-time stochastic realization problem," Journal A, vol. 19, No. 2, pp. 90-100, 1978.
- [6] P. Bernhard, "From classical to modern signal processing. Some aspects of the theory," in: Mathematical Techniques of Optimization, Control and Decision, Ed., J. P. Aubin, A. Bensoussan, and I. Ekeland, Birkhauser, Massachusetts, 1981.
- [7] W. E. Larimore, "System identification, reduced-order filtering and modeling via canonical variate method," Proc. Amer. Control Conf., San Francisco, CA, pp. 445-451, 1983.
- [8] U. B. Desai and D. Pal, "A realization approach to stochastic model reduction and balanced stochastic realizations," Proc. of the 21st IEEE Conf. on Decision and Control, Orlando, FL, pp. 1105-1112, 1982.
- [9] K. S. Arun and S. Y. Kung, "A new algorithm for approximate stochastic realization," Proc. of the 22nd Conf. on Decision and Control, San Antonio, TX, pp. 1353 - , 1983.
- [10] J. A. Ramos, "A stochastic realization and model reduction approach to streamflow modeling," Ph.D. dissertation, Dept. of Civil Eng., Georgia Institute of Technology, Atlanta, GA, 1985.
- [11] Y. Baram, "Realization and reduction of Markovian models from nonstationary data," IEEE Trans. Automat. Contr., vol. AC-26, pp. 1225-1231, June 1981.
- [12] H. Akaike, "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes," Ann. Inst. Statis. Math., vol. 26, pp. 363-387, 1974.
- [13] H. Akaike, "Markovian representation of stochastic processes by canonical variables," SIAM Journal of Control, Vol. 13, No. 1, pp. 162-173, 1975.

- [14] B. D. O. Anderson and T. Kailath, "Forwards, backwards and dynamically reversible markovian models of second-order processes," Proc. Int. Conf. Circuits and Systems, New York, pp. 981-986, 1978.
- [15] G. S. Sidhu and U. B. Desai, "New smoothing algorithms based on reversed-time lumped models," IEEE Trans. Automat. Contr., vol. AC-21, pp. 538-541, Aug. 1976.
- [16] L. Ljung and T. Kailath, "Backwards markovian models for second-order stochastic processes," IEEE Trans. Inf. Theory, vol. IT-22, pp. 488-491, April 1976.
- [17] B. L. Ho and R. E. Kalman, "Effective construction of linear state-variable models from input/output functions," Regelungstechnik, vol. 14, pp. 454-548, 1966.
- [18] S. Fujishige, H. Nagai, and Y. Sawaragi, "System theoeretical approach to model reduction and system-order determination," Int. Journal of Control, vol. 22, No. 6, pp. 807-819, 1975.
- [19] J. V. White, "Stochastic state-space models from empirical data," Proc. Conf. Acoust., Speech and Signal Proc., Boston, MA, pp. 243-247, 1983.
- [20] R. J. Vaccaro, "Modeling of perturbed covariance sequences," Proc. IEEE, vol. 74, pp. 617-619, April, 1986.
- [21] Y. Escoufier, "Le traitement des variables vectorielles," Biometrics, vol. 29, pp. 751-760, 1973.
- [22] E. I. Verriest, "A unified theory of model reduction via Gealson measures," Proc. IMA Conf. Math. in Signal Processing, Bath, U.K., 1985 (to be published).
- [23] E. I. Verriest, "Projection techniques for model reduction," in: Modeling, Identification and Robust Control, Ed., C. I. Byrnes and A. Lindquist, North-Holland, 1986.
- [24] H. Hotelling, "Relations between two sets of variates," Biometrika, vol. 28, pp. 321-377, 1936.
- [25] E. M. Cramer and W. A. Nicewander, "Some symmetric, invariant measures of multivariate association," Psychometrika, vol. 44, No. 1, pp. 43-54, 1979.
- [26] C. R. Rao, "The use and interpretation of principal component analysis in applied research," SANKHYA, A, vol. 26, pp. 329-358, 1966.
- [27] W. E. Larimore, "Generalized canonical variables with minimum quadratic prediction error," preprint.
- [28] A. L. van den Vollenberg, "Redundancy analysis. An alternative for canonical correlation analysis," in: A Second Generation of Multivariate Analysis, Vol. 1, Ed., Claess Fornell, Praeger, New York, 1982.

- [29] C. Fornell, "External single set components analysis of multiple criterion/multiple predictor variables," in: A Second Generation of Multivariate Analysis, vol. 1, Ed., C. Fornell, Praeger, New York, 1982.
- [30] J. A. Ramos and E. I. Verriest, "A unifying tool for comparing stochastic realization algorithms and model reduction techniques," Proc. Amer. Contr. Conf., San Diego, CA, pp. 150-155, 1984.
- [31] P. Robert and Y. Escoufier, "A unifying tool for linear multivariate statistical methods: the RV-coefficient," Appl. Statist., vol. 25, No. 3, pp. 257-265, 1976.
- [32] G. H. Golub and C. F. Van Loan, Matrix Computations, John Hopkins University Press, Baltimore, MD, 1983.
- [33] K. S. Srikantan, "Canonical association between nominal measurements," Journal Amer. Stat. Assoc., vol. 65, No. 329, pp. 284-292, 1970.
- [34] K. E. Muller, "Relations between redundancy analysis, canonical correlation, and multivariate regression," Psychometrika, vol. 46, No. 2, pp. 139-142, 1981.
- [35] D. G. Kabe, "On some multivariate statistical methodology with applications to statistics, psychology, and mathematical programming," Journal Ind. Math Soc., vol. 30, pt. 1, pp. 1-18, 1985.
- [36] V. J. Yohai and M. S. Garcia Ben, "Canonical variables as optimal predictors," Ann. Statist., vol. 8, No. 4, pp. 865-869, 1980.
- [37] G. P. McCabe, "Principal variables," Technometrics, vol. 26, No. 2, pp. 137-144, 1984.
- [38] I. M. Gelfand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function," Amer. Math Soc. Transl., vol. 2, No. 12, pp. 199-246, 1959.
- [39] K. S. Arun, D. V. Bhaskar Rao, and S. Y. Kung, "A new prediction efficiency criterion for approximate stochastic realization," Proc. 22nd Conf. on Decision and Control, San Antonio, TX, pp. 1353-1365, 1983.
- [40] S. Y. Kung and K. S. Arun, "Approximate realization methods for ARMA spectral estimation," IEEE Int. Symp. Circuits and Syst., vol. 1, pp. 105-109, 1983.
- [41] S. Y. Kung, K. S. Arun, and D. V. Bhaskar Rao, "State-space and singular-value decomposition-based approximation method for the harmonic retrieval problem," Journal Opt. Soc. Am., vol. 73, No. 12, pp. 1799-1811, 1983.
- [42] Y. Escoufier and P. Robert, "Optimizing RV-coefficient," in: Optimizing Methods in Statistics, Ed., J. S. Rustagi, pp. 205-219, Academic Press, New York, 1979.

- [43] M. Vallee and P. Robert, "A forward multivariate regression procedure based on maximization of the RV-coefficient," in: COMPSTAT 82, pp. 436-441, Physica-Verlag, Vienna, 1982.
- [44] G. A. F. Seber, Multivariate Observations, Wiley, New York, 1984.
- [45] J. C. Gower, "Statistical methods of comparing different multivariate analyses of the same data," in: Mathematics in the Archaeological and Historical Sciences, Ed. F. R. Hodson, D. G. Kendall, and P. Toutu, pp. 138-149, University Press, Edinburgh, 1971.
- [46] Sibson, "Studies in the robustness of multidimensional scaling procrustes statistics," Journal Royal Stat Soc., Series B, vol. 4, pp. 234-238, 1978.
- [47] A. Bjork and G. H. Golub, "Numerical methods for computing angles between linear subspaces," Math. Comp., vol. 27, pp. 579-594, 1973.
- [48] U. B. Desai, D. Pal, and R. D. Kirkpatrick, "A transformation approach to stochastic model reduction," IEEE Trans. Automat. Contr., vol. AC-29, pp. 1097-1100, Dec., 1984.
- [49] B. C. Moore, "Singular value analysis of linear systems, pt. I," Department of Electrical Engineering, Systems Control Report 7801, University of Toronto, Ontario, Canada, 1978.
- [50] K. V. Fernando and H. Nicholson, "Minimality of SISO Linear Systems," Proc. IEEE, vol. 70, No. 10, pp. 1241-1243, 1982.
- [51] A. J. Laub, L. M. Silverman, and M. Verma, "A note on cross Grammians for symmetric realizations," Proc. IEEE, vol. 71, No. 7, pp. 904-905, 1983.
- [52] J. A. de Abreu-Garcia and P. W. Fairman, "A note on cross Grammians for orthogonally symmetric realizations," IEEE Trans. Automat. Contr., vol. AC-31, pp. 866-868, Sept. 1986.
- [53] R. J. Vaccaro and B. W. Dickinson, "Pseudo-balanced realizations for deterministic and stochastic discrete-time systems," Proc. Allerton Conf. on Comm., Control, and Comp., pp. 874-883, 1982.
- [54] P. Harshavardhana, E. A. Jonckheere, and L. M. Silverman, "Stochastic balancing and approximation - stability and minimality," in: Proc. 22nd Conf. on Decision and Control, San Antonio, TX, pp. 1983.
- [55] U. B. Desai, D. Pal, and C. S. Hsu, "An invariant parameter in linear estimation and control," Proc. 21st Conf. on Decision and Control, Orlando, FL, pp. 1238-1288, 1982.
- [56] L. M. Silverman and M. Bettayeb, "Optimal approximation of linear systems," Proc. Joint Automat. Control Conf., San Francisco, CA, 1980.
- [57] P. Faurre, "Identification par minimisation d'une representation markovienne de processus aleatoire," Symp. on Optimization, Nice, Vienna, Springer-Verlag, Lecture Notes in Mathematics 132, pp. 85-107, 1969.

- [58] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

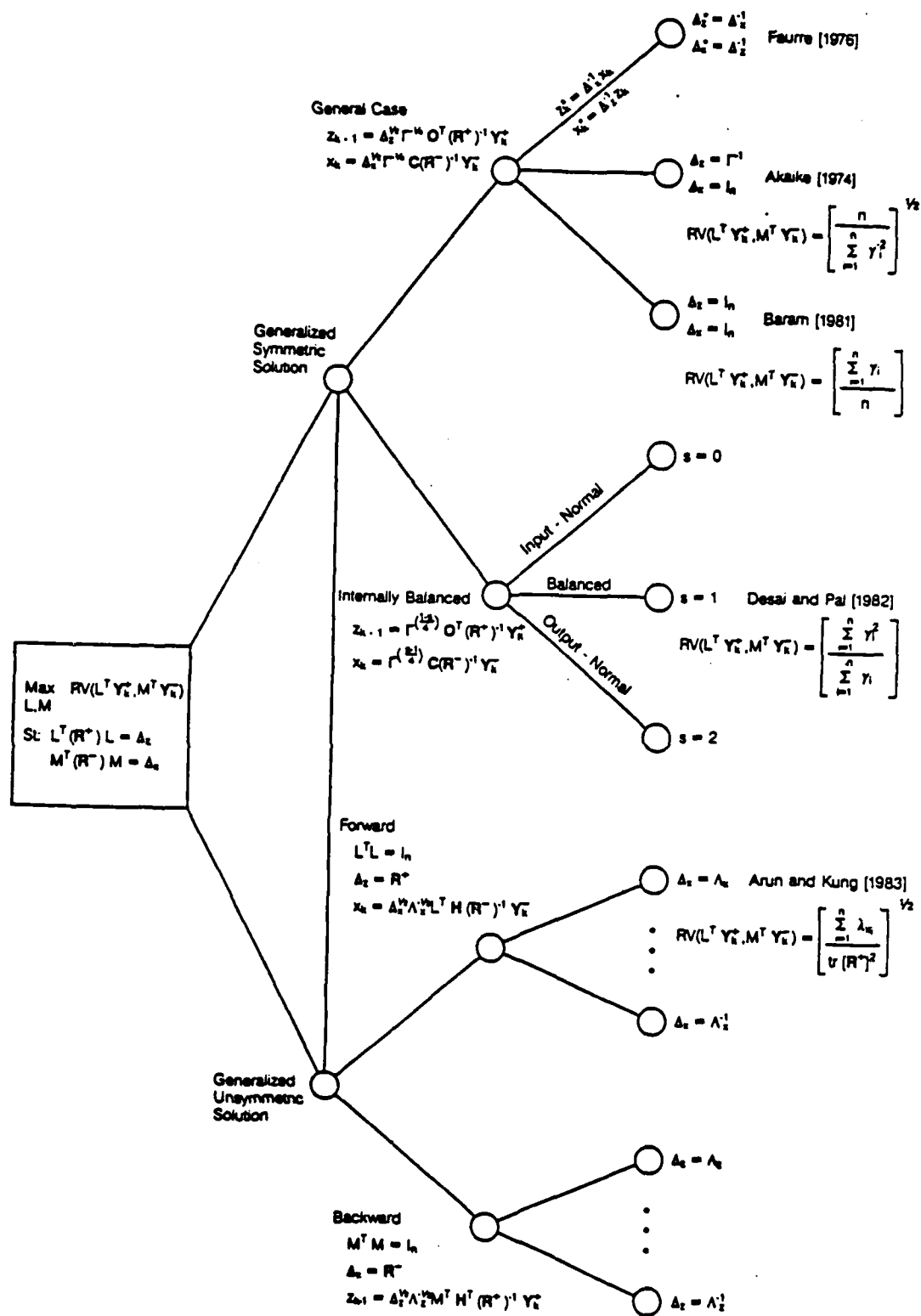


Figure 1. A hierarchy of solutions to the stochastic realization problem.

APPENDIX R

A NOTE ON THE CROSS RICCATIAN AND RELATED PROPERTIES FOR

SYMMETRIC STOCHASTIC REALIZATIONS

A NOTE ON A CROSS RICCATIAN AND RELATED PROPERTIES FOR
SYMMETRIC STOCHASTIC REALIZATIONS

Jose A. Ramos
United Technologies Optical Systems, Inc.
Optics and Applied Technology Laboratory
P.O. Box 109660
West Palm Beach, Florida 33410-9660
(305) 863-4271

Erik I. Verriest
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
(404) 894-2949

ABSTRACT

Recent results [1] - [7] concerning cross Grammians for both SISO and symmetric MIMO systems are extended for stochastic systems. The extension is based on a cross Riccatian matrix and several results from stochastic realization theory. It is shown that the cross Riccatian can be obtained from the solution to a cross Riccati equation carrying properties from both a forward and a backwards innovations representation. Other symmetry properties similar to those for the cross Grammian are introduced, along with connections to balanced stochastic realizations.

I. INTRODUCTION

In a recent series of papers [1] - [3], a cross Grammian matrix W_{co} was introduced for both continuous and discrete-time SISO linear dynamical systems. This matrix, which is the product of the controllability and observability matrices, i.e., $W_{co} = CO$, shares some interesting properties which are of concern in linear systems theory. One such property is that if W_{co} is of full rank, then the system is observable, controllable, and minimal. It has also been shown [1] that W_{co} can be determined from the solution to a matrix Lyapunov equation, and further, that $W_{co}^2 = W_c W_o$, where W_c and W_o are respectively the controllability and observability Grammians. In [4], these results were extended to MIMO systems having a symmetric transfer function matrix (equivalently symmetric Markov parameters). More recently [5], the definition of the cross Grammian has been extended to a more general class of transfer function matrices referred to as orthogonally symmetric transfer function matrices. Other properties of W_{co} are given in [6] and its connection to balanced realizations have been reported recently in [2], [3], [7].

In this note we introduce a cross Riccattian matrix, Σ_{PN} , for MIMO linear dynamical stochastic systems with symmetric Markov parameters. This type of symmetry arises naturally in SISO systems such as univariate ARMA and state-space models, as well as in MIMO systems representing electrical circuits and power networks. Other symmetry properties similar to those for W_{co} are introduced, including the solution to a cross Riccati equation and its connection to balanced stochastic realizations.

II. PRELIMINARIES. FORWARD - BACKWARDS STOCHASTIC REALIZATIONS

Here we are interested in stochastic processes that have a finite dimensional Markovian (state-space) representation. Thus, suppose we can represent a given m -dimensional zero-mean stationary stochastic process $\{y_k\}$ by a forward state-space model of the form

$$\begin{aligned} x_{k+1} &= Fx_k + v_k^f \\ y_k &= Hx_k + v_k^f \end{aligned} \quad k = 0, 1, \dots \quad (1)$$

where x_k is the $(n \times 1)$ state vector, y_k is the $(m \times 1)$ output vector, and $\{v_k^f\}$ and $\{v_k^b\}$ are zero-mean white Gaussian noise processes with joint covariance matrix

$$E \begin{bmatrix} v_k^f \\ v_k^b \end{bmatrix} \begin{bmatrix} v_s^f & v_s^b \end{bmatrix}^T = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{ks} \geq 0, \quad Q \geq 0, \quad R > 0 \quad (2)$$

Furthermore, v_k^f and v_k^b satisfy the forward propagation property

$$E\{x_k (v_s^f)^T\} = E\{x_k (v_s^b)^T\} = 0, \quad s \geq k \quad (3)$$

We assume that the parameter triple $(F, G, H)_n$ is minimal, i.e., observable and controllable such that the state covariance matrix and its inverse exist, i.e., $\Sigma = E\{x_k x_k^T\}$ is the unique positive definite solution to the following matrix Lyapunov equation [8]

$$\Sigma = F \Sigma F^T + Q \quad (4)$$

It is well known [8] that the output autocovariance matrix $\Lambda(k)$ is parametrically represented by

$$\Lambda(k) = [HF^{k-1}G]1_{(k)} + [G^T F^{(-k-1)T} H^T]1_{(-k)} + \Lambda(0)\delta_{k0} \quad (5)$$

where

$$1_{(k)} = \begin{cases} 1 & \text{if } k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\Lambda(0) = H\bar{E}H^T + R \quad (6)$$

$$G = S - P\bar{E}H^T \quad (7)$$

From duality one can also represent $\{y_k\}$ by a backwards state-space model [9], [10], i.e.,

$$z_{k-1} = F^T z_k + v_k^b \quad k = 0, -1, -2, \dots \quad (8)$$

$$y_k = G^T z_k + v_k^b$$

where z_k is an $(n \times 1)$ state vector evolving in the opposite direction of time, and $\{v_k^b\}$ and $\{v_k^b\}$ are zero-mean white Gaussian noise processes with joint covariance matrix

$$\begin{bmatrix} v_k^b \\ v_k^b \end{bmatrix} \begin{bmatrix} v_s^b, v_s^b \end{bmatrix}^T = \begin{bmatrix} Q_b & S_b \\ S_b^T & R_b \end{bmatrix} \quad \delta_{ks} \geq 0, Q_b \geq 0, R_b > 0 \quad (9)$$

and uncorrelated with z_s for $s \geq k$. In addition, $\Sigma^{-1} = E\{z_k z_k^T\}$ satisfies the following matrix Lyapunov equation

$$\Sigma^{-1} = F^T \Sigma^{-1} F + Q_b \quad (10)$$

while R_b and S_b satisfy

$$R_b = \Lambda(0) - G^T \Sigma^{-1} G = R + H E H^T - G^T \Sigma^{-1} G \quad (11)$$

$$S_b = H^T - F^T \Sigma G \quad (12)$$

A pair of models satisfying (1) - (12) is called a forward-backwards dual pair. These types of models have been studied recently in the light of balanced stochastic realizations [9] - [12] and in smoothing problems [13], [14].

III. SYMMETRIC STOCHASTIC REALIZATIONS

When the stochastic process $\{y_k\}$ is a scalar, or when given the model, the triple $(F, G, H)_n$ yields a symmetric stochastic realization, then some interesting properties common to both the forward and backwards model take place. These properties are summarized below in a series of Lemmas, along with their proofs.

Lemma 1:

The following conditions are equivalent for a scalar or symmetric stochastic realization:

- i. $(F, G, H)_n$ is symmetric
- ii. $\Lambda(k) = HF^{k-1}G$ is symmetric for all $k \geq 1$
- iii. $R^+ = R^- = (R^+)^{1/2}(R^-)^{T/2} = R$
- iv. $H = (R^+)^{-1/2}H(R^-)^{-T/2}$ is symmetric

where R^- and R^+ are semi-infinite toeplitz covariance matrices with respective first rows $[\Lambda(0), \Lambda(1), \Lambda(2), \dots]$ and $[\Lambda(0), \Lambda^T(1), \Lambda^T(2), \dots]$, H is a semi-infinite Hankel matrix with first row $[\Lambda(1), \Lambda(2), \Lambda(3), \dots]$, $(A)^{1/2}$ denotes the matrix square root of A , and A^T denotes the transpose of A .

proof: A condition for symmetry is that the parameters satisfy the following relation

$$F = DF^TD \quad \text{and} \quad H = G^TD \quad (13)$$

where D is a sign matrix. Substituting these into $\Lambda(k) = HF^{k-1}G$, we get

$$\begin{aligned} \Lambda(k) &= G^TDD(F^{k-1})^TDD^{-1}H^T \\ &= G^T(F^{k-1})^TH^T \\ &= \Lambda^T(k) = \Lambda(-k) \end{aligned} \quad (14)$$

where $DD = I_n$. From (14), properties (1-4) then become obvious.

Lemma 2:

Given a symmetric stochastic realization characterized by the triple

where

$$A_1 = \Lambda^{1/2}(0)$$

$$A_2 = CH^T \Lambda^{-1/2}(0)$$

$$A_3 = [R^- - CH^T \Lambda^{-1}(0)HC]^{1/2}$$

$$A_4 = [R^+ - OG \Lambda^{-1}(0)G^T O^T]^{1/2}$$

$$A_5 = \Lambda^{-1/2}(0)G^T O^T$$

and by applying the matrix inversion Lemma, along with the symmetry properties from Lemma 1, we get the desired result.

It should be noted that the above cross Riccati equation shares properties from both a forward and a backwards innovations representation. Furthermore, it is the stochastic counterpart to the cross Grammian equation introduced in [1] for deterministic systems.

Lemma 3:

For a symmetric stochastic realization $(F, G, H)_n$, the following relations always hold true:

$$\Sigma_{PN}^2 = PN \tag{17}$$

$$\Sigma_{PN} = PD \tag{18}$$

$$\Sigma_{PN} = DN \tag{19}$$

$(F, G, H)_N$, if P and N are positive definite solutions to a matrix Riccati equation corresponding respectively to the forward and backwards Kalman filter, then

$\Sigma_{PN} = P^{1/2} N^{T/2}$ satisfies a cross Riccati equation of the form

$$\Sigma_{PN} = F \Sigma_{PN} F^T + [G - F \Sigma_{PN} G] \{ \Lambda(0) - H \Sigma_{PN} G \}^{-1} [H^T - F^T \Sigma_{PN} H^T] \quad (15)$$

proof: Since $P = C(R^-)^{-1} C^T$ and $N = O^T(R^+)^{-1} O$ [12], we have

$$\Sigma_{PN} = C(R^-)^{-1/2} (R^+)^{-T/2} O = C(R)^{-1} O$$

or equivalently, by making use of the semi-infinity properties of O , C , R^- , and R^+ ,

$$\begin{aligned} \Sigma_{PN} &= \begin{bmatrix} G & FC \end{bmatrix} \begin{bmatrix} \Lambda(0) & HC \\ C^T H^T & R^- \end{bmatrix}^{-1/2} \begin{bmatrix} \Lambda(0) & G^T O^T \\ OG & R^+ \end{bmatrix}^{-T/2} \begin{bmatrix} H \\ OF \end{bmatrix} \\ &= \begin{bmatrix} G & FC \end{bmatrix} \begin{bmatrix} A1 & 0 \\ A2 & A3 \end{bmatrix}^{-1} \begin{bmatrix} A1 & A4 \\ 0 & A5 \end{bmatrix}^{-1} \begin{bmatrix} H \\ OF \end{bmatrix} \end{aligned} \quad (16)$$

proof: from property (3) of Lemma 1, we have

$$\begin{aligned} C(R^-)^{-1} C^T O^T (R^+)^{-1} O &= C(R^-)^{-T/2} [(R^+)^{-1/2} O C(R^-)^{-T/2}] (R^+)^{-1/2} O \\ &= C(R)^{-1} O C(R)^{-1} O \\ &= \Sigma_{PN}^2 \end{aligned}$$

and recalling that $\Sigma_{PN} = C(R)^{-1} O$ and by using the identity $C^T D = 0$, the last two relations follow easily.

Lemma 4:

A necessary and sufficient condition for a symmetric system to be observable and controllable, hence minimal, is that $\text{rank } [\Sigma_{PN}] = n$.

proof: It suffices to show that the nonzero eigenvalues of the weighted Hankel matrix H are also the eigenvalues of Σ_{PN} . Let

$$|\Sigma_{PN} - \lambda I_n| = |C(R)^{-1} O - \lambda I_n| = 0 \quad (20)$$

be the characteristic equation for Σ_{PN} . Then it follows that

$$\begin{aligned} |C(R)^{-1} O - \lambda I_n| &= (-\lambda)^{n-p} |O C(R)^{-1} - \lambda I_p| = 0 \\ &= |(R)^{-1/2} H (R)^{-1/2} - \lambda I_p| \\ &= |H - \lambda I_p| = 0 \end{aligned} \quad (21)$$

where $p > n$ is the dimension of H and R . Hence, Σ_{PN} and H have the same nonzero eigenvalues. Now, since $\text{rank}[H] = \text{rank}[\Sigma_{PN}] = n$ is a necessary and sufficient

condition for minimality (equivalently observability and controllability), then $\text{rank}[\Sigma_{PN}] = n$ since H has $(p-n)$ zero eigenvalues. This completes the proof.

Lemma 5:

The Cauchy index of a symmetric stochastic realization $(F, G, H)_n$ is given by the signature of the cross Riccatian Σ_{PN} .

proof: The Cauchy index is given by the sum of the positive real minus the negative real eigenvalues of the Hankel matrix. Hence, from (21) it is clear that this is equivalent to the signature of Σ_{PN} .

The conditions of Lemma 3 imply that the computation of the cross Riccatian reduces to the problem of determining either P or N . Whereas Lemmas 4 and 5 reveal that the cross Riccatian contains the same information as the Hankel matrix but in a more compact form.

Lemma 6:

The eigenvalues of Σ_{PN}^2 are invariant under similarity transformations and equal to the squared canonical correlation coefficients $\{\gamma_i\}_{i=1}^n$ between the past and future of the stochastic process $\{y_k\}$.

proof: Let T be a similarity transformation such that $[F, G, H]_n \xrightarrow{T} [TFT^{-1}, TG, HT^{-1}]_n = [F, G, H]_n$, $P \xrightarrow{T} TPT^T = P$, and $N \xrightarrow{T} T^{-T}NT^{-1} = N$ are the similarity relations between the original and transformed systems. Then

it follows that $\Sigma_{PN}^2 \xrightarrow{T} T\Sigma_{PN}^2T^{-1} = \Sigma_{PN}^2$ are similar matrices; therefore, they must have the same eigenvalues. This proves the first part. The second

part follows by substituting for $P = C(R^-)^{-1}C^T$ and $N = O^T(R^+)^{-1}O$ in (17) and making use of property (4) of Lemma 1, followed by a singular value decomposition of H , and finally, applying some simple arguments from [12].

Lemma 7:

Given a symmetric stochastic realization characterized by the triple $(F, G, H)_n$, if the system is transformed to any one of the following balanced coordinates

- a) input-normal: $P = \Gamma$ and $N = I_n$
- b) output-normal: $P = I_n$ and $N = \Gamma$
- c) internally balanced: $P = N = \Gamma^{1/2}$

where $\Gamma = \text{diag}[\gamma_1^{1/2}, \gamma_2^{1/2}, \dots, \gamma_n^{1/2}]$ is the diagonal matrix of canonical correlation coefficients, then the condition $\Sigma_{PN}^2 = \Gamma$ always holds.

proof: Follows by inspection

Lemmas 6 and 7 are useful in balanced stochastic realization theory [10]-[12] since finding the balancing transformation amounts to finding an eigenvalue - eigenvector decomposition for Σ_{PN}^2 . For symmetric systems, however, one can use the properties of Lemma 3 to find Σ_{PN}^2 by solving only one Riccati equation as opposed to two in the general case. Furthermore, the results of Lemma 7 show that Σ_{PN} is an invariant parameter when the system is in balanced coordinates. We finally remark that the above symmetry results are the stochastic counterpart to the results in [1] - [4] for deterministic systems.

With the exception of [10], very little work has been done on symmetric

stochastic realizations. We hope that the above symmetry results, in addition to those in [10], motivate further work on symmetric balanced stochastic realizations. In [7], an algorithm has been developed for obtaining deterministic balanced realizations from the solution to the cross-Grammian equation. The authors are currently investigating the possibility of extending this algorithm to use the cross Riccattian for obtaining balanced stochastic realizations.

CONCLUSIONS

We have defined a new cross Riccattian matrix, Σ_{PN} , which contains properties from both a forward and a backwards innovations representation. It was shown that Σ_{PN} satisfies a cross Riccati equation which is related to the forward and backwards Riccati equations by a sign matrix. For balanced stochastic realizations, this implies some computational savings since one need not solve a pair of Riccati equations while computing the balancing transformations. Furthermore, the cross Riccattian matrix and the associated Hankel matrix share some common properties which arise naturally in realization theory.

REFERENCES

- [1] K. V. Fernando and H. Nicholson, "Minimality of SISO Linear Systems", Proc. IEEE, Vol. 70, pp. 1241-1242, Oct. 1982.
- [2] K. V. Fernando and H. Nicholson, "On the Structure of Balanced and Other Principal Representations of Linear Systems", IEEE Trans. Automat. Contr., vol. AC-28, pp. 228-231, Feb. 1983.
- [6] K. V. Fernando and H. Nicholson, "On the Cauchy Index of Linear Systems", IEEE Trans. Automat. Contr. Vol. AC-28, pp. 222-224, Feb. 1983.
- [4] A. J. Laub, L. M. Silverman, and M. Verma, "A Note on Cross Grammians for Symmetric Realizations", Proc. IEEE, Vol. 71, pp. 904-905, July 1983.
- [5] J. A. DeAbreu - Garcia and F. W. Fairman, "A Note on Cross Grammians for Orthogonally Symmetric Realizations", IEEE Trans. Automat. Contr., vol. AC-31, pp. 866-867, Sept. 1986.
- [3] K. V. Fernando and H. Nicholson, "On the Cross Gramian for Symmetric MIMO Systems", IEEE Trans. Circuits Syst., vo. CAS-32, pp. 487-489, May 1985.
- [7] K. V. Fernando and H. Nicholson, "Computation of Balanced Realizations from Transfer Functions", Proc. IEE, Vol. 131, Pt. D, pp. 203-205, Sept. 1984.
- [8] B. D. O. Anderson and J. B. Moore, Optimal Filtering. Englewood Cliffs, N.J.: Prentice-Hall, 1979.
- [9] J. A. Ramos, "A Stochastic Realization and Model Reduction Approach to Streamflow Modeling", Ph.D dissertation, Dept. of Civil Eng., Georgia Institute of Technology, Atlanta, GA 1985.
- [10] U. B. Desai and D. Pal, "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations", Proc. of the 21st IEEE Conf. on Decision and Contr., Orlando, FL, pp. 1105-1112, 1982.
- [11] J. A. Ramos and E. I. Verriest, "A Unifying Tool for Comparing Stochastic Realization Algorithms and Model Reduction Techniques", Proc. Amer. Contr. Conf., San Diego, CA, pp. 150-155, 1984.
- [12] U. B. Desai, D. Pal, and R. D. Kirkpatrick, "A Transformation Approach to Stochastic Model Reduction", IEEE Trans. Automat. Contr., Vol. AC-29, pp. 1097-1100, Dec. 1984.
- [13] G. S. Sidhu and U. B. Desai, "New Smoothing Algorithms Based on Reversed-time Lumped Models", IEEE Trans. Automat. Contr., Vol. AC-21, pp. 538-541, Aug. 1976.
- [14] B. D. O. Anderson and T. Kailath, "Forwards, Backwards, and Dynamically Reversible Markovian Models of Second-order Processes", Proc. Int. Conf. Circuits and Syst., New York, N.Y. pp 981-986, 1978.

APPENDIX S

PROJECTION TECHNIQUES FOR MODEL REDUCTION

PROJECTION TECHNIQUES FOR MODEL REDUCTION

Erik Verriest

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

The emphasis of this paper is on the geometrical theory behind the stochastic realization problem, and its application to (stochastic) model reduction. It is shown that known stochastic realization algorithms based on canonical correlations or principal components can be analyzed in a common framework using certain operator valued measures. This analysis is based on the concept of the RV-coefficient as introduced in multivariable statistics as a measure for the similarity between two sets of random variables. It is shown that the theory has some parallels with the foundations of (geometric) quantum mechanics.

INTRODUCTION

The stochastic realization problem deals with the quest for a finite dimensional Markovian representation for a stochastic process from the known covariance information. If the covariances of the intervening random variables are exactly known, then we deal with the exact stochastic realization problem, which has received great attention [1,3,4,5,10]. This is primarily due to its fundamental importance in system identification, digital filtering, signal processing, and time series modeling [16]. 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 partially motivates the search for smaller dimensional Markovian realizations which approximate the original (or exact) one in some sense. The high dimensionality of the original (exact) state space model can, for instance, be caused by the incorporation of weakly coupled superfluous state components. These components may mask any underlying physical principles or tendencies hidden in the dynamics.

Another difficulty with the stochastic realization problem is the necessity of the exact covariance information. In most practical situations, all one has available is an estimate of the covariances based on the real data (i.e., sample covariances). Not only would the noise fluctuations in the covariance structure lead to models of high dimensions, but what is more essential, the sample covariance sequence may not be positive-real. In such a case, the exact realization algorithm applied to inexact data may not have a solution at all [3].

Akaike has 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. Here the pair of canonical vectors with positive canonical correlations form a minimal interface between the past and the future of the process. It is shown that these two canonical vectors are the states of the forward and backward innovations representation (extreme Markovian representations) introduced by Faurre [10] and are also basis vectors of what Akaike calls the forward and backward predictor spaces, respectively. Moreover, the canonical correlations coefficients provide a rational basis for obtaining the

reduced order model. Baram [4] extended Akaike's result to the nonstationary case and considered the model reduction problem by deleting the insignificant singular values from a singular value decomposition of the Hankel covariance matrix. A similar algorithm for obtaining the stochastic realization and reduced order model called the Canonical Realization Algorithm (CRA) was introduced by Desai and Pal [5]. It is a further extension to Akaike's work by introducing the concept of balanced stochastic realizations. Here a forward-backwards dual pair is obtained with state covariance matrices being equal and diagonal and these diagonal elements are the canonical correlation coefficients. A forward-backward pair that satisfies these conditions is said to be in balanced form. These balancing conditions are the stochastic counterpart to the deterministic balancing conditions originally proposed by Moore [17] and for the time varying case by Verriest and Kailath [23].

Recently Arun and Kung [3] introduced the Karhunen-Loeve Method (KLM) which also has its grounds on Multivariate statistics and in the context used here is shown to be equivalent to Principal Components of instrumental variables as discussed in Rao [21]. KLM optimizes the approximation of the information interface between the past and the future of a stochastic process via a one sided Karhunen-Loeve Expansion (KLE) of the predictor space. A break in the diagonal elements of the state covariance matrix dictates the reduced order model. As opposed to CA, KLM is not symmetric in the sense that either a forward or a backwards Markovian representation is obtained. This would constitute two separate problems. Arun and Kung [3] also pointed out that CRA is not well suited for model reduction due to the smallness of the canonical correlations and the fact that it works with unapproximated data. Despite the theoretical facts, no direct comparison between KLM and CRA has been reported justifying Arun and Kung's argument nor has there been any statistical measure of information common to both models that would aid the modeler to discriminate against CRA and KLM when dealing with the model reduction problem.

Ramos and Verriest explored in an earlier paper [19] the combination of the canonical correlation and principal component analyses, given the exact covariance information, in a common framework using the RV-coefficient introduced by Escoufier [6]. It was shown that this common statistical measure of information provides a rationale for drawing inferences about the performance of the algorithms. In the RV-coefficient framework, linear transformations on sets of random variables are found, so that the transformed sets are as similar as possible in a certain sense. The RV-measure attains values in $[0,1]$ and the closer to one, the greater the similarity of the sets of random variables.

The motivation for this paper is to derive some more geometrical insight in the problem, in order to adapt the method for use in the approximate stochastic realization case, based on real data.

In the following, we therefore briefly summarize the stochastic realization to set the necessary background. This is followed by a brief discussion of the canonical correlation and principal component analysis. Next, the RV-coefficient is introduced in a geometrical context, for the exact covariance data and for the real data cases. A unifying framework is developed, and a connection is made with some of the foundations of quantum mechanics. Finally, the RV-technique is illustrated for the CCA and PCA.

THE DISCRETE STOCHASTIC REALIZATION PROBLEM

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

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

$$\bar{y}_k = Hx_k + v_k \quad (2)$$

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

$$E \begin{bmatrix} w_k \\ v_k \end{bmatrix} \begin{bmatrix} w'_e & v'_e \end{bmatrix} = \begin{bmatrix} Q & S \\ S' & R \end{bmatrix} \delta_{k,l} ; \quad v_{k,l} \quad (3)$$

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

The solution to this problem is well known [10]. Given the covariance sequence, one forms the (infinite) Hankel matrix

$$\hat{H} = \begin{pmatrix} \Lambda(1) & \Lambda(2) & \dots \\ \Lambda(2) & \Lambda(3) & \\ \vdots & & \end{pmatrix} \quad (4)$$

The time sequence is rational if and only if this Hankel matrix has finite rank (say n). It follows then from the deterministic realization theory [10] that the order of any minimal Markovian representation of $\{y_k\}$ is precisely n , and a triple (F,G,H) can be constructed such that

$$\begin{aligned} \Lambda(k) &= HF^{k-1}G + \Lambda_0 \delta_{k0} , & k > 0 \\ \Lambda(k) &= \Lambda'(-k) , & k < 0 \end{aligned} \quad (5)$$

where in order to have a Markovian representation, the following needs to be satisfied.

$$P - FPF' = Q \quad (6)$$

$$G - FPH' = S \quad (7)$$

$$\Lambda_0 - HPH' = R \quad (8)$$

$$\begin{bmatrix} Q & S \\ S' & R \end{bmatrix} > 0 , \quad P > 0 \quad (9)$$

Here P is interpreted as the state covariance matrix

$$P = E(x_k x'_k) \quad (10)$$

The triple (F,G,H) together with Λ_0 do not uniquely specify the covariances P,Q,S , and R . However, P completely specifies Q,S , and R , and therefore characterizes the Markovian representation. Furthermore, note that any minimal realization of the covariance sequence is unique, modulo a similarity transformation.

In the stationary sequence realization problem, the past and the future can be brought on equal footing, since the statistical properties of the given sequence are invariant with respect to time inversion. There are, thus, two classes of representations: the forward and the backward representations. The forward Markovian representations have the causal structure; or forward propagation property:

$$E(x_t v'_s) = 0 \quad \forall s > t$$

Similarly, the backwards models have the anti-causal structure.

$$E(x_t v'_s) = 0 \quad \forall s < t$$

Denoting by Π the set of state covariances defining a forward Markovian model with triple (F,G,H) , and by $\bar{\Pi}$ the corresponding set of state covariances for the

backward models with triple (F, G, H) , the following can be asserted about the sets Π and $\bar{\Pi}$. Note that both sets only contain positive definite matrices.

1. Both sets are closed, bounded, convex, and have two extreme points.

$$\begin{aligned} P_* &< P < P^* \\ P^* &> P_* \end{aligned} \quad \text{for } \Pi$$

$$\begin{aligned} \bar{P}_* &< \bar{P} < \bar{P}^* \\ \bar{P}^* &> \bar{P}_* \end{aligned} \quad \text{for } \bar{\Pi}$$

2. There exists an order isomorphism (matrix inversion) between the partially ordered sets $(\Pi, <)$ to $(\bar{\Pi}, >)$. Thus,

$$P \in \Pi \Leftrightarrow P^{-1} \in \bar{\Pi} \quad (11)$$

$$P^* = \bar{P}_*^{-1} \quad (12)$$

$$\bar{P}^* = P_*^{-1} \quad (13)$$

It can be shown [1] that the extreme points P_* and $\bar{P}_*$ respectively correspond to the forward and backward innovations-representations.

APPROACHES TO MARKOVIAN MODELING

Assume that the stochastic time-series $\{y_k\}$ is Gaussian (with zero mean). The relevant random variables are then in the Hilbert space $L_2(\Omega, \mathcal{B}, P)$ and conditional

expectations can be interpreted as orthogonal projections onto subspaces

$$L_2(\Omega, \mathcal{F}_k^{\{y_k\}}, P).$$

For the time-series $\{y_k\}$ define as in [10] the infinite vectors.

$$y_k^+ = \begin{pmatrix} y_k \\ y_{k+1} \\ \vdots \end{pmatrix}, \text{ the future} \quad (14)$$

$$y_k^- = \begin{pmatrix} y_{k-1}^* \\ y_{k-2}^* \\ \vdots \end{pmatrix}, \text{ the past} \quad (15)$$

and define the semi-infinite covariance matrices

$$\hat{H} = E\{y_k^+(y_k^-)'\} \quad (16)$$

$$R^+ = E\{y_k^+(y_k^+)'\} \quad (17)$$

$$R^- = E\{y_k^-(y_k^-)'\} \quad (18)$$

Within this representation, the forward and backward predictor subspaces are

$$X_k = \text{Span}(Y_k^+ | Y_k^-) \quad (19)$$

$$Z_{k-1} = \text{Span}(Y_k^- | Y_k^+) \quad (20)$$

(A|B) denotes the projection of span (A) onto the Hilbert space spanned by the Components of B. From the projection theorem, one obtains

$$X_k = \hat{H}(R^-)^{-1} Y_k^- \quad (21)$$

$$Z_{k-1} = \hat{H}'(R^+)^{-1} Y_k^+ \quad (22)$$

Under the assumption that a finite dimensional Markovian representation exists, finite dimensional bases X_k^* and Z_k^* can be found such that they respectively generate X_k and Z_k , i.e., there exists operations A and B such that

$$X_k^* = A' X_k = A' \hat{H}(R^-)^{-1} Y_k^- = M' Y_k^- \quad (23)$$

$$Z_{k-1}^* = B' Z_{k-1} = B' \hat{H}'(R^+)^{-1} Y_k^+ = L' Y_k^+ \quad (24)$$

Since the basis vectors must be linearly independent, their covariances must be nonsingular. We may impose the constraints

$$E X_k^* X_k^{*'} = \Delta_x = \text{diag}(\delta_{x_1}) \quad (25)$$

$$E Z_{k-1}^* Z_{k-1}^{*'} = \Delta_z = \text{diag}(\delta_{z_1}) \quad (26)$$

by suitably redefining the A and B (multiplication with an orthogonal matrix on the right.) The stochastic realization problem is thus equivalent to the problem of finding matrices L and M such that the similarity between the predictor spaces is as large as possible, while satisfying the constraints (25) and (26). For the stochastic model reduction problem, different statistical techniques have been proposed, notably the canonical correlation method and the principal component analysis.

The application of the Canonical Correlation Analysis to the realization problem has been pioneered by Akaike [1], Baram [4], and Desai and Pal [5], who tied this in with balancing techniques. By the above analysis, the state is in fact the information interface between past and future. The canonical correlations lead, therefore, to a natural distance measure between the past and the future, which in the Gaussian case is nothing else than the Kullback-Leibler mutual information.

Arun and Kung [3] used a different approach. In [3], the past is treated as instrumental values for predicting the future. (The Principal Component Analysis is also named Instrumental Variable, or Karhunen-Loeve method [21]). The model reduction method is then based on retaining the components of the past that have a significant contribution to its efficiency in predicting the future.

The two methods are obviously not equivalent, and therefore problems and some critique on each have been pointed out. First of all, the interpretation of the canonical correlations has been questioned [3]. A relatively strong canonical correlation between two components is possible, however, they may not extract significant portions of the (total) variance. Another critique is that the canonical correlations are rather small numbers to compare, since they must obviously belong to the interval [0,1]. This is probably not such a sharp disadvantage, since only relative magnitudes are important anyway. Finally, since the

canonical correlation technique is an exact covariance realization technique, and only sample covariances are available in any real situation, the robustness of the realization procedure may be at stake. On the other hand, the proponents of the canonical correlation analysis point out the nice way in which the connection between Y^- and Y^+ is displayed. This paper will hopefully resolve such a dispute between opponents and proponents of either method, by giving a common framework which puts both analyses on equal footing. As was already remarked in the introduction, the analyst can then decide which features are important for the situation at hand, and make an appropriate decision.

STATISTICAL PRELUDE: THE GEOMETRIC NATURE OF MULTIVARIATE ANALYSIS

Let X be a random vector of dimension p over a probability space (Ω, β, P) with finite second order moments. Thus X belongs to the Hilbert space $L_2^p(\Omega, \beta, P)$. In this setting, the Principal Component Analysis consists of finding a transformation in $L_2^p(\Omega, \beta, P)$ taking the vector X into Y such that the components of Y are uncorrelated and

$$E(y_i y_j) = \lambda_i \delta_{ij} \quad (27)$$

where λ_i is the i -th eigenvalue of the matrix $E XX'$. Note that if $X = AY$, then

$$E XX' = A A A' \quad (28)$$

and thus

$$E x_i x_j = \sum_{k=1}^p a_{ik} a_{jk} \lambda_k \quad (29)$$

which "explains" the (co)variance of the components of X .

The Canonical Correlation Analysis on two random vectors $X^{(1)}$ and $X^{(2)}$, not necessarily of the same dimension, finds linear combinations of $X^{(1)}$ and $X^{(2)}$ with the largest covariances in a certain way. It, therefore, "explains the connection" between the two random vectors.

The important remark that we want to make here is that both techniques rely on the Hilbert space structure: the inner product in the space is derived from the covariance. In the statistical literature, a new measure has been introduced by Escoufier [6] which is scalar and expresses the "similarity" between subspaces. Let X be the p dimensional random vector partitioned as

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \in L_2^p \quad (30)$$

with covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad (31)$$

Escoufier defined the following scalar measures [3]:

$$\text{COVV}(X^{(1)}, X^{(2)}) = \text{Tr} \Sigma_{21} \Sigma_{12} \quad (32)$$

$$\text{VARV}(X^{(1)}) = \text{Tr} \Sigma_{11}^2 \quad (33)$$

$$\text{RV}(X^{(1)}, X^{(2)}) = \frac{\text{COVV}(X^{(1)}, X^{(2)})}{[\text{VARV}(X^{(1)}) \text{VARV}(X^{(2)})]^{1/2}} \quad (34)$$

The properties of the above measures are analogous to the usual covariance properties since for any other related $x^{(3)}$

1. $\text{COVV}(x, x^{(3)}) = \text{COVV}(x^{(1)}, x^{(3)}) + \text{COVV}(x^{(2)}, x^{(3)})$
2. If $x^{(1)} = AX^{(2)}$; where $A \in \mathbb{R}^{n_1 \times n_2}$; $AA' = kI$ and $n_1 < n_2$, then

$$\text{COVV}(x^{(2)}, x^{(3)}) = k \text{COVV}(x^{(1)}, x^{(3)})$$

Note that it follows at once from 2 that the RV-measure is invariant with respect to scaling and orthogonal transformations. However, the introduced scalar measures are not quite the same as a covariance, since if $x^{(1)}$ and $x^{(2)}$ are scalar, then

$$\text{COVV}(x_1, x_2) = (E x_1 x_2)^2$$

Other measures have been introduced previously, e.g., Hotelling's Vector Correlation Coefficient (VCC), defined as [12]

$$(\text{VCC})^2 = \frac{\begin{vmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}}{\begin{vmatrix} \Sigma_{11} & \\ & \Sigma_{22} \end{vmatrix}} \quad (35)$$

This measure has several drawbacks. First of all, it does not support the degenerate case $x^{(1)'} = (x^{(2)'} | z^{'})'$. The RV-measure on the other hand supports this degenerate case. Furthermore, any intuitive notion of similarity between two random vectors (of arbitrary, not necessarily equal, dimension) should be nonzero as long as there is some correlation between at least one of the components of each vector. Moreover, it may be advantageous to let the difference in dimensionality of each vector influence the similarity. An increasing difference in dimensionality should decrease the similarity. In order to compare the two defined measures in this respect, let the covariance matrix be partitioned as

$$\Sigma = \begin{bmatrix} I_{n_1} & \Delta \\ \Delta' & I_{n_2} \end{bmatrix}; \quad n_1 > n_2 \quad (36)$$

where

$$\Delta = \text{diag}(\delta_1 \cdots \delta_{n_2})$$

then

$$1. \quad \text{RV}(x^{(1)}, x^{(2)}) = \sum_{i=1}^{n_2} \delta_i^2 \frac{1}{\sqrt{n_1 n_2}} \quad (37)$$

$$2. \quad \text{If } n_2 = 1, \text{ then } \text{RV}(x^{(1)}, x^{(2)}) = \delta_1^2 / \sqrt{n_1} \quad (38)$$

the right hand side of (38) decreases as the dimension n increases, thus indicating the decreasing "similarity" of the random vectors as n increases. The Vector Correlation Coefficient gives respectively

$$1. \quad (\text{VCC})^2 = \prod_{i=1}^{n_2} \delta_i^2 \quad (39)$$

$$2. \quad \text{VCC is independent of } n_1 \text{ (if } n_1 \geq n_2 \text{)}.$$

These properties may provide the prime motivation to consider RV as a measure of similarity, but are not sufficient for a justification to take it out of the ad-hoc status.

The theoretical justification is due to Escoufier. The rigorous mathematical construction is as follows. First, characterize the random vector X in L_2^p by an operator on L_2 . Next, show that the set of such operators has the structure of a Hilbert space under the inner product $\text{COVV}(\cdot, \cdot)$. Then the usual induction

inner product $\rightarrow$ norm $\rightarrow$ distance

leads to a rigorous definition of the "similarity." In particular, we have the following definition of the operator associated to X .

Definition: Associated Operator. With $X \in L_2^p$ associate an operator $U_X : L_2 \rightarrow L_2$ defined by

$$\forall y \in L_2 : U_X(y) \triangleq \sum_{i=1}^n [E(x_i y)] x_i \quad (40)$$

The following propositions follow then at once from the definition. The proofs are in [3].

Proposition 1: U_X characterizes X in the sense of eigenvalues and eigenvectors. More precisely, if Σ is the covariance matrix of X , then

$$1. \quad \forall z \in \mathbb{R}^n \text{ s.t. } \Sigma z = \lambda z \quad (41)$$

the random variable $y = x'z$ satisfies

$$U_X(y) = \lambda y \quad (42)$$

$$2. \quad \text{Conversely, } \forall y \in L_2 \text{ such that } U_X(y) = \lambda y$$

$$y = x'z \text{ where } \Sigma z = \lambda z$$

Proposition 2:

$$1. \quad U_X \text{ is a Hilbert-Schmidt operator.}$$

$$2. \quad \text{The set of operators } \{U_X\} \text{ forms a Hilbert space under the inner product}$$

$$\langle U_1, U_2 \rangle = \sum_{\text{CONS}} E(U_1(\phi_k), U_2(\phi_k)) \quad (43)$$

where "CONS" stands for a complete orthonormal system of basisvectors $\{\phi_k\}$ in the space L_2 (which is separable). Note that one has also

$$\langle U_{x_1}, U_{x_2} \rangle = \sum_{\text{CONS}_1} \sum_{\text{CONS}_2} \mu_i \lambda_j E(\phi_i \psi_j) \quad (44)$$

where the (μ_i, ϕ_i) and (λ_j, ψ_j) solve the eigen problems (41) associated respectively with X_1 and X_2 .

Proposition 3:

$$\langle U_{x_1}, U_{x_2} \rangle = \text{Tr}(\Sigma_{21} \Sigma_{12}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} E[x_1^{(1)} x_j^{(2)}]^2 \quad (45)$$

This proposition follows from the invariance of the inner product (43) with respect to orthogonal transformations.

Define now a "correlation coefficient" $\gamma(U_1, U_2)$ by

$$\gamma(U_1, U_2) = \frac{\langle U_1, U_2 \rangle}{\|U_1\| \|U_2\|} \quad (46)$$

Note that the special case $\gamma = 0$ occurs iff the eigenspaces of U_1 and U_2 are orthogonal to each other, while $\gamma = 1$ iff U_1 and U_2 have the same eigenspace, the same eigenvectors, and proportional eigenvalues. The correlation between the operators associated with the random vectors is thus

$$\frac{\text{Tr}(\Sigma_{21} \Sigma_{12})}{\sqrt{\text{Tr} \Sigma_{11}^2 \text{Tr} \Sigma_{22}^2}} \quad (47)$$

which is defined as the RV-coefficient between X_1 and X_2 .

REAL DATA CASE (SIGNAL PROCESSING CONTEXT)

Consider the case where a p -dimensional vector is measured n times. It is customary to say that we have p variables, and n samples. Let these then be organized in a datamatrix

$$X = [x_1 \cdots x_n] \text{ in } R^{p \times n} \quad (48)$$

This can also be represented geometrically as a configuration $C(X)$ of n points in R^p (and of course dually as a configuration of p points in R^n , but only the first will be discussed). Let the distance between points in R^p be derived from the metric

$$\langle x_i, x_j \rangle = x_i' Q x_j \quad (49)$$

with Q positive definite. Then $Q = LL'$, leading to the equivalent interpretation as the Euclidean metric on the transformed variables.

$$y = L'x$$

Different weights may be attached to the different samples (e.g. according to some measure of the accuracy of the obtained measurement). Let $p = \{p_i, i=1, \dots, n\}$ be such a set of weights for which $p_i > 0$;

$$\sum_{i=1}^n p_i = 1$$

(i.e. $\{p_i\}$ is a probability measure). A weighted average of the data points can be defined

$$\langle x \rangle_p = \sum_{i=1}^n p_i x_i$$

Whenever one has the configuration $C(x)$ together with a measure $\{p_i\}$ it will be advantageous to consider the "centered" data matrix whose columns are $x_i - \langle x \rangle_p = x_i$. Its importance, of course, is to obtain translation invariant properties.

The classical multivariable methods consist now in a search for linear transformations of the original variables that minimize (under some constraints) the "closeness" of the configurations $C(X)$ and $C(Y)$. But how can a distance between configurations of points be defined? Ideally, such a measure should be invariant with respect to translation, rotation and scaling, and this should thus a priori hold for the "self-distance" or distance of the relative positions in $C(X)$. It

is easily seen that the Euclidean distance matrix (i.e., with ij -element $[(x_i - x_j)'Q(x_i - x_j)]^{1/2}$) is a (matrix-valued) measure which is invariant with respect to translation and rotation, but not with respect to scaling. On the other hand, the matrix-valued measure

$$\frac{S_p(X)}{\sqrt{\text{Tr } S_p(X)^2}} \quad (51)$$

where $S_p(X) = \text{diag}(\sqrt{p}) \tilde{X}' \tilde{X} \text{diag}(\sqrt{p})$ is invariant with respect to translation, rotation, and scaling. Note that for A, B in $R^{n \times n}$, $\langle A, B \rangle = \text{Tr } A'B$ defines an inner product on $R^{n \times n}$ for which the induced norm is the Frobenius or F-norm. The distance between the (measured) configurations $(C(X), p)$ and $(C(Y), p)$ is then induced by this norm., i.e.,

$$\begin{aligned} d^2((C(X), p)(C(Y), q)) &= \left\| \frac{S_p(X)}{\sqrt{\text{Tr } S_p(X)^2}} - \frac{S_p(Y)}{\sqrt{\text{Tr } S_p(Y)^2}} \right\|_F^2 \\ &= 2(1 - RV(X, Y)) \end{aligned} \quad (52)$$

if one defines

$$RV(X, Y) = \frac{\text{Tr } S_p(X) S_p(Y)}{\sqrt{\text{Tr } S_p(X)^2} \sqrt{\text{Tr } S_p(Y)^2}} = \frac{\text{Tr } S_{12} S_{21}}{\sqrt{\text{Tr } S_{11}^2} \sqrt{\text{Tr } S_{22}^2}} \quad (53)$$

Since S can be interpreted as a sample covariance, this last definition shows the neat similarity between the signal processing (53) and stochastic realization (47) contexts.

UNIFYING FRAMEWORK

In the previous sections, we have shown how the (exact) stochastic realization and the (real) signal modeling benefit from the use of a certain measure with similar interpretation. Both have been introduced by Escoufier [7,8]. In this section, a more abstract representation is developed. The motivation starts from the observation that for the stochastic realization problem, the underlying space $L_2^p(\Omega, \mathcal{B}, \rho)$ and in the real data case, the space $R^{p \times n}$ are isomorphic with the tensorproduct spaces, respectively

$$L_2^p(\Omega, \mathcal{B}, \rho) \sim R^p \otimes L_2(\Omega, \mathcal{B}, \rho) \quad (54)$$

$$R^{p \times n} \sim R^p \otimes R^n \quad (55)$$

In general now, let G and H be separable Hilbert spaces. Let $\{\phi_i\}$ be a CONS in G , and $\{\psi_i\}$ a CONS in H , then any vector x in the tensorproduct space $G \otimes H$ has a decomposition

$$x = \sum_i x_i \otimes \phi_i \quad (56)$$

with $x_i \in H$. In this framework, we define the Associated Operator and Gramian:

Definition: With the decomposition of $x \in G \otimes H$ as in (56), the associated operator U_x from H to H is defined as

$$U_x : H \rightarrow H : y \mapsto \sum_i \langle y, \phi_i \rangle x_i \quad (57)$$

Although this definition is given with reference to a specific coordinate system, the following theorem is easily shown.

Theorem: The associated operator U_x is

1. Independent of the choice of a CONS in G .
2. Bounded, self-adjoint and positive semidefinite.
3. Hilbert Schmidt (i.e., $\sum_i \|U_x(\psi_i)\|_H^2 < \infty$)

Definition: Given x in $G \otimes H$, the Gramian G_x is the operator G_x from G to G , $g \mapsto G_x(g)$ such that

$$G_x(\phi_i) = \sum_j \langle x_i, x_j \rangle \phi_j \quad (58)$$

Note that the above gives a definition of G_x via its action on a CONS. Again, the following is easy to show:

Theorem: The above defined Gramian G_x

1. Is coordinate independent.
2. Is bounded, self-adjoint, positive semidefinite and Hilbert-Schmidt.
3. Has normalized eigenvectors which form a CONS in G .

Main Property: The operators U_x and G_x have the same eigenvalues (counting their multiplicity). The eigenspace of G_x corresponds to the eigenspace of U_x under the mapping

$$\begin{aligned} X : G \rightarrow H : Z \mapsto X(Z) \\ X(Z) = \sum_j Z_j x_j \end{aligned} \quad (59)$$

Definition: The "correlation" between x_1 in $G_1 \otimes H$ and x_2 in $G_2 \otimes H$ is (with G_1 and G_2 subspaces of G)

$$(x_1, x_2) = \frac{\Delta}{\|x_1\| \|x_2\|} Y(U_{x_1}, U_{x_2}) \quad (58)$$

where $Y(U_{x_1}, U_{x_2})$ is the correlation in the Hilbert space $L(H)$ of the associated operators.

Theorem: The correlation defined above is

1. Independent of the chosen CONS.
2. Invariant w.r.t. orthogonal transformations, scale transformations, and translations.
3. $\langle U_{x_1}, U_{x_2} \rangle_{L(H)} = \text{Tr}(G_{21} G_{12})$ (61)

It follows now from this last theorem that

$$(x_1, x_2) = \frac{\text{Tr } G_{21} G_{12}}{\sqrt{\text{Tr } G_{11}^2 \text{Tr } G_{12}^2}} \quad (62)$$

Remark: A complete duality exists between the roles played by G and H .

GEOMETRICAL INTERPRETATION: CORRELATION BETWEEN SUBSPACES

Let G be the Hilbert space containing the "observations" $\{y_i\}$. Denoting by $M \wedge N$ the largest subspace contained in both M and N , and by $M \vee N$ the smallest subspace containing both M and N , the set of subspaces which are closed under $\wedge$ and $\vee$ have a lattice structure. If $M^\perp$ denotes the usual orthogonal complement in G , then the subspaces of G form a complete orthocomplemented lattice (or logic [22]). It is postulated that the propositions of a physical system are a complete ortho

complemented lattice [15]. Since the propositional calculus of a physical system has a similarity to the corresponding calculus of formal logic, one refers to this often as "quantum logic."

The set of subspaces of a Hilbert space is isomorphic to the set of orthoprojectors $\text{Proj } G$ on G . With each subspace, there corresponds one projector (namely the one whose range is exactly that subspace), and vice versa. Consider now the vector valued measure on the subspaces (projectors)

$$\xi_Y(A) \stackrel{\Delta}{=} P^A Y \quad (63)$$

where P^A denotes the orthoprojector on subspace A . This ξ has the interesting properties that if $A \perp B$, then $\xi(A) \perp \xi(B)$, which is the orthogonal scattering property. Furthermore, if $\{P^i\}$ is a set of pairwise orthogonal subspaces (projectors), then

$$\sum_i \xi_Y(P^i) = \xi_Y\left(\sum_i P^i\right) \quad (64)$$

The last property is characteristic for a Gleason measure [13], so that ξ is referred to as an Orthogonally Scattered Gleason (OSG) measure [14]. This OSG-measure induces a scalar Gleason measure

$$\mu_Y(A) = \|\xi_Y(A)\|_G^2 \quad (65)$$

Interpreting the right hand side of (65) as a variance, it is natural to introduce a "covariance" between subspaces as

$$\langle \xi(A), \xi(B) \rangle_G \quad (66)$$

By Gleason's theorem [11], there exists a self-adjoint, trace class operator T_Y such that

$$\mu_Y(A) = \text{Tr}(T_Y P^A) \quad (67)$$

Note that here simply $T_Y = YY'$.

Consider now a family of vectors y_i and their corresponding measures ξ_i . Introducing a measure p on the y_i a "mixture" gleason measure (or linear superposition) is obtained

$$\xi = \sum_i p_i \xi_i \quad (68)$$

Aragon and Couot [2] show that with any such superposition, there corresponds again an operator

$$T = \sum_i p_i T_i = \sum_i p_i y_i y_i' \quad (69)$$

The correlation between the subspaces A and B is then [14]

$$\frac{\langle \xi(A), \xi(B) \rangle}{\sqrt{\mu(A) \mu(B)}} = \frac{\text{Tr}(TP^A P^B)}{\sqrt{\text{Tr}(TP^A) \text{Tr}(TP^B)}} \quad (70)$$

Note that T can be interpreted as sample covariance S . Hence finding the subspace B of fixed dimension which best "approximates" the given space entails the maximization of (note that since $A=G$, then $P^A = I$)

$$\frac{\text{Tr}(SP^B)}{\sqrt{\text{Tr}(S) \text{Tr}(SP^B)}} = \frac{\text{Tr}(S_B)}{\sqrt{\text{Tr}(S) \text{Tr}(S_B)}} = \sqrt{\frac{\text{Tr}(S_B)}{\text{Tr}(S)}} \quad (71)$$

for $S_B = P^B S P^B$. Clearly the optimal projection P^B is the one projecting on the eigenspace of S with the largest principal components. If instead of a discrete

measure P , one has a continuous measure p , then (70) remains valid if one replaces the (sample) covariance S by the exact covariance matrix Σ for $G = L_2^p(\Omega, \mathcal{B}, P)$.

The principal component analysis for the exact covariance and real data cases follow thus nicely from this formalism. What is the connection between this formalism and the operators developed in the general RV-theory? Rewriting (69) as

$$T = \sum_i \sqrt{p_i} y_i \sqrt{p_i} y_i' = \sum_i \hat{y}_i \hat{y}_i' \quad (72)$$

and defining the data matrix Y as $\text{diag}(\sqrt{p_i}) Y'$ the operator can be written as

$$T(\cdot) = \sum_{i=1}^n \langle \hat{y}_i, \cdot \rangle_{R^P} \hat{y}_i = U_Y(\cdot) \quad (73)$$

where (57) is used (for the dual case) for $G = R^n$ and $H = R^P$. But then for $A, B \in \text{Proj } R^P$ we get

$$\begin{aligned} \langle TP^A, TP^B \rangle_{L(R^P)} &= \text{Tr } TP^A TP^B \\ &= \text{covv}(A, B) \end{aligned} \quad (74)$$

thus leading to the earlier defined RV coefficient. Note that (74) defines a "covariance" between operators whose existence follows from Gleason's theorem, while in (66) we considered directly the "covariance" between the OSG-measures itself. T plays the role of the exact or sample covariance matrix, which is in fact the representation of the Gramian defined earlier.

APPLICATIONS

The results in this section have been reported in [18] and [19]. They are included for completeness. Consider the:

General Problem: Given $x \in R^{p \times n}$ and $y \in R^{q \times n}$. Find transformations L and M (i.e., metrics), such that $L'X$ and $M'Y$ are as "similar" as possible. As a measure for similarity, the use of the RV-coefficient was motivated,

$$RV(L'X, M'Y) = \frac{\text{Tr}(L'S_{12}MM'S_{21}L)}{\sqrt{\text{Tr}(L'S_{11}L)^2 \text{Tr}(M'S_{22}M)^2}} \quad (75)$$

Remarks:

1. Since for any orthogonal matrix R , the substitution of L by LR leaves RV invariant, one may assume without loss of generality that the matrices $L'S_{11}L$ and $M'S_{22}M$ are diagonal.
2. Instead of finding optimal transformations, equivalent problems are to find a proper metric or a set of projection hyperplanes.

The familiar statistical modeling techniques can now be brought on a common ground using this powerful tool [8] and applied to the stochastic realization problem.

1. Generalized Double-Sided Stochastic Realization

Here we let $X = Y^+$ and $Y = Y^-$, then

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} R^+ & \hat{H} \\ \hat{H}^* & R^- \end{bmatrix} \quad (76)$$

With the states for the forward and the backward realization as in (23) and (24) and diagonal Δ_x and Δ_y the generalized double-sided stochastic realization problem [18,19] can be formulated as the maximization of

$$RV(L'Y^+, M'Y^-)$$

with respect to L and M and subject to

$$L'R^+L = \Delta_z \quad (77)$$

$$M'R^-M = \Delta_x \quad (78)$$

Using (diagonal) Lagrange multipliers Λ and Ψ , the problem is reduced to an unconstrained optimization problem. Its conditions for optimality are

$$\hat{H}M\hat{M}'HL - R^+\Lambda = 0 \quad (79)$$

$$\hat{H}'LL'\hat{H}M - R^-\Psi = 0 \quad (80)$$

It follows that

$$\Delta_z\Lambda = \Delta_x\Psi \quad (81)$$

which together with (79) and (80) leads to the generalized eigenvalue - eigenvector problem

$$H(R^-)^{-1}H'L = R^+\Gamma \quad (82)$$

$$H'(R^+)^{-1}HM = R^-\Gamma \quad (83)$$

with Γ the eigenvalue matrix. Γ is related to Λ and Ψ by [8]

$$\Lambda = \Delta_x\Gamma \quad (84)$$

$$\Psi = \Delta_z\Gamma \quad (85)$$

The optimal transformations L and M are the solution to

$$L = (R^+)^{-1}\hat{H}M\Gamma^{-1/2}\Delta_x^{-1/2}\Delta_z^{1/2} \quad (86)$$

$$M = (R^-)^{-1}\hat{H}L\Gamma^{-1/2}\Delta_z^{-1/2}\Delta_x^{1/2} \quad (87)$$

For the particular choices of Δ_x and Δ_z made, the maximal RV coefficients

$$RV(L'Y^+, M'Y^-) = \frac{\text{Tr } \Delta_x\Gamma\Delta_z}{\sqrt{\text{Tr } \Delta_x^2 \text{Tr } \Delta_z^2}} \quad (88)$$

Note that if one of Δ_x and Δ_z is the identity, then choosing the other as Γ , the diagonal of the squared canonical correlations, maximizes RV. Ramos [18] has shown that the input normal, the balanced [5], and the output normal [1] stochastic realization follow from this choice. Baram's realization [4] follows by taking both equal to I .

2. The General Double-Sided Stochastic Realization

This corresponds with the principal component analysis [3], constraining L to be the identity, thus maximizing

$$RV(Y^+, M'Y^-)$$

subject to (78).

The solution is

$$M = (R^-)^{-1} \hat{H}' Q \quad (89)$$

for some orthogonal matrix Q . The maximal RV is

$$RV(Y^+, M_*' Y^-) = \sqrt{\frac{\text{Tr } \hat{H} (R^-)^{-1} \hat{H}'}{\text{Tr } (R^+)^2}} \quad (90)$$

CONCLUSION

The RV -coefficient framework enables the unification of the theory of stochastic realization, and in fact provides a valuable tool to directly compare different modeling or model reduction schemes. The double sided solution involves canonical correlations, as opposed to variances on the one-sided solution. The formalism applies to the exact covariance as well as the real data case.

The realizations of the RV -coefficient to projection measures have been explored. This is currently being further investigated, and it is hoped that it will yield more insight into the dynamical aspects of the real data case.

REFERENCES

- [1] Akaike, A., Markovian Representation of Stochastic Processes by Canonical Variables, SIAM J. Control 13, No. 4, (1975) 162-173.
- [2] Aragon, Y. and Couot, J., Une Definition de l'Operateur d'Escoufier, Comptes rendus, 283, serie A (1976) 867-869.
- [3] Arun, K. S. and Kung, S. Y., A New Algorithm for Approximate Stochastic Realization, to be published.
- [4] Baram, Y., Realization and Reduction of Markovian Models from Nonstationary Data, IEEE Trans. Automatic Control AC-26, no. 6 (1981) 1225-1231.
- [5] Desai, U. B. and Pal, D., A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations, Proc. 21st Conf. on Decision and Control (1982) 1105-1111.
- [6] Escoufier, Y., Le Traitement des Variables Vectorielles, Biometrics 29, (1973) 751-760.
- [7] Escoufier, Y., Operators related to a Data Matrix, in: Barra J. R. et al. (eds.), Recent Developments in Statistics (North-Holland, Amsterdam, 1977).
- [8] Escoufier, Y. and Robert, P., Choosing Variables and Metrics by Optimizing the RV -Coefficient, in: Opt. Meth. in Statistics (Academic Press, 1979) 205-219.

- [9] Escoufier, Y., Robert, P. and Cambon, J., Construction of a Vector Equivalent to a Given Vector from the Point of View of the Analysis of Principal Components, in: Bruckmann et al. (eds.), Computational Statistics (Springer-Verlag, Wien, 1974) 155-164.
- [10] Faurre P. L., Stochastic Realization Algorithms, in: Mehra, R. K. and Lainiotis, D. G., (eds.), System Identification: Advances and Case Studies, (Academic Press, 1976).
- [11] Gleason A. M., Measures on the Closed Subspaces of a Hilbert Space, J. Math. Mech. 6 (1957), 885-893.
- [12] Hotelling H., Relations between two Sets of Variates, Biometrika 28 (1936) 321-377.
- [13] Jajte R., Gleason Measures, in: Bharucha-Reid (ed.), Probabilistic Analysis and Related Topics, Vol. 2 (Academic Press, 1979).
- [14] Jajte J. and Paszkiewicz A., Vector Measures on the Closed Subspaces of a Hilbert Space, Studia Mathematica, 58 (1978), 229-251.
- [15] Jauch J., Foundations of Quantum Mechanics (Addison-Wesley, Reading, 1968).
- [16] Kung, S. Y., Deterministic and Stochastic Linear System Approximations, with Applications to Model Reduction and Signal Processing, Joint US-Japan Math. Systems Seminar, Gainesville, Florida (1983).
- [17] Moore, B. C., Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, IEEE Trans. Automatic Control, AC-26, No. 5 (1) 17-32.
- [18] Ramos, J. A., A Stochastic Approach to Streamflow Modeling, Ph.D. Dissertation, School of Civil Engineering, Georgia Institute of Technology (1985).
- [19] Ramos J. A. and Verriest E. I., A Unifying Tool for Comparing Stochastic Realization Algorithms and Model Reduction Techniques, Proc. 1984 ACC.
- [20] Robert P. and Escoufier Y., A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient, Appl. Statist. 25, no. 3 (1976) 257-265.
- [21] Rao C. R., The Use and Interpretation of Principal Component Analysis in Applied Research, Sankhya A. 26 (1964) 329-358.
- [22] Varadarajan, V.S., Geometry of Quantum Theory (Van Nostrand, 1968).
- [23] Verriest, E. I. and Kailath, T., On Generalized Balanced Realizations, IEEE Trans. Automatic Control, AC-28, no. 8 (1983) 833-844.

APPENDIX T

OPTIMALITY PROPERTIES OF BALANCED REALIZATIONS: MINIMUM

SENSITIVITY

W. Steven Gray and Erik I. Verriest

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

Abstract. It has been well established that the so-called balanced state space realizations of linear time-invariant systems have certain desirable computational properties. But the complete relationship between balanced realizations and the general parameter sensitivity problem is not yet fully appreciated.

In this paper we present a geometric approach to the robust design problem and its solution using a specific optimality criterion. In particular, for discrete-time linear systems, a minimum sensitivity parameterization of a linear system with a given input-output relationship will be linked to the balanced realization.

Notation

(The following is a summary of the notation used in this paper. All nonstandard terminology will be defined in the body of the paper.)

R	set of real numbers
N	set of natural numbers
$\{...\}$	set of parameters
R^{ixj}	set of ixj real matrices
$\lambda_i(T)$	i th eigenvalue of matrix T
n	system order
m, p	number of system inputs, outputs
r	minimum of $(n+1)m$ and $(n+1)p$
s	equivalent to $(n+1)nm$
t	equivalent to $(n+1)np$
q	equivalent to $s+t$
θ	q -dimensional parameter space
θ	point in θ , i.e., a parameter set
θ^*	extremal sensitivity point in θ
$f(\cdot)$	observable functional
$f_\theta(\cdot)$	gradient vector of functional f
$f_{\theta\theta}(\cdot)$	Hessian matrix of functional f
$f_{\theta\theta\theta}(\cdot)$	third order tensor of functional f with respect to θ
k	observable value
$M_k(f)$	manifold in θ induced by functional f and observable value k
$\sim$	equivalence relation on θ
$L(\cdot)$	sensitivity performance index
λ	Lagrange multiplier constant
$H_k(\cdot)$	sensitivity Hamiltonian with respect to k
(A, B, C)	state space realization triple
R, O	reachability, observability matrices
O_i	i th column of matrix O
$O_{i,j}$	i, j th element of matrix O
H	system Hankel matrix
I_n	$n \times n$ identity matrix
I_{nm}	$n \times m$ identity matrix
σ_j, u_j	j th singular value
$\delta_{i,j}$	Kronecker delta function
$\otimes$	Kronecker product
$\text{vec}(\cdot)$	column stacking operator
$\text{Tr}(\cdot)$	trace operator
$\ \cdot\ _F$	Frobenius norm

1. The Problem Definition and Historical Content

This paper deals with a new geometric approach to the robust design problem. Classically, the sensitivity properties of a given realization have been investigated either via a "sensitivity system," which is usually prohibitively large [12], or alternatively, via the operator form [11]. Muller and Weber determine the control and observation sensitivity in [10] and maximize scalar measures for the "quality" with respect to certain structural parameters. The question of robustness, with respect to variations of certain structural parameters, is closely related to this problem and treated by Ackermann in [1]. Finally, sensitivity analysis from a geometric point of view was recently introduced by Delchamps [4] and applied to compensation and feedback. Our approach will be geometric in nature as well, but strictly from a realization perspective.

Consider a linear time-invariant system with m inputs and p outputs. For our applications, this may be a model for a real system one wants to simulate, or the implementation of a digital or analog filter, or an observer-controller for implementing an optimal regulator in some given plant. In all these applications, only the relationship between the input and output of the implemented device is important. It is now well known that many equivalent state space realizations exist for one and the same input-output behavior. Usually the so-called "canonical forms" are implemented because they minimize the number of parameters that are required and allow for a pipelined realization of the devices (e.g., the "direct forms" in digital signal processing). A minimal number of parameters corresponds to minimal complexity, a quality that may be important if the operation count is significant. However, a minimal set of parameters has no redundancy, and therefore, one may expect high sensitivity with respect to these parameters.

This paper investigates how the nonuniqueness of the state space realizations can be utilized to determine optimal parameterizations under various measures of "optimality." In particular, we shall look at realizations which exhibit a minimal sensitivity with respect to variations in parameters derived from the reachability and observability matrices.

In general, an n -dimensional linear system with m inputs and p outputs has a full parameterization consisting of $n(n+m+p)$ parameters [6]. Furthermore, it has been demonstrated that such a system can be minimally parameterized by no more than $n(m+p)$ parameters [2]. We shall find it mathematically convenient, however, to work with a larger, albeit redundant, parameter set. Specifically, we will consider a $q = (n+1)n(m+p)$ -dimensional affine space such that each point in the space represents a particular reachability and observability matrix for a particular linear system. The parameter space is not required to have a vector space structure. This is because additions and scalar multiples of such matrices have no meaningful interpretation. The space is given the structure of a Riemannian manifold by introducing an Euclidean metric in the tangent space at each point.

Clearly, in this space certain connected subsets will exist such that all points on such a subset correspond to a system with the same input-output behavior. In fact, we can resolve the whole space into disjoint sets corresponding to different input-output behavior. For a particular parameterization, the proximity of neighboring sets or "leaves" will be an indication of the robustness and sensitivity of the parameterization with respect to perturbations of the individual parameters. Hence, optimal parameterizations are those for which the "inter-leaf" distances are maximal.

2. Geometric Approach to the Robust Design Problem

Suppose we have an affine space θ and some smooth functional $f: \theta \rightarrow \mathbb{R}$. Then we can define $M_k(f) = \{\theta \in \theta: f(\theta) = k\}$ as the "level surfaces" induced by f and the scalar k on the space θ . Essentially, f generates equivalence classes on θ with $\theta^i \sim \theta^j$ if and only if $f(\theta^i) = f(\theta^j)$ for $\theta^i, \theta^j \in \theta$. This notion motivates the following definition.

Definition 1: The functional f will be called an observable functional over the parameter space θ . The scalar $f(\theta)$ at a particular $\theta \in \theta$ will be referred to as the observable value at θ .

Using the construction above, we see that parameter sets which give the same observable value, k , are essentially indistinguishable with respect to the given observable f . In the context of the study of parameter sensitivity, an important question herein is to determine which parameter sets, if any, in a given equivalence class, M_k , have the least or greatest propensity to change the observable value when the individual parameters in the sets are perturbed. This problem can obviously be addressed by examining the gradient of f over the manifold M_k .

Suppose we are given a parameter set $\theta_0 \in M_k$ for some fixed k . It is clear that we can perturb θ_0 in an infinite number of directions and, in general, the observable value will be perturbed as well. The direction with the greatest influence on k , i.e., the greatest directional derivative in magnitude, is in the direction of the vector $f_\theta(\theta_0)$, where

$$f_\theta(\theta_0) = \text{gradient of } f \text{ evaluated at } \theta_0.$$

Alternatively, we can view the space θ as being composed of level surfaces or "leaves," where a given leaf corresponds to a specific equivalence class M_k . The relative spacing between such leaves is measured by moving normal to the surface of one leaf until another is reached. This normal direction at a point on a surface is given by the gradient at that point, and the normal distance in this case can be best interpreted as the magnitude of the gradient. Thus, we immediately get the definition below.

Definition 2: A parameter set $\theta^* \in M_k$ is an extremal sensitivity point in M_k if and only if θ^* minimizes or maximizes $L(\theta)$ over M_k , where

$$L(\theta) = \frac{1}{2} \|f_\theta(\theta)\|^2. \quad (1)$$

Note, the extra square function and factor of $\frac{1}{2}$ does not change our conclusion above and were added only as an algebraic convenience.

Using methods from the calculus of variations, we can further characterize and determine the extremal sensitivity points of M_k . Specifically, let

$$H_k(\theta) = L(\theta) - \lambda(f(\theta) - k), \quad (2)$$

where λ is a Lagrange multiplier constant. Then we have the following definition.

Definition 3: A point θ on M_k is said to be an extremal sensitivity point if

$$dH_k(\theta)/d\theta = (f_{\theta\theta}(\theta) - \lambda I)f_\theta(\theta) = 0. \quad (3)$$

In other words, the extremal sensitivity points have gradient vectors which are eigenvectors of their Hessian matrix $f_{\theta\theta}$. We can ascertain whether a particular extremal sensitivity point has maximum sensitivity or minimum sensitivity by examining the definiteness of the Hessian matrix.

We now apply these general results in the context of discrete-time linear system theory and show how they can be used to solve the robust design problem.

3. An Application for Robust Design of Discrete-Time Systems

The parameter sensitivity of state space realizations of linear time-invariant systems has been studied by many investigators during the past decade [7,9-12]. Much of the research has been motivated by the desire of system designers to take into account the effects of uncertainty inherent in all practical systems. Frequent sources of uncertainty include imprecise knowledge of the "given" plant parameters, external disturbances to the plant, and roundoff and quantization error in the finite precision components of the control system.

It has been well established that balanced (in the sense of Moore [8,13]) state space realizations of linear time-invariant systems have certain desirable computational properties [7,9]. But the complete relationship between balanced realizations and the general parameter sensitivity problem is not yet fully understood. What follows is an attempt to use the results of the previous section to demonstrate further linkage between the two concepts.

Consider a system described by the state space model (A, B, C) where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$. At this point, we need not assume that the triple describes specifically a continuous or discrete-time system. It is well known that such a linear system can also be uniquely specified by an $(n+1)p \times (n+1)m$ matrix of Markov parameters known as the Hankel matrix, say H . Furthermore, the Rankel matrix can be shown to always permit the factorization

$$H = OR \quad (4)$$

where

$$O = [C^T \ A^T C^T \ (A^T)^2 C^T \ \dots \ (A^T)^n C^T]^T,$$

$$R = [B \ AB \ A^2 B \ \dots \ A^n B].$$

Now assume that we wish to realize a linear system specified uniquely by a given Hankel matrix, H . The parameterization we shall consider is a parameter set consisting of the $(n+1)n(m+p)$ components in the matrices O and R .

Definition 4: Let τ be a class of $(n+1)m \times (n+1)p$ matrices with the defining property that $A \in \tau$ if and only if A has the singular value decomposition

$$\Lambda = \sum_{i=1}^r v_i u_i^T, \quad (5)$$

where $r = \min((n+1)m, (n+1)p)$ and v_i and u_i are the appropriately sized column vectors such that $v_i^T v_j = u_i^T u_j = \delta_{i,j}$ [5].

Note that if $\Lambda \in \mathcal{A}$, then by definition it follows that

$$\Lambda \Lambda^T = I_{(n+1)m} \quad \text{when } p > m, \quad (6a)$$

$$\Lambda^T \Lambda = I_{(n+1)p} \quad \text{when } m > p. \quad (6b)$$

Clearly, when $m=p$, then $\mathcal{A}$ is simply the class of $(n+1)m \times (n+1)m$ orthogonal matrices.

Lemma 1: Let Z be an $(n+1)p \times (n+1)m$ matrix such that $\text{Tr}(\Lambda Z) = 0$ for all $\Lambda \in \mathcal{A}$. Then it follows that $Z = 0$.

Proof: Let Z have the singular value decomposition

$$Z = \sum_{j=1}^s \sigma_j u_j v_j^T. \quad (7)$$

Decomposing Λ , then as in (5) we have that

$$\begin{aligned} \text{Tr}(\Lambda Z) &= \text{Tr}\left(\sum_{i=1}^r v_i u_i^T \cdot \sum_{j=1}^s \sigma_j u_j v_j^T\right) \\ &= \sum_{i=1}^r \sigma_j \sum_{j=1}^s \text{Tr}(v_i u_i^T u_j v_j^T) \\ &= \sum_{i=1}^r \sigma_i = 0. \end{aligned}$$

But by the definition of the singular value decomposition $\sigma_i > 0$ for all $i \in N$ so $Z = 0$. ■

Lemma 1 provides us with an observable functional that can be used successfully in the context of the previous section. Specifically, let

$$f(\theta) = \text{Tr} \Lambda (E - O\theta), \quad \Lambda \in \mathcal{A} \quad (8)$$

$$\theta = \begin{bmatrix} \text{vec}(R^T) \\ \text{vec}(O) \end{bmatrix}, \quad (9)$$

where $\text{vec}(\cdot)$ is the column stacking operator from matrix calculus [3]. Clearly then we shall be interested in the equivalence class $M_0(f)$ where by Lemma 1 it follows that $E = O\theta$.

Theorem 1: The extremal sensitivity points of $M_0(f)$ have the property that

$$RR^T = O^T \Lambda^T \Lambda O \quad \text{when } p > m,$$

$$R \Lambda \Lambda^T R^T = O^T O \quad \text{when } m > p.$$

Proof: There are several ways to prove this result. The following method takes advantage of some matrix calculus "shorthand" notation to do the matrix manipulations [3]. From equation (8), we have

$$f(\theta) = \text{Tr} \Lambda (E - O\theta), \quad \Lambda \in \mathcal{A} \quad (10a)$$

$$= \text{Tr} \Lambda E - \text{Tr} R \Lambda O$$

$$= \text{Tr} \Lambda E - (\text{vec}(R^T))^T \cdot (I_n \otimes \Lambda) \cdot \text{vec}(O)$$

(see T3.8, [3])

$$f_\theta(\theta) = \begin{bmatrix} (I_n \otimes \Lambda) \text{vec}(O) \\ (I_n \otimes \Lambda)^T \text{vec}(R^T) \end{bmatrix} \quad (10b)$$

$$f_{\theta\theta}(\theta) = \begin{bmatrix} 0 & (I_n \otimes \Lambda) \\ (I_n \otimes \Lambda)^T & 0 \end{bmatrix}. \quad (10c)$$

The optimality condition is then

$$(f_{\theta\theta} - \lambda I_Q) f_\theta = 0, \quad (11)$$

Making the appropriate substitutions, it follows that

$$\begin{bmatrix} -\lambda I_s & (I_n \otimes \Lambda) \\ (I_n \otimes \Lambda)^T & -\lambda I_t \end{bmatrix} \cdot \begin{bmatrix} (I_n \otimes \Lambda) \text{vec}(O) \\ (I_n \otimes \Lambda)^T \text{vec}(R^T) \end{bmatrix} = 0, \quad (12)$$

where $s = (n+1)nm$ and $t = (n+1)np$. Equation (12) gives

$$\pm (I_n \otimes \Lambda) \text{vec}(O) = \text{vec}(R^T) \quad \text{when } p > m, \quad (13a)$$

$$\text{vec}(O) = \pm (I_n \otimes \Lambda)^T \text{vec}(R^T) \quad \text{when } m > p. \quad (13b)$$

We have used in (13a,b) the requirement that $\lambda = \pm 1$ for the system given by (12) to be consistent.

Now expand (13a):

$$\begin{aligned} O &= [O_1 \ O_2 \ \dots \ O_n]; \quad R^T = [R_1^T \ R_2^T \ \dots \ R_n^T] \\ O_i &= \text{columns of } O; \quad R_i^T = \text{columns of } R^T \end{aligned}$$

$$\pm \begin{bmatrix} \Lambda & & & \\ & \Lambda & & \\ & & \ddots & \\ & & & \Lambda \end{bmatrix} \begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_n \end{bmatrix} = \begin{bmatrix} R_1^T \\ R_2^T \\ \vdots \\ R_n^T \end{bmatrix}$$

$$\pm \Lambda O_i = R_i^T; \quad i = 1, 2, \dots, n$$

$$\pm \Lambda [O_1 \ O_2 \ \dots \ O_n] = [R_1^T \ R_2^T \ \dots \ R_n^T]$$

$$\pm \Lambda O = R^T$$

$$RR^T = O^T \Lambda^T \Lambda O.$$

Likewise, from (13b) we get

$$R \Lambda \Lambda^T R^T = O^T O.$$

We now restrict our attention to the case where $m = p$ such that Λ and, consequently, $f_{\theta\theta}$ are simply orthogonal matrices.

Definition 5: Realization (A, B, C) with $m = p$ is said to be an essentially balanced realization if $RR^T = O^T O$.

The definition above is motivated by the fact that in the discrete-time case RR^T and $O^T O$ are respectively, the controllability and observability grammians at time

$t = n$. Thus, an essentially balanced realization can be converted to a true finite-interval discrete-time balanced realization (see [14]) by simply an orthogonal state space transformation T such that

$$T(RR^T)T^T = T(O^TO)T^T = \text{diag}(u_1, u_2, \dots, u_n) = I.$$

In this case, T can be interpreted as a series of rotations of the state space's coordinate axes. From this observation it follows then that only the scaling of the state variables has an effect on the sensitivity of the parameters $[R_{i,j}]$ and $[O_{i,j}]$. Another important property of essentially balanced realizations is illustrated in the following lemma.

Lemma 2: An essentially balanced realization has the minimum sensitivity property.

Proof: As alluded to in the previous section, we can test for the sense of the optimality by examining the definiteness of the Hessian matrix

$$H_{\theta\theta} = (f_{\theta\theta})f_{\theta} + (f_{\theta\theta} - \lambda I_q)f_{\theta\theta}. \quad (14)$$

For our particular choice of observable, f , we have $f_{\theta\theta} = 0$ since it is bilinear in $R_{i,j}$ and $O_{i,j}$. Since $f_{\theta\theta}$ is an orthogonal symmetric matrix, it follows that

$$\lambda_i [f_{\theta\theta}]_{\theta=\text{extremal}} = \pm 1, \quad i = 1, 2, \dots, q,$$

i.e., the matrix $f_{\theta\theta}$ is similar to a signature matrix. Thus there exists a similarity transform T such that

$$Tf_{\theta\theta}T^{-1} = \text{diag}(\pm 1, \pm 1, \pm 1, \dots, \pm 1)$$

$$T(f_{\theta\theta} - \lambda I_q)T^{-1} = \text{diag}(\pm 1 - \lambda, \pm 1 - \lambda, \pm 1 - \lambda, \dots, \pm 1 - \lambda).$$

Hence,

$$T(f_{\theta\theta} - \lambda I_q)f_{\theta\theta}T^{-1} = \text{diag}(1 \pm \lambda, 1 \pm \lambda, 1 \pm \lambda, \dots, 1 \pm \lambda).$$

For $\lambda = 1$ and $\lambda = -1$, we have

$$T(f_{\theta\theta} - \lambda I_q)f_{\theta\theta}T^{-1} = \text{diag}\left(\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}, \begin{smallmatrix} 2 \\ 0 \end{smallmatrix}, \begin{smallmatrix} 2 \\ 0 \end{smallmatrix}, \dots, \begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right), \quad (15)$$

and thus $H_{\theta\theta} > 0$.

The fact that the Hessian matrix is semi-definite leaves open the possibility that there may be some coordinate directions one can move in the parameter space θ that have no influence on the observable value. Hence, the observable would be completely insensitive to these parameters.

There are several interesting observations to be made with respect to the minimum sensitivity criterion. First, note that we worked specifically with a Hankel matrix $H \in \mathbb{R}^{(n+1)p \times (n+1)m}$. It is well known that the Hankel matrix must be at least this large in order to specify the system uniquely. However, there is no reason why the above derivations could not have been carried out using a larger system Hankel matrix, i.e., using Markov parameters of higher degree. In fact, if it is assumed that we are working with a stable system, then we could have used the doubly infinite Hankel matrix in the derivations. In this later case, the criterion given by Theorem 1 is satisfied by requiring that the system realization be an infinite-interval discrete-time balanced realization. Second, note that the minimum sensitivity criterion did not use implicitly the fact that the Hankel matrix given corresponded to that of a discrete-time system. This assumption

simply allowed us to put the results immediately in a balanced realization framework. The continuous-time problem is slightly more difficult to solve and will be presented in a later publication [15].

Finally, it is instructive to display the performance index explicitly for the observable functional given in equation (8). Substituting equation (10b) into (1), it follows that

$$L(\theta) = \frac{1}{2} \left\| \frac{\text{vec}(O)}{\text{vec}(R^T)} \right\|^2. \quad (16)$$

Hence, the effect of minimizing $L(\theta)$ while constraining $H = OR$ is to make the components of O and R roughly the same order of magnitude. Such a situation would be desirable if these parameters were to be quantized for fixed point data registers.

4. Computational (Algorithmic) Aspects

A few comments are in order concerning how an essentially balanced realization might be computed given a specific Hankel matrix. Clearly, there are matrix algebraic methods possible similar in flavor to those used to compute balanced realizations. We shall concentrate in this case, however, on a numerical optimization approach that follows very naturally from the theoretical structure put forth up to this point.

From our discussion, it follows that an essentially balanced realization will minimize the performance index given in equation (16), subject to the constraint that $H = OR$. Lemma 1 gave us a convenient way to adjoin this constraint while keeping the Hamiltonian simple and differentiable in the components of O and R . The price paid, however, was the introduction of an arbitrary orthogonal matrix λ . Clearly, such a formulation does not lend itself easily to the use of numerical algorithms. To remedy this problem, form the Hamiltonian instead as

$$H_0(\theta) = H_0(O, R) = \frac{1}{2} \left\| \frac{\text{vec}(O)}{\text{vec}(R^T)} \right\|^2 - \lambda \|H - OR\|_F^2, \quad (17)$$

where $\|\cdot\|_F$ is the Frobenius norm. Now the optimization can be carried out explicitly on H_0 with respect to O and R and the scalar λ . It should be noted, however, that from an analytic viewpoint equation (16) is often more difficult to work with. Finding an essentially balanced realization is now reduced to solving an optimization problem where H_0 is to be minimized with respect to the components of O and R . One approach to solving such a problem is to employ a relaxation-type algorithm. Specifically, one can perform a sequence of one-dimensional gradient based optimizations until an essentially balanced pair (O, R) is computed to desired precision. The B and C matrices of the realization can then be immediately identified from the first block column of R and the first block row of O , respectively. The A matrix follows, for example, from

$$A = \hat{R} \hat{R}^+ \quad (18)$$

where

$$R = \begin{bmatrix} R \\ \vdots \\ R_{n+1} \end{bmatrix}, \quad (19)$$

$\hat{R}^+$ is the pseudo-inverse of $\hat{R}$ and $\hat{R}$ is found by shifting the right-most block column of R into $\hat{R}$ [7].

An obvious limitation to this method is that H_0 must be optimized with respect to a large number of parameters. In fact, for long or infinite time intervals, such an approach will not be practical.

5. Conclusions and Future Research

Given a square system Hankel matrix, it was shown that a realization, which satisfies the discrete-time balanced realization criterion to within an orthogonal state space transformation, has minimum sensitivity with respect to perturbations in the components of the reachability and observability matrices. This was observed to be true regardless of the size of the Hankel matrix used provided it was large enough to specify the system uniquely. It was also suggested that an optimization-type algorithm could be used to determine explicitly such a state space realization.

A topic for future research is to determine whether the nonuniqueness of the essentially balanced realization can be exploited to find subsets of this class which have other desirable properties. For example, in the context of finite wordlength effects, there is the deterministic effect due to coefficient truncation and the stochastically modeled effect due to computation roundoff. Here we have shown that essentially balanced realizations, i.e., a realization where the controllability and observability grammians are equal, have a certain minimal sensitivity property, with respect to parametric perturbation. On the other hand, Mullis and Roberts have demonstrated that, with respect to roundoff noise, an optimal wordlength filter follows from a state space realization where such grammians are simultaneously diagonal (though not necessarily equal) [9]. Hence, both optimality properties are possessed by the usual balanced realization.

Another topic for future research is to relate the results stated herein concerning the parametric sensitivity of the matrix factorization $\Sigma = QR$ to the parametric sensitivity of the state space triple (A, B, C) .

Acknowledgement

This research is supported by the U.S. Air Force under Contract F08635-84-C-0273.

References

- [1] J. Ackermann, "Parameter Space Desig. of Robust Control Systems," *IEEE Trans. Auto. Control*, vol. AC-25, Dec. 1980, pp. 1058-1072.
- [2] S.P. Bingulac, "On the Minimal Number of Parameters in Linear Multivariable Systems," *IEEE Trans. Auto. Control*, vol. AC-21, Aug. 1976, pp. 604-605.
- [3] J.W. Brewer, "Kronecker Products and Matrix Calculus in System Theory," *IEEE Trans. Circuits Syst.*, vol. CAS-25, Sept. 1978, pp. 772-781.
- [4] D.P. DeChamps, "New Geometric Approaches to Parameter Sensitivity in Feedback Systems," Modelling, Identification and Robust Control, C.I. Byrne and A. Lindquist, eds., North-Holland, 1986, pp. 445-456.
- [5] G.E. Golub and C.F. Van Loan, Matrix Computations, John Hopkins Univ. Press, 1983.
- [6] M. Hazewinkel, "Moduli and Canonical Forms for Linear Dynamical Systems II: The Topological Case," *Math. Syst. Thy.*, vol. 10, 1977, pp. 363-385.
- [7] S. Kung, "A New Identification and Model Reduction Algorithm via Singular Value Decomposition," *Proc. 12th Asilomar Conf. on Circuit, Systems and Computers*, Pacific Grove, CA, Nov. 1978.
- [8] B.C. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Trans. Auto. Control*, vol. AC-26, Feb. 1981, pp. 17-32.
- [9] C.T. Mullis and R.A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, Sept. 1976, pp. 551-562.
- [10] P.C. Muller and H.I. Weber, "Analysis and Optimization of Certain Qualities of Controllability and Observability for Linear Dynamical Systems," *Automatica*, vol. 8, 1972, pp. 237-246.
- [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.*, vol. 301, Jan.-Feb. 1976, pp. 123-141.
- [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] E.I. Verriest, "On Generalized Balanced Realizations," *IEEE Trans. Auto. Control*, vol. AC-28, June 1983, pp. 833-844.
- [15] E.I. Verriest and W.S. Gray, "Robust Design Problems: A Geometric Approach," *Proc. 1987 MTNS (to appear)*.

APPENDIX U

ROBUST DESIGN PROBLEMS: A GEOMETRIC APPROACH

ROBUST DESIGN PROBLEMS:
A GEOMETRIC APPROACH

Erik I. Verriest
and
W. Steven Gray

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

(404) 894-2949

Abstract

There exists a certain correspondence between the problems of observability and identification. A less familiar correspondence is the one relating controllability to a "dual" of the identification problem: the "DESIGN"-problem. This amounts to the choice of a realization or approximation of a desired system response, using parameters that can only be approximately adjusted, e.g. due to quantization. A particular application is in the design of digital filters, simulators and controllers, which minimize the effects of component tolerances in analog systems or finite wordlength effects in the digital discrete case.

A geometric approach to the design problem is presented, and its solution given under a useful criterium for optimality. For linear time invariant systems, the minimum sensitivity realizations of a desired Hankel matrix are linked to the Balanced Realizations.

Acknowledgement

This research is supported by the U.S. Air Force under Contract No. FO8635-84-C-0273

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, either via a "sensitivity system", which is usually prohibitively large [TV,F], or alternatively, via the operator form [RMAD]. Muller and Weber determine the control and observation sensitivity in [MW], and maximize scalar measures for the "duality" with respect to certain structural parameters. The questions of robustness with respect to variations of certain structural parameters is closely related to this problem, and treated by Ackermann in [A]. Finally, sensitivity analysis from a geometric point of view was recently introduced by Delchamps [D], 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. For our applications, 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 implementing 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 that is required, and allow for a pipelined realization of the devices, e.g. the "Direct Forms" in digital signal processing. A minimal number of parameters 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 one may expect high sensitivity with respect to these parameters.

This paper investigates how the nonuniqueness of the state space

realizations can be utilized to determine optimal parametrizations under various measures of "optimality". In particular, two issues seem to be important for the practical implementation of a given transfermatrix: sensitivity and clustering. The minimal sensitivity requirement guarantees that the actual realized transfermatrix is "close" to the nominal transfer matrix. Clustering deals with the desired parameter values. If for instance a fixed point implementation is used, then it is desirable to have all parameter values in some range or ranges. It relates to the problem of realizing an approximation to a given system with parameters chosen from a finite set with fixed values. This paper focusses on the first problem.

Our approach to the problem is geometric, as in [D]. A full parametrization of the system has $n^2+np+nm$ parameters. A minimal parametrization on the other hand requires $n(1+m)$ or $n(1+p)$ parameters if $p=1$ or $m=1$ [H], or if p and m are both larger than 1, somewhere between $n(m+1)+p(p+1)$ and $n(m+p)$ parameters [B].

Because addition and scalar multiplication of systems have no meaningful natural interpretations, the parameter space is simply assumed to have the structure of an affine space of dimension $n(n+m+p)$. Each point in this space represents a particular realization of an m input, p output linear time invariant system of order n (or less). 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.

Clearly, in this space, certain connected subsets will exist such that all points on such a subset correspond to a system with one and the same input output behavior. In fact we can resolve (i.e. partition into

equivalence classes) the whole space 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. Hence optimal realizations are those at which the "inner-sheet" distances are maximal. 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

A rather unusual viewpoint due to Root [R], considers the phenomenon "linear system" as a mapping σ from a suitable subset of the cartesian product of input functions (U) and realizations (Σ) to the set of output functions (Y). The restriction to a certain subset (we will not go into the details of this) is necessary for the convergence issues. For continuous linear time-invariant systems, the mapping stands for the convolution operator

$$\sigma : U \times \Sigma \longrightarrow Y : (u(.), S) \longrightarrow y(.)$$

$$y(t) = \int_{-\infty}^t C e^{A(t-r)} B u(r) dr$$

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

$$\sigma_S : U \times \{S\} \longrightarrow Y : u(.) \longrightarrow y(.)$$

On the other hand, for a fixed input $u(.)$, the marginal maps

$$\sigma_u : \{u\} \times \Sigma \longrightarrow Y : S \longrightarrow y(.)$$

associate with each realization S e.g. the impulse response $h(t)$ if $u(t) = \delta(t)$, or the transferfunction $H(p)$ characterizing the steady state response to a sinusoid $u(t) = e^{pt}$ of complex radial frequency p .

The control and deconvolution problems are inverse problems for the map σ_u 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 is necessarily corrupted with uncertainties due to the finite observation time and finite memory effects. It may even be impossible to isolate the phenomenon of interest from the rest of the universe. Similarly, uncertainties interfere with the design problem: the parameter values necessarily must have a finite precision. In order to find "uniquely" an "optimal" solution to these problems, one introduces a suitable distance measure or norm in the domain and range spaces $[W]$.

3. MAIN RESULTS

A summary of some known results on the geometry of systems and their realizations is first given. It is established that the subsets of the realization space of system realizations which exhibit identical input/output behavior form smooth manifolds. However the totality of all

such "sheets" does not have the structure of a foliation since not all these subsets are manifolds of the same dimension. A restriction on the set of systems is required in order to avoid such degeneracies. 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 dimensions $n \times n$, $n \times m$, $p \times n$. Only realizations over the real field R will be dealt with here. Since there is little significance to the addition and/or scalar multiplication of realizations, this space is not endowed with a vector space structure, but rather that of an affine space with vector space $R^{n(m+n+p)}$. Hence at each point S , there is an attached vectorspace $T_S L$ (the tangentspace 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

$$(A,B,C) \longrightarrow (A,B,C)^T = (TAT^{-1}, TB, CT^{-1})$$

which of course corresponds to a change of base in the state space $z = Tx$.

The quotient topology is non Hausdorff in general. The restriction to the completely reachable (or dually, the completely observable) pairs eliminates many problems, in particular, the action of $GL_n(R)$ is free (as a consequence of reachability/observability) and the quotient space (set of orbits)

$M_{m,n,p}^{cr} = L_{m,n,p}^{cr} / GL_n(R)$ is a smooth (real) analytic (thus certainly C^∞) differentiable manifold (hence Hausdorff) of dimension $n(m+p)$ [H2]. The set of equivalence classes of minimal realizations $M_{m,n,p}^{co,cr}$ are analytic open sub-manifolds, [H2]. In the system identification problem, this space, called parameter space, plays a crucial role. Its properties have been well studied. (e.g. in relation to the (non)existence of continuous canonical forms [H2], and degeneration phenomena [H3]. The best one can do based on input-output data alone is to identify the orbit of a minimal realization S ,

i.e. parametrize the input-output operator in some (minimal) way, by choosing a point in $M_{m,n,p}^{co,cr}$. An obvious question is: "Does a sequence (obtained as more data comes in) of points in $M_{m,n,p}^{co,cr}$ necessarily converge to a point in $M_{m,n,p}^{co,cr}$?" The answer is no: degeneration phenomena occur, and so-called generalized systems appear [H2,H3]. It has been established that there is a continuous canonical form on $L_{m,n,p}^{co,cr}$ if and only if $\min(p,m) = 1$.

Since the isotropy subgroup is trivial for all reachable or observable realizations, its dimension is constant on $L_{m,n,p}^{co,cr}$, and hence the orbits of $GL_n(R)$ form a foliation F of $L_{m,n,p}^{co,cr}$ of dimension $n(m+p)$ [L]. The field of tangent spaces to the leaves form an $n(m+p)$ dimensional subbundle $\tau(F)$ of the bundle, called the tangent bundle to F . The quotient bundle $\nu(F) = TL/\tau(F)$ is called the normal bundle to F .

Our interest is not in the universal parametrization, but in the orbits under the action of $GL_n(R)$ itself. These orbits are open, and the boundary points of reachable realizations are non reachable realizations. The explicit form of the closure of the orbits was addressed in [KM]. We shall endow the tangentbundle $TL_{m,n,p}^{co,cr}$ with a positive definite metric.

$$\langle \dots \rangle_S: T_S L \times T_S L \longrightarrow R \quad \text{for all } S \text{ in } L_{m,n,p}^{co,cr}$$

This metric is associated with the tolerance in the components of the realization and is different from the (observability, reachability and riccati metrics) on the associated vector bundle (the state bundle) of the principal fibre bundle $\pi: L \longrightarrow M$ with structural group $GL_n(R)$, discussed by Delchamps [D]. A uniform metric, $P(S) = I_{n(m+n+p)}$ would for instance be useful in the design of fixed point computer realizations of a given $m \times p$

regulator of McMillan degree n . Its induced norm is the Frobenius norm of the realization matrix $\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$.

3.2 The Robust Design Problem: A Geometric Approach.

Before proceeding with our system design, we shall prove a general result on sensitivity:

Definition: Let Θ be an N - dimensional open subset of an affine space A^N of design parameters (configurations). By an Observable, we shall mean any smooth function $f: \Theta \rightarrow \mathbb{R}$ which has no critical points (i.e. the gradient is never zero).

Remarks : i) The reason for considering open subsets is that typically for our applications, but not exclusively, this situation occurs if the inverse image of a finite set of points by a continuous map from this set to the reals is cut out.

ii) The significance of an observable is that any two configurations θ_1 and θ_2 in the parameterspace are indiscernable by observation of f if $f(\theta_1) = f(\theta_2)$. This allows us to regard two parametrizations yielding the same observable(s) as being the same (or equivalent) for some purpose. In a systems context, an observable is for instance the value of the transfer function (scalar case) at a particular frequency, or the impulse response evaluated at a specific instant. They are also referred to as "system functions" [F].

iii) The gradient of a function depends on the metric of the space. The vanishing of a gradient at a point is independent of the chosen metric.

Every such map induces a partition of Θ into equivalence classes, in fact, these equivalence classes form what is commonly known as a foliation.

In this case, the submanifolds are the level surfaces of f , and have dimension $N-1$. There exist 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. Of course, this notion needs to be made precise, since the leaves are densely stacked.

3.2.1 Riemannian Metrics

If Θ is paracompact, then a Riemannian structure G can be put on Θ (or, more exactly, on its tangentbundle). This means that for each $\theta \in \Theta$, a symmetric, positive definite bilinear form G_θ is defined on the vector space $T_\theta\Theta$, such that G defines a metric on $T\Theta$, i.e. is a smooth section of the vector bundle $T_\theta^*\Theta$. Let $\sharp: T^*\Theta \rightarrow T\Theta$ 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 $df^\sharp$ of $T\Theta$ (i.e. the vector field corresponding under the map $\sharp$ to the differential form df). In the local 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})

$$g_{ij} = G \left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \theta^j} \right)$$

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 tangentspace 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 .

The problem is to find the points on the leaves for which the effects of an infinitesimal perturbation are minimized. A worst case analysis leads to the minimization of the gradient norm $\|\nabla_G f\| = G(\nabla_G f, \nabla_G f)^{\frac{1}{2}}$, or equivalently, but mathematically more convenient, the map $\bar{h}$

$$\bar{h} = \frac{1}{2}\|\nabla_G f\|^2$$

This scalar field induces a vector field in the tangent space Δ_θ of the leaf. However, note that $d\bar{h}^* = dG(df^*, df^*)^*$ is in general not tangent to the leaf. Its projection on the tangentspace to the leaf at θ yields the tangent vector $dG(df^*, df^*)^* - \lambda df^*$ to the leaf through θ , for some $\lambda \in \mathbb{R}$. Equivalently, we could have worked directly with the Hamiltonian for the constrained problem, as the points are constrained by the leaves of the foliation ($f=\text{cst}$) of θ .

$$h = \frac{1}{2}\|\nabla_G f\|^2 - \lambda f$$

Either way, it leads in coordinate free form to

Theorem 2: 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 $\bar{h}$ 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

$$[\text{diag}(\theta)\text{diag}(f_{\theta}) + \text{diag}(\theta^2)f_{\theta\theta}] \text{diag}(\theta^2)f_{\theta} = \lambda \text{diag}(\theta^2)f_{\theta}$$

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 $\hat{H} = \text{diag}(\hat{\nabla}f) + \text{diag}(\theta)f_{\theta\theta}\text{diag}(\theta)$

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.

$$(\hat{H}(f) - \lambda I)df^* = 0$$

Example 1. Consider in R^2 the foliation $\theta_1\theta_2 = \text{constant}$. With the uniform metric, the extremal sensitivity points are on the diagonals $\theta_1 = \pm\theta_2$. With the relative metric, the generalized gradient is $\hat{\nabla}f = \theta_1\theta_2 (1,1)'$ and the generalized Hessian $\hat{H} = \theta_1\theta_2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

The problem is degenerate. Every point is an extremal sensitivity point.

Example 2. Consider in R^2 the leaf $f(\theta) = 1$ of an ellipsoidal foliation given by $f(\theta) = \theta_1^2/a^2 + \theta_2^2/b^2$. For the uniform metric, the gradient is $(2\theta_1/a^2, 2\theta_2/b^2)'$ and the Hessian $\text{diag}(2/a^2, 2/b^2)$. The eigenvectors of the

Hessian are (1,0) and (0,1) corresponding to the extremal sensitivity points $(\pm a, 0)$ and $(0, \pm b)$. If $|a| \leq |b|$ then the former is a maximal sensitivity, and the latter a minimal sensitivity point. With the relative metric, one finds the extremal sensitivity points at $\theta_1 = \pm a/\sqrt{2}$ and $\theta_2 = \pm b/\sqrt{2}$, i.e. the points where the diagonal of the enclosing rectangle, with sides $2a$ and $2b$, intersects the ellipse.

4 APPLICATION TO ROBUST REALIZATIONS

For the application to linear system realizations and design it was found fruitful to express the parametrization 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. This does not quite solve the optimal realization problem, but it provides a suboptimal solution, which is mathematically more tractable. The Hankel matrix defined as a map with domain $L_{m,n,p}$ plays the role of a multidimensional observable. The details for discrete time systems are given in [GV]. In this paper the continuous time systems design under the uniform metric is discussed, for square ($p=m$) systems only. The general case is more tedious and will be published elsewhere.

Definitions: Let $L_2^m[0, \infty)$ be the Hilbertspace of m -vectorfunctions with inner product $\langle x(\cdot), y(\cdot) \rangle = \int_0^\infty x(t)' y(t) dt$. The reachability operator $R: L_2^m[0, \infty) \rightarrow R^n$ for a realization (A, B, C) is defined by $Ru(t) = \int_0^\infty e^{At} B u(t) dt$. Its adjoint R^* is the operator $R^*: R^n \rightarrow L_2^m[0, \infty) : R^*x = B' e^{A' t} x$.

The observability operator is $Q: R^n \rightarrow L_2^p[0, \infty) : Qx = C e^{At} x$. Since R and Q have a finite dimensional range and domain respectively, they are compact operators [K, p.157]. Furthermore, their composition QR is also

compact [K, p. 158]. Finally, we introduce the Hankel operator H : $L_2^{\mathbb{M}}[0, \infty) \rightarrow L_2^{\mathbb{P}}[0, \infty) : Hu(t) = \int_0^{\infty} h(t+r)u(r)dr$ where $h(t) = Ce^{At}B$ is the impulse response of the realization (A, B, C) . It is readily verified that indeed $H = QR$. An operator $\Lambda: L_2^{\mathbb{M}}[0, \infty) \rightarrow L_2^{\mathbb{M}}[0, \infty)$ satisfying $\Lambda\Lambda^* = \Lambda^*\Lambda = \text{Id}$ (the Identity operator) is called isometric. We shall also assume that the set $\{e_i\}_{i=1}^{\infty}$ is the standard basis for $\mathbb{R}^n$ and that the functions $\{\psi_i\}_{i=1}^{\infty}$ form a complete orthonormal basis in $L_2^{\mathbb{M}}[0, \infty)$. Many notations are used in such a setting. We found it relatively easy to use the Dirac notation for the vectors and their duals (i.e. the bra and ket notation). In this form we have:

$$H = \sum_{uv} h_{uv} |\psi_u\rangle \langle \psi_v|$$

$$R = \sum_{ij} r_{ij} |e_i\rangle \langle \psi_j|$$

$$Q = \sum_{kl} o_{kl} |\psi_k\rangle \langle e_l|$$

The matrix representations $[h_{ij}]$, $[r_{ij}]$ and $[o_{ij}]$ will be respectively denoted by $\text{Mat}(H)$, $\text{Mat}(R)$ and $\text{Mat}(Q)$. By $\text{Vec}(M)$ 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 $E: L_2^{\mathbb{M}}[0, \infty) \rightarrow L_2^{\mathbb{M}}[0, \infty)$ be such that $\text{Tr } \Lambda E = 0$ for all isometric operators Λ , then $E = 0$.

proof: Suppose E has the singular value decomposition [K, p.261]

$$E = \sum \theta_i |u_i\rangle \langle v_i|$$

where $\{u_i\}$ and $\{v_i\}$ are orthonormal sets in $L_2^{\mathbb{M}}[0, \infty)$, then choosing Λ as $\sum |v_j\rangle \langle u_j|$ yields $\sum \theta_i = 0$. Since the singular values 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 matricielements of Q and R , so that the parameter vector is $\theta' = [\text{Vec}(\text{Mat}(R))', \text{Vec}(\text{Mat}(Q))']$.

Analogous to the discrete case [GV], we shall consider the observables:

$f_\Lambda(\theta) = \text{Tr} \Lambda(\underline{H} - \underline{O}R)$. Denote by $M_O(f_\Lambda)$ the leaf on which f_Λ is constant,

zero say, then we have the

Theorem 3: The extremal sensitivity points of $M_O(f_\Lambda)$ have the property that $\underline{R}\underline{R}^* = \underline{Q}^*\underline{Q}$.

proof: 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 [GV], where it was shown, based on the corollary to Theorem 2, that the extremal sensitivity points satisfy

$$\text{Mat}(\underline{R})\text{Mat}(\underline{R})' = \text{Mat}(\underline{Q})'\text{Mat}(\underline{Q})$$

Re-expressing $\text{Mat}(\underline{R})\text{Mat}(\underline{R})'$ and $\text{Mat}(\underline{Q})'\text{Mat}(\underline{Q})$ in the basis $(e_i)_{i=1}^n$ gives then the condition in terms of the original operators: $\underline{R}\underline{R}^* = \underline{Q}^*\underline{Q}$ ■

Corollary: 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.

proof: 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

$M_0(f_A)$ is the submanifold characterized by $H = QR$, i.e. the orbit of the system with Hankel operator H under the action of $GL_n(R)$.

Then, by the previous theorem, $RR^* = Q^*Q$ so that

$$\begin{aligned}\langle x, RR^*y \rangle &= \langle x, Q^*Qy \rangle & \forall x, y \in R^n \\ \langle R^*x, R^*y \rangle &= \langle Qx, Qy \rangle\end{aligned}$$

which in integral form is

$$x' \int_0^\infty e^{At} BB' e^{A't} dt y = x' \int_0^\infty e^{A't} C' C e^{At} dt y$$

By definition, this states the equality of the Reachability Gramian with the Observability Gramian. Realizations having this property are essentially balanced, hence their name, as an orthogonal similarity transformation will make them truly balanced (equal and diagonal gramians), in the sense of Moore [M]. The observables $f_A(\theta)$ are invariant with respect to such transformations. The second variation property is used to show that the extremal solutions obtained indeed 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. It follows thus that any essentially balanced realization is truly a minimum sensitivity realization! ■

As shown by the main theorem, it suffices to find an essentially balanced realization of the given system. The characterization as a factorization of the Hankel matrix is therefore independent of the size of the Hankel matrix considered, as long as it is large enough to specify the given input-output relation.

5 CONCLUSIONS

The optimal sensitivity properties for the continuous time realizations have been derived. By using expansions in a complete orthonormal basis in the function space L_2 , the problem was reduced to the discrete timeoptimality problem, solved in [GV]. In both cases the balanced realizations are therefore optimal. The balanced realizations have been widely used in model reduction methods, even though no clear optimality properties were known about the resulting reduced order models [M,G1].

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. Finally, the idea in the proof of the main sensitivity theorem leads to gradient type algorithms for the optimal sensitivity realizations. Some preliminary remarks regarding these appear in [GV].

REFERENCES

- [A] J. Ackerman, "Parameter Space Design of Robust Control Systems", IEEE Trans. Auto. Control, Vol. AC-25, No.6, 1980.
- [B] S.P. Bingulac, "On the Minimal Number of Parameters in Linear Multivariable Systems", IEEE Trans. Auto. Control, August 1976
- [D] 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.
- [F] P.M. Frank, "Introduction to System Sensitivity Theory", Academic Press 1978.
- [G] K. Glover, "All Optimal Hankel-Norm Approximations of Linear Multivariable Systems and their L^∞ Bounds" Int. J. Control, 1986.
- [GV] W.S. Gray and E.I. Verriest, "Optimality Properties of Balanced Realizations: Minimum Sensitivity", 1987 IEEE Conf. Dec. Control, Los Angeles, CA, December 1987.
- [H1] M. Hazewinkel, "Moduli and Canonical Forms for Linear Dynamical Systems II: The Topological Case", Math. Syst. Thy., Vol. 10, 1977.
- [H2] 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, 1977 125-193.

- [H3] 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, Volume 18, 1980, 157-189.
- [K] T. Kato, "Perturbation Theory for Linear Operators" Springer-Verlag, 1976.
- [KM] A.S. Khadr and C. Martin, "On the $\text{GL}_n(\mathbb{R})$ 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.
- [L] H.B. Lawson, Jr. "The Quantitative Theory of Foliations", Conf. Board of the Mathematical Sciences, 1977.
- [M] 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.
- [MW] P.C. Muller and H. I. Weber, "Analysis and Optimization of Certain Qualities of Controllability and Observability for Linear Dynamical Systems", Automatica, Vol. 8, 1972.
- [PMAD] 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., Vol. 301, Nos. 1 and 2, Januari-Februari 1976.
- [R] W.L. Root, "On the Identifiability of Abstract Systems", Proc. 1986 Conf. on Information Systems and Sciences.
- [IV] R. Tomovic and M. Vukobratovic, "General Sensitivity Theory", American Elsevier, 1972.
- [V] E.I. Verriest, "The Structure of Multivariable Balanced Realizations", Proc. 1983 Int'l. Symp. Circuits and Systems, Newport Beach, CA.
- [W] J.L. Willems, "Models for Dynamic Systems", submitted to Dynamics Reported, 1986.

APPENDIX V
ON SINGULARLY PERTURBED SWITCHED PARAMETER SYSTEMS

ON SINGULARLY PERTURBED SWITCHED PARAMETER SYSTEMS

M. V. Jose and A. H. Haddad

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

ABSTRACT

This paper considers a switched parameter stochastic linear system whose state equations depend on a finite state Markov process. The decomposition of the system and the process together into fast and slow components is investigated in the paper when both are singularly perturbed. The results can be shown to hold when the process is independent of the original system, is ergodic, and the matrices of the different models of the system commute.

INTRODUCTION

This paper considers the limiting behavior of singularly perturbed switched parameter linear stochastic models. Such models occur in many detection-estimation schemes [1] and in the study of multiplex control systems [2]. They have also been described more recently as hybrid systems [3]. The paper is concerned with the properties of such models when both the continuous and the underlying discrete-event processes are singularly perturbed. Stochastic linear singularly perturbed systems have already been considered, and their properties are well documented [4]. Similarly, [5] considers the aggregation of states for singularly perturbed discrete-event Markov processes. Here we study the combination of both types of behavior for singularly perturbed linear stochastic models which depend on a discrete-event Markov process, also singularly perturbed.

The system model is assumed to have the following set of state equations

$$\dot{x}_1 = A_1[u(t)] x_1 + A_{12}[u(t)] x_2 + B_1 w(t) \quad (1)$$

$$\mu \dot{x}_2 = A_{21}[u(t)] x_1 + A_2[u(t)] x_2 + B_2 w(t) \quad (2)$$

where $w(t)$ is a white Gaussian noise vector, and where $\mu > 0$ is a small parameter. Thus $x_1(t)$ represents the slow mode of the system, and $x_2(t)$ the fast mode. The process $u(t)$ is a discrete-event Markov chain with transition probability matrix $P(\tau)$, and is assumed to be independent of $x(t)$ and ergodic. The singularly perturbed case assumes that the $M \times M$ transition probability matrix $P(\tau)$ may be block partitioned into $P_{11}(\tau)$ (of dimension $M_1 \times M_1$), $P_{12}(\tau)$, $P_{21}(\tau)$, and $P_{22}(\tau/c)$ (of dimension $M_2 \times M_2$), where $c > 0$ is a small parameter. The parameter c represents the fact that some of the states of $u(t)$ are assumed to have fast

transition probabilities. Thus $u(t)$ may be described as having two sets of discrete states, S_1 containing M_1 states, and S_2 containing M_2 states, representing the slow and fast components, respectively. The methods developed in [5] may be used to aggregate the fast states into a single state so that as far as the slow time-scale is concerned the process may be approximated by an M_1+1 state slow Markov process.

The limiting behavior of the overall system depends on the relative size of μ and c as they both tend to zero. This reflects the fact that the fast time-scales of the discrete process and the continuous one may not be equally fast. While the general case is considered in the paper, this summary will only address the problem when μ and c have the same order of magnitude, and are assumed to be equal. Furthermore, for simplicity (and due to the possibility of transforming the original system into a decoupled one) we shall concentrate on the decoupled case, i.e., $A_{12} = 0$, and $A_{21} = 0$. The summary of the main results are given in the following sections.

FAST MODE SYSTEM

The analysis of the fast subsystem is relatively straightforward and depends on whether it is of interest in its own right or as input to slow subsystems as discussed in [4]. If it is of interest as an input to slow subsystem, then as μ tends to zero the process $x_2(t)$ tends to a white noise process

$$x_2(t) = -A_2^{-1} B_2 w(t) + \text{error} \quad (3)$$

where it is assumed that all the values of A_2 are stable, and where the dependence on $u(t)$ is omitted for simplicity. Hence, the fast process behaves in the limit as a switched parameter white Gaussian noise whose covariance $A_2^{-1} B_2 Q B_2' A_2^{-1}$ switches among the M_1 values based on the slow states of the underlying Markov process $u(t)$. The question of the involvement in this limit of the fast component of $u(t)$ is still open. The error can be expressed in terms of the integral of $x_2(t)$ and it can be shown that the mean-squared error is $O(\mu)$ as in the standard singularly perturbed case.

When the limit is required for the purpose of analyzing the fast state $x_2(t)$ directly, then the analysis need to be carried in the stretched time-scale $(t-t_1)/\mu$ after appropriate scaling (as mentioned in [6], for example) of the white noise

process to obtain finite variance for $x_2(t)$. The time instants $\{t_i\}$ represent the transition times from the slow states of the process to a fast state of the process. In this case it can be shown that in the stretched time-scale the fast subsystem may be modeled as a switched parameter process depending only on the fast states of the Markov chain. When $u(t)$ takes values among the slow states, then in the stretched time-scale, $x_2(t)$ behaves approximately as any time invariant process with constant parameters held to their values at the last slow transition.

SLOW MODE SYSTEM

The analysis of the slow modes of the system is based on writing the solution of $x_1(t)$ as a function of the fast states for the duration of intervals of transitions among the fast states of $u(t)$. The solution is given as a standard state equation solution of a time varying linear system with white noise input $w(t)$. The time varying nature stems from the dependence of the system matrix on the different values of the fast states of $u(t)$. Let these values be denoted by $A_1[u_{f_i}]$ where $\{u_{f_i}, i = 1, 2, \dots, M_2\}$ are the fast states of $u(t)$. A crucial assumption for the proof of the results is that these matrices $A_1[u_{f_i}]$ commute with each other. It is shown by finding the mean-squared error of the approximation and taking its limit as μ tends to zero that the approximate model for $x_1(t)$ during the fast transitions is as follows

$$\dot{\bar{x}}_1(t) = \bar{A}_1 \bar{x}_1(t) + B_1 w(t) \quad (4)$$

where $\bar{A}_1$ is the statistical average value of $A_1[u_{f_i}]$. Hence during the fast transition intervals the mean-squared error between x_1 and $\bar{x}_1$ is $O(\mu)$ and tends to zero as μ tends to zero.

The resulting approximation implies that the slow process can be approximated by a switched parameter process depending on the aggregated slow process $\bar{u}(t)$ that has M_1+1 slow states with given by $\bar{u}_{s_i}, i = 1, 2, \dots, M_1+1$, and where $\bar{u}_{s_i} \in S_1$ for $i \leq M_1$, and where the value of A_1 as a function of the last state (the aggregated state of the fast states) is equal to $\bar{A}_1$.

It is important to note that the proof is derived by showing a mean-squared limiting procedure and not a weaker form of convergence. The derivation is based on finding a set of differential equations for the conditional means of the various terms of the mean-squared error, conditional on the last value of the state of the Markov chain. The differential equations become a set of singularly perturbed differential equations due to the dependence on $P_{22}(t/\mu)$. The resulting expectations therefore tend to zero as μ tends to zero. The proof also allows the derivation of first order approximations to the error for nonzero μ .

CONCLUDING REMARKS

This note considered the limiting behavior of a singularly perturbed stochastic linear system with switched parameters which depend on a singularly perturbed finite state Markov process. The

two main approximations that become possible with such a model are: The fast continuous process can be approximated by a white noise with random covariance that switches according to the slow states of the Markov process. The slow continuous process can be approximated by adding an additional state to the slow Markov process that yields an average value of the system matrix over all their values based on the fast states. The results depend on two crucial assumptions, the ergodicity of the Markov chain, and the fact that the values of the system matrix commute. The results allow the derivation of higher order correcting terms to the approximations.

It is crucial to relax some of the restrictions imposed by the proofs derived in this case. Also, the case when the time-scales of the fast Markov process are different from the time-scales of the fast subsystem need further study. The motivation for studying this problem is in its potential for deriving simplified filtering schemes for switched parameter systems. Typically these schemes require expanding memory as more observation samples are taken. Such aggregations and approximations derived in this note may be helpful in obtaining approximate implementable schemes.

ACKNOWLEDGEMENTS

This research is supported by the U. S. Air Force Armament Laboratory, under Contract F08635-84-C-0273.

REFERENCES

- [1] J. K. Tugnait, "Detection and Estimation for Abruptly Changing Systems", *Automatica*, Vol. 18, pp. 607-615, 1982.
- [2] G. S. Ladde and D. D. Siljak, "Multiplex Control Systems: Stochastic Stability and Dynamic Reliability", *Int. J. Control*, Vol. 38, pp. 515-524, 1983.
- [3] T. L. Johnson, "Synchronous Switched Linear Systems," *Proc. 24th IEEE Conf. on Decision and Control*, Ft. Lauderdale, December 1985, pp. 1699-1700.
- [4] A. H. Haddad, "Linear Filtering of Singularly Perturbed Systems", *IEEE Trans. Automatic Control*, Vol. AC-21, pp. 515-519, 1976.
- [5] M. Coderch, A. S. Willsky, S. S. Sastry, and D. A. Castanon, "Hierarchical Aggregation of Singularly Perturbed Finite State Markov Processes", *Stochastics*, Vol. 8, pp. 259-289, 1983.
- [6] H.K. Khalil, A.H. Haddad, and G.L. Blankenship, "Parameter Scaling and Well-Posedness of Stochastic Singularly Perturbed Control Systems," *Proc. 12th Asilomar Conference on Circuits, Systems and Computers*, November 6-8, 1978, Pacific Grove, CA.

APPENDIX W

ON LINEAR SINGULARLY PERTURBED SYSTEMS WITH QUANTIZED CONTROL

ON LINEAR SINGULARLY PERTURBED SYSTEMS WITH QUANTIZED CONTROL

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

ABSTRACT

The effect of quantized control on an otherwise linear singularly perturbed system is analyzed. Three cases are studied: open loop control, closed loop control with small quantization step size, and closed loop control with large quantization step size. The results of a numerical example are also given to demonstrate the analytical techniques described in the paper.

1. INTRODUCTION

This work examines the effect of quantization on a singularly perturbed linear control system. Singular perturbation theory is often used for dynamic systems possessing both slow and fast dynamics to simplify the analysis and control design [1,2]. One of the requirements under which a dynamic system may be separated into slow and fast models using standard singular perturbation techniques is that the system is smooth with respect to its variables including the control input and continuous in time [1,3]. However, in many applications the actuators controlling the system have discrete states supplying piecewise-constant or quantized control so that the smoothness requirement is not met. This type of actuator includes stepper motors and certain types of hydraulic and pneumatic devices [4,5]. Quantization may also occur as a result of a digital controller quantizing either the measurements or the signal to the actuator. As shown in this paper, these systems may still be separated into slow and fast models using the techniques described here.

The system under consideration is linear, time-invariant and is represented by

$$\dot{x}(t) = A_{11}x(t) + A_{12}z(t) + B_1u(t) \quad (1)$$

$$\mu\dot{z}(t) = A_{21}x(t) + A_{22}z(t) + B_2u(t) \quad (2)$$

where $\mu > 0$ is small [1]. The quantized control input, $u(t)$, may represent an open loop command signal or a closed loop feedback signal. Both cases are analyzed in this paper.

If A_{22} is invertible, the system in (1)-(2) may be decoupled using the transformation cited in [6] with the resulting form,

$$\dot{\xi}(t) = A_0\xi + B_0u; \quad \xi(t_0) = \xi_0 \quad (3)$$

$$\mu\dot{n}(t) = A_2n + B_2u; \quad n(t_0) = n_0 \quad (4)$$

where $A_0 = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $B_0 = B_1 - A_{12}A_{22}^{-1}B_2$ and $A_2 = A_{22}$. The new variables ξ and n are expressed up to an error of order $O(\mu)$ by

$$\begin{bmatrix} \xi \\ n \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ A_{22}^{-1}A_{21} & I_m \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} \quad (5)$$

where I_n and I_m are $n \times n$ and $m \times m$ identity matrices, respectively. The natural modes of (3) correspond to the slow modes of the original system and the natural modes of (4) correspond to the fast modes of the original system.

If a stretched time-scale $\tau = (t - t_0)/\mu$ is defined, (4) may be expressed as

$$\frac{dn(\tau)}{d\tau} = A_2n(\tau) + B_2u; \quad n(0) = n_0 \quad (6)$$

where $n(\tau)$ denotes the transformed function $n(\mu\tau + t_0) = n(t)$. In this equation, $n(\tau)$ is composed of a transient due to the initial conditions and may have a steady-state value with respect to τ . This fast transient in τ is known as the boundary layer solution and its contribution to $n(t)$ is only significant for t in a short interval $[t_0, t_1]$ known as the boundary layer. Also, note that ξ remains relatively constant with respect to τ so that $\xi(\tau) = \xi_0$ [1]. Without loss of generality the paper considers decoupled systems of the form given in (3), (4), or (6).

The following is an outline of the paper. Section 2 discusses the effect of open loop piecewise-constant control input on the decoupled system. The effect of closed loop quantized control on (3)-(4) is shown in Section 3 for cases involving both small and large quantization steps. The latter case is demonstrated via a numerical example in Section 4 involving systems (1)-(2). Section 5 concludes the paper.

2. OPEN LOOP QUANTIZED CONTROL

This section considers singularly perturbed linear systems excited by an open loop piecewise constant control input. This control input is assumed to be separable into a slow switching component, u_s , and a fast switching component, u_f . The slow and fast modes of the system are assumed to be decoupled as in (3)-(4) with input $u = u_s + u_f$. The magnitude of the input is restricted to be in a finite class of allowable levels, i.e., quantization levels. Any change in the input is characterized by a discontinuity with minimum height equal to the smallest quantization step.

The decomposition of u may be defined as follows

$u_s(t) = u(t) = \text{constant for } t \in [t_i, t_{i+1}]$, such that the switching interval $[t_i, t_{i+1}]$ is large relative to μ .

$$u_f(t) = u(t) - u_s(t).$$

The response of the decoupled system (3)-(4) is examined due to the switching of u_s and u_f . It is found that any change to u_s excites the fast modes of (4) requiring analysis of the boundary layer. Furthermore, control of the fast modes in the boundary layer requires application of the u_f control immediately following a switch in u_s . If $\bar{u}_f$ represents the steady state component of u_f (with respect to the fast time-scale), then the control $(u_f - \bar{u}_f)$ has $O(\mu)$ effect on the slow variable $\xi(t)$. The $\bar{u}_f$ control affects $\xi(t)$ in the same manner as u_s . (Actually, the original u may be decomposed so that u_s contains $\bar{u}_f$ and u_f has no steady state value). Some details of this analysis are given below.

The response of $n(t)$ and $\xi(t)$ due to a switch in u_s is examined first. Let u_s switch from u_{s1} to u_{s2} at t_1 and define the expanded time variable $\tau = (t - t_1)/\mu$. Equation (6) becomes

$$\frac{dn(\tau)}{d\tau} = A_2 n(\tau) + B_2 u_{s2}; \quad n(0) = -A_2^{-1} B_2 u_{s1}. \quad (7)$$

In the boundary layer $\xi(\tau) = \xi(0) + O(\mu)$. The solution to (7) is given by

$$n(\tau) = \Phi_f(\tau, 0) n(0) + [\Phi_f(\tau, 0) - I] A_2^{-1} B_2 u_{s2} \quad (8)$$

where $\Phi_f(\tau, t_1) = \exp[A_2(\tau - t_1)]$. Note that if A_2 is a stable matrix then,

$$\lim_{\tau \rightarrow \infty} n(\tau) = -A_2^{-1} B_2 u_{s2} = \lim_{\mu \rightarrow 0} n(t)$$

for t outside of the boundary layer (i.e. for $t \in (t_1 + \delta, t_2]$, where $\delta > 0$). The response of $\xi(t)$ for $t \in (t_1 + \delta, t_2]$ can be expressed as

$$\xi(t) = \Phi_s(t, t_1) \xi(t_1) + [\Phi_s(t, t_1) - I] A_0^{-1} B_0 u_{s2} \quad (9)$$

where $\Phi_s(t, t_1) = \exp[A_0(t - t_1)]$.

If the fast system is to be controlled in the boundary layer $[t_1, t_1 + \delta]$, $u_f(\tau)$ must be added in equation (7) ($\bar{u}_f$ is assumed to be zero). The resulting system becomes

$$\begin{aligned} \frac{dn(\tau)}{d\tau} &= A_2 n(\tau) + B_2 (u_{s2} + u_f(\tau)); \\ n(0) &= -A_2^{-1} B_2 u_{s1}. \end{aligned} \quad (10)$$

Assuming $u_f(\tau)$ is bounded, $\xi(\tau) = \xi(0) + O(\mu)$. Let τ_1 and τ_{f1} denote the switching times and the corresponding values, $u_f(\tau_1)$, of $u_f(\tau)$, respectively. Then the solution to (10) becomes

$$\begin{aligned} n(\tau) &= \Phi_f(\tau, 0) n(0) + [\Phi_f(\tau, 0) - I] A_2^{-1} B_2 u_{s2} \\ &+ \sum_{i=0}^{n-1} [\Phi_f(\tau, \tau_i) - \Phi_f(\tau, \tau_{i+1})] A_2^{-1} B_2 u_{fi}, \quad \tau \geq \tau_n \end{aligned} \quad (11)$$

where n is the number of switches and $u_{fn} = 0$. Again note that $n(\tau) \rightarrow -A_2^{-1} B_2 u_{s2}$. The effect of $u_f(\tau)$ on $\xi(t)$ is $O(\mu)$ as can be seen from the solution of (3)

$$\begin{aligned} \xi(t) &= \Phi_s(t, t_1) \xi(t_1) + [\Phi_s(t, t_1) - I] A_0^{-1} B_0 u_{s2} \\ &+ \sum_{i=0}^{n-1} (\tau_{i+1} - \tau_i) B_0 u_{fi}. \end{aligned} \quad (12)$$

A time-varying version of the system in (3)-(4) with open loop piecewise constant control may be similarly analyzed. The constraint on the system is that $|A_2|$, $|B_2|$, $|A_0|$, $|B_0|$ are bounded.

Therefore, linear systems exhibiting a separation into fast and slow dynamics without control input also exhibit this behavior when the input is piecewise constant and separable into fast and slow components. Such systems can be reduced into a slow model in the slow time-scale t , and a fast model in the fast time-scale τ .

3. CLOSED LOOP QUANTIZED CONTROL

3.1 Small Quantization Steps

Quantization effects in the closed loop system may be represented by a white Gaussian noise input if the quantization steps are small [7]. Singularly perturbed systems with white noise input have been studied previously, see, for example [8,9]. This work differs from other papers on stochastic singularly perturbed systems in that along with the noise input which is white in the normal time-scale, t , there is also a noise input which is white in the expanded time scale, τ (used to implement the fast controller).

The control, $u(t)$, in equations (1)-(2) for linear state feedback may be represented by

$$u = -K_s x - K_f z + w_s(t) + w_f(\tau) \quad (13)$$

where w_s is Gaussian white noise in t and w_f is white noise in τ . These noises are assumed to be independent with autocorrelations given by

$$\begin{aligned} E(w_s(t_1) w_s(t_2)) &= Q_s \delta(t_1 - t_2) \text{ and} \\ E(w_f(\tau_1) w_f(\tau_2)) &= Q_f \delta(\tau_1 - \tau_2) \end{aligned} \quad (14)$$

(The independence assumption may not be very accurate and merits further investigation). The effects of w_s and w_f on the mean and covariance of n and ξ are examined below.

Substituting (13) into (1)-(2) yields the new closed loop system

$$\dot{x} = (A_{11} - B_1 K_s) x + (A_{12} - B_1 K_f) z + B_1 (w_s + w_f) \quad (15)$$

$$\dot{z} = (A_{21} - B_2 K_s) x + (A_{22} - B_2 K_f) z + B_2 (w_s + w_f) \quad (16)$$

This system may then be decoupled using the transformation mentioned in Section 1. The resulting system will be as in (3)-(4) with input $u = (w_s + w_f)$ and system matrices given by

$$\begin{aligned} A_0 &= A_{11} - B_1 K_s - \\ &\quad (A_{12} - B_1 K_f)(A_{22} - B_2 K_f)^{-1}(A_{21} - B_2 K_s) B_2 \\ B_0 &= B_1 - A_{12}(A_{22} - B_2 K_f)^{-1} B_2 \\ A_2 &= A_{22} - B_2 K_f \end{aligned} \quad (17)$$

In order to examine the variance of n , the time scale is expanded by defining $\tau = (t - t_0)/\mu$ and (6) is rewritten using the definitions in (17)

$$\frac{dn(\tau)}{d\tau} = A_2 n(\tau) + B_2(w_s + w_f); n(0) = n_0 \quad (18)$$

Since w_s and w_f are independent, the variance of $n(\tau)$, $P_n(\tau)$, consists of a component due to w_s , $P_{ns}(\tau)$, and a component due to w_f , $P_{nf}(\tau)$, i.e.,

$$P_n = P_{ns} + P_{nf} \quad (19)$$

P_{nf} is the solution of the following Liapunov equation [10]

$$\frac{dP_{nf}}{d\tau} = A_2 P_{nf} + P_{nf} A_2' + B_2 Q_f B_2' \quad (20)$$

Note that if A_2 is asymptotically stable, P_{nf} approaches a steady state limit, $\bar{P}_{nf}$ as $\tau \rightarrow \infty$. The variance P_{ns} has been shown previously to be $O(1/\mu)$ [8]. In fact, as $\mu \rightarrow 0$ n approaches a white noise process. Clearly, P_{ns} dominates P_{nf} for small μ ; therefore, $P_n(\tau) \approx P_{ns}(\tau)$.

Similarly, the variance of $\xi(t)$, $P_\xi(t)$, consists of a component due to w_s , $P_{\xi s}(t)$, and a component due to w_f , $P_{\xi f}(t)$. $P_{\xi s}$ is easily found to be the solution of the following Liapunov equation [10]

$$\dot{P}_{\xi s} = A_2 P_{\xi s} + P_{\xi s} A_2' + B_1 Q_s B_1' \quad (21)$$

In order to find $P_{\xi f}$, the solution to (3) is required with input w_f

$$\xi(t) = \phi_s(t, t_0) \xi(t_0) + \int_{t_0}^t \phi_s(t, \sigma) B_1 w_f \left(\frac{\sigma - t_0}{\mu} \right) d\sigma \quad (22)$$

Performing a change of variables in the integral and post-multiplying by $\xi(t)'$ and taking the expected value yields

$$P_{\xi f}(t) = \phi_s(t, t_0) P_\xi(t_0) \phi_s'(t, t_0) + \mu^2 \int_0^\beta \phi_s(t, \mu\lambda + t_0) B_1 Q_f B_1' \phi_s'(t, \mu\lambda + t_0) d\lambda \quad (23)$$

where $\beta = (t - t_0)/\mu$. Expanding this equation in a Taylor series about $\mu = 0$ yields

$$P_{\xi f}(t) = \phi_s(t, t_0) P_\xi(t_0) \phi_s'(t, t_0) + O(\mu). \quad (24)$$

Therefore, the effect of $w_f(\tau)$ on the covariance of $\xi(t)$ is $O(\mu)$ and $P_\xi(t) \approx P_{\xi s}(t)$.

Evidently, the variance of both the fast and the slow time variables are dominated by the white noise in t . Furthermore, as $\mu \rightarrow 0$, the white noise in τ has negligible effect on either the slow or the fast systems. Therefore, standard singular perturbation techniques for stochastic systems may be applied to this problem if the quantization steps are small.

3.2 Large Quantization Steps

When the quantization steps are large compared to the feedback signal, the stochastic error model described in Section 3.1 does not apply. Also, the usual perturbation series expansion techniques are not applicable to this system due to the discontinuity in the quantifier function [3]. However, the system in equations (1)-(2) can still be separated into fast and slow subsystems with $O(\mu)$ errors in the variables. The analysis for this separation is carried out for the decoupled system in equations (3)-(4) with a scalar input and can suitably be applied to the original system.

In the case of scalar quantized feedback, the control input u in (3)-(4) can be written $u = Q(-K_1 \xi - K_2 n)$ where K_1 and K_2 are row vectors representing gain and $Q(\cdot)$ is a quantizer function defined as follows

$$Q(x) = c_i, \quad d_i \leq x < d_{i+1} \quad (25)$$

for $i = 1, 2, \dots, n$.

It is assumed that c_i, d_i are defined such that $d_1 = -\infty, d_{n+1} = +\infty$, and $c_i < c_{i+1}, d_i < d_{i+1}$. An example of this function is given in Fig. 1.

The fast system is written in the form of equation (6) where $u = Q(-K_1 \xi_0 - K_2 n(\tau))$ assuming that $\xi(\tau) = \xi_0$ is constant with respect to τ . The dependence of $Q(\cdot)$ on ξ_0 may be removed by defining the following modified quantizer function $Q'(x) = Q(-K_1 \xi_0 - x)$

$$Q'(x) = c_i, \quad -(d_i + K_1 \xi_0) \leq x < -(d_{i+1} + K_1 \xi_0) \quad (26)$$

for $i = 1, 2, \dots, n$.

Since u can take on only a finite number of values, this system is actually piecewise linear. The phase space corresponding to this system is partitioned into regions associated with each possible value c_i of u as seen from the definition of $Q'(K_2 n)$. The boundaries of these regions are defined by the solutions n_i of the following set of expressions:

$$(K_2 n = -(d_i + K_1 \xi_0); i = 1, 2, \dots, n) \quad (27)$$

Regional equilibrium points are defined to be $n_i = -A_2^{-1} B_2 c_i$ so that n_i governs the behavior of $n(\tau)$ inside the i th region. Note that if a regional equilibrium point does not lie in its associated region, it is not a global equilibrium point. Also, the assumption that A_2 is stable means that there may exist a global equilibrium point on one of the boundaries. Inspection of the function $Q'(K_2 n)$ shows that only one possible such equilibrium exists for a given value of ξ_0 .

As $\tau \rightarrow \infty$, n approaches its global equilibrium point. Let n_s be the equilibrium point and define

$$n_f(\tau) = n(\tau) - n_s. \quad (28)$$

In the normal time scale, t , n_f is seen to decay rapidly from its initial condition to zero in the initial boundary layer. As the slow variable $\xi(t)$ begins to decay, the "equilibrium point" n_s of the fast system changes values. In fact, n_s can be written entirely as a continuous function, $n_s = f(K_1 \xi(t))$, where $f(x)$ is as shown

$$\begin{aligned} \text{i) } f(x) &= n_i, \text{ if } -d_{i+1} - K_2 n_i \leq x \leq -d_{i+1} - K_2 n_{i+1} \\ &\quad \text{for some } i, \\ \text{ii) } f(x) &= m(x + d_{i+1} + K_2 n_i) + n_i, \text{ if} \\ &\quad -d_{i+1} - K_2 n_i \leq x \leq -d_{i+1} - K_2 n_{i+1} \text{ for some } i, \end{aligned} \quad (29)$$

where $m = (n_i - n_{i+1}) / (K_2(n_{i+1} - n_i))$.

Part i) corresponds to a regional equilibrium lying inside its associated region and part ii) corresponds to an equilibrium point lying on one of the boundaries. An example of this function is shown in Fig. 2 for a second order system. The flat portions of the graphs correspond to part i) of the function. The sloped lines are a result of simple interpolation between the flat parts and correspond to part ii). Note that the slope of these lines is roughly inversely proportional to K_2 . If this slope is less than $O(1/\mu)$ in order of magnitude, the variation of n_s outside the boundary layer does not excite n_f by more than $O(\mu)$.

Therefore, since the fast variable satisfies

$$\dot{n}(t) = n_f(t) + n_s(t) + O(\mu), \quad (30)$$

it is approximated by $n_f(t) + n_s(t_0)$ in the boundary layer and by $n_s(t)$ otherwise. The slow variable, ξ , is approximated up to $O(\mu)$ error by

$$\begin{aligned} \frac{d\xi}{dt} &= A_0 \xi + B_0 u; \quad \xi(t_0) = \xi_0 \\ u &= Q(-K_1 \xi - K_2 n_s) \end{aligned} \quad (31)$$

where $n_s(t) = f(K_1 \xi(t))$.

This procedure can be extended to the multiple input case. However, the complexity of finding the equilibrium points increases.

A remark can be made about the comparison between analysis by time integration of the actual system (3)-(4) and by time integration of the reduced systems (6) (where $u = Q(-K_1 \xi_0 - K_2 n(\tau))$) and (31). The usual benefits of using the reduced models include integrating lower order systems and using a larger time step for the slow model. A peculiarity of this system is that u_f does not necessarily go to zero as n_f goes to zero. This is a consequence of an equilibrium point lying on the boundary between regions in the phase space. If n spirals into such an equilibrium point, it keeps crossing into different regions characterized by u switching between two values. In other words, n_f is driven to zero by u_f which must keep switching to maintain the zero value of n_f . In the reduced models, the boundary layer need only be evaluated until n_f decays sufficiently regardless of u_f . The slow model does not depend on u_f . However, the constant fast switching of u_f may cause numerical problems in time-integrating (3)-(4)

directly due to time step considerations.

4. NUMERICAL EXAMPLE

An example of the separation method discussed in Section 3.2 is demonstrated here. The system is given by equations (1)-(2) where [11]

$$\begin{aligned} A_{11} &= \begin{bmatrix} 0 & 0.4 \\ 0 & 0 \end{bmatrix} & A_{12} &= \begin{bmatrix} 0 & 0 \\ 0.345 & 0 \end{bmatrix} & B_1 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ A_{21} &= \begin{bmatrix} 0 & -0.524 \\ 0 & 0 \end{bmatrix} & A_{22} &= \begin{bmatrix} -0.465 & 0.262 \\ 0 & -1 \end{bmatrix} & B_2 &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ K_s &= [1 \ 0.89202] & K_f &= [0.24396 \ 0.061996] \end{aligned}$$

$$x(0) = z(0) = [2.5 \ 0]' \quad (32)$$

and $\mu = .1$ and $u = Q(-K_s x - K_f z)$. The quantizer is uniform with five possible states,

$$Q(x) = \begin{cases} -2 & x < -1.5 \\ -1 & -1.5 \leq x < -.5 \\ 0 & -.5 \leq x < .5 \\ 1 & .5 \leq x < 1.5 \\ 2 & 1.5 \leq x \end{cases} \quad (33)$$

Decoupling the system using the transformation of Section 1, the regional equilibrium points are found to be

$$\begin{aligned} n_1 &= [-1.127 \ -2]' & n_4 &= [.563 \ 1]' \\ n_2 &= [-.563 \ -1]' & n_5 &= [1.127 \ 2]' \\ n_3 &= [0 \ 0]' \end{aligned} \quad (34)$$

and the initial equilibrium $n_s(0) = n_1$. The function $n_s = f(K_1 \xi)$ is shown in Fig. 2. Note that the slope for n_{s1} is approximately 2.8 and for n_{s2} is 5. This is less than $10 = 1/\mu$ but it is large enough to cause excitation of the fast variable n_f as seen in the results.

The phase plane plot of $n_1(t)$ versus $n_2(t)$ found by integrating the actual system is shown in Fig. 3 along with the initial equilibrium point $n_s(0)$. The trajectory is attracted to $n_s(0)$, then follows along the line of possible equilibrium points to zero. This system is approximated by two systems: a fast one and a slow one. The fast system given by equation (6) with appropriate parameter matrices starts at a value of $n(0) - n_s(0)$ and decays to zero in τ as $n(\tau) - n_s(0)$. This is superimposed with the slow system given by equation (31) integrated with respect to t , so that $n(t) = n_f(t) + n_s(t)$. The coordinates then are transformed back to x and z for comparison.

Comparison between time integration of the original system and the approximated system are shown in Fig. 4 to Fig. 7. Note that the initial boundary layer is tracked very accurately. With the exception of z_2 , the errors between models are $O(\mu)$. The fast variable, n_f , does appear to be excited in both z_1 and z_2 near the transitions corresponding to $n_s(t)$ moving along one of the slopes in Fig. 2. The CPU time required for time-integration of the actual system was 5-10

times longer than required for the approximated system. Also, convergence problems occurred in the actual system time integration while none appeared in the approximated system. As μ decreases, numerical evaluation of the actual system encounters more problems while the approximated system becomes more accurate.

5. SUMMARY

The effect of quantization on a singularly perturbed linear system is discussed. Three cases of quantized control are studied: open loop, closed loop with small quantization steps, and closed loop with large quantization steps. The open loop quantized control is decomposed into a slow switching and a fast switching component. The fast switching has $O(\mu)$ effect on the slow subsystem, yet any change in the slow control excites the fast subsystem and requires evaluation of the boundary layer. The closed loop system with small quantization steps is evaluated by modelling the quantizer errors as white noise in t and white noise in τ . The white noise in t dominates the behavior of both the slow and the fast subsystems. Therefore, the system may be evaluated using standard singular perturbation theory for stochastic systems. A closed loop system with large quantization steps in the feedback violates the smoothness requirement of standard perturbation methods, so a new technique for separating this system into slow and fast models is provided. The technique is successfully illustrated via a numerical example.

ACKNOWLEDGEMENT

This research is supported by the U. S. Air Force Armament Laboratory, under Contract F08635-84-C-0273.

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] 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, 1984, pp. 273-293.
- [3] J.J. Levin, "The Asymptotic Behavior of the Stable Initial Manifolds of a System of Non-linear Differential Equations," *Trans. Am. Math. Soc.*, Vol. 85, 1957, pp. 357-368.
- [4] T.J. Harned, "Making Stepper Motors Behave," *Machine Design*, Vol. 57, Sept. 26, 1985, pp. 54-58.
- [5] Parker Pneutronics, July 1984 Bulletin, Pepperall, MA.

- [6] P.V. Kokotovic and A.H. Haddad, "Controllability and Time-Optimal Control of Systems with Slow and Fast Modes," *IEEE Trans. Auto. Control*, Vol. AC-20, 1975, pp. 111-113.
- [7] R.E. Curry, *Estimation and Control with Quantized Measurements*, MIT Press, 1970.
- [8] H.K. Khalil, A.H. Haddad, and G.L. Blankenship, "Parameter Scaling and Well-Posedness of Stochastic Singularly Perturbed Control Systems," *Proc. 12th Asilomar Conference on Circuits, Systems and Computers*, November 6-8, 1978, Pacific Grove, CA.
- [9] A.H. Haddad and P.V. Kokotovic, "Stochastic Control of Linear Singularly Perturbed Systems," *IEEE Trans. Auto. Control*, Vol. AC-22, October 1977, pp. 815-820.
- [10] Kwakernaak and Sivan, *Linear Optimal Control Systems*, Wiley, N.Y., 1972.
- [11] J.H. Chow and P.V. Kokotovic, "A Decomposition of Near-Optimum Regulators for Systems with Slow and Fast Modes," *IEEE Trans. Auto. Control*, Vol. AC-21, Oct. 1976, pp. 701-704.

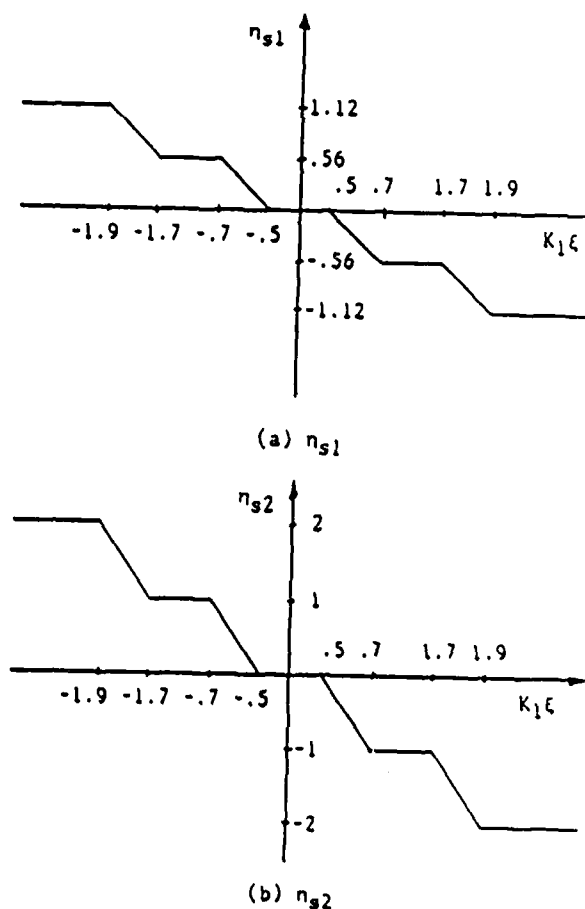


Fig. 2. n_s as a function of $K_1 \epsilon$

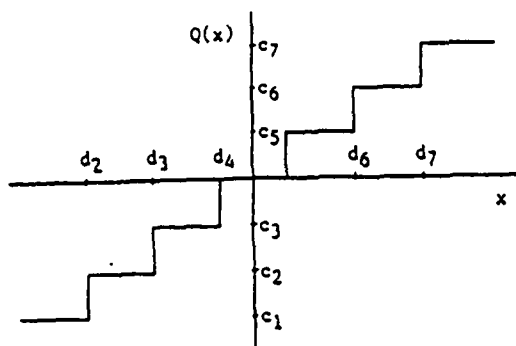


Fig. 1. Quantizer function for $n = 7$

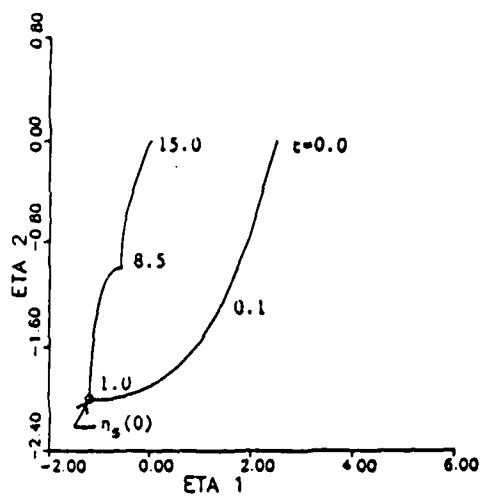


Fig. 3. Phase plane plot of n_2 vs. n_1 for the actual system

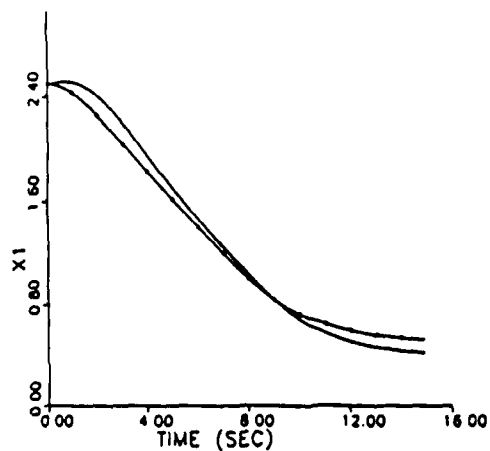


Fig. 4. Response of x_1 to initial condition for actual system (solid line) and approximate system

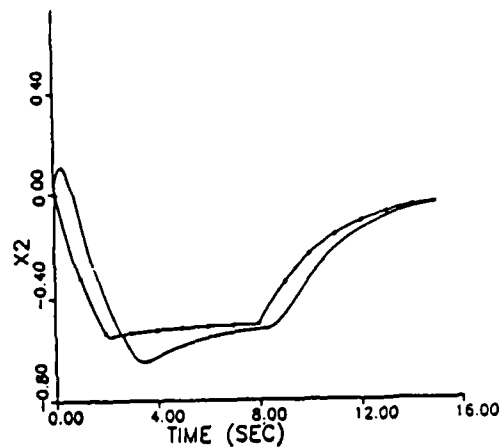


Fig. 5. Response of x_2 to initial condition for actual system (solid line) and approximate system

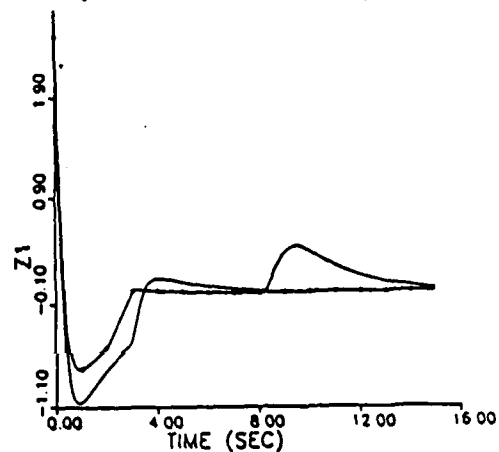


Fig. 6. Response of z_1 to initial condition for actual system (solid line) and approximate system

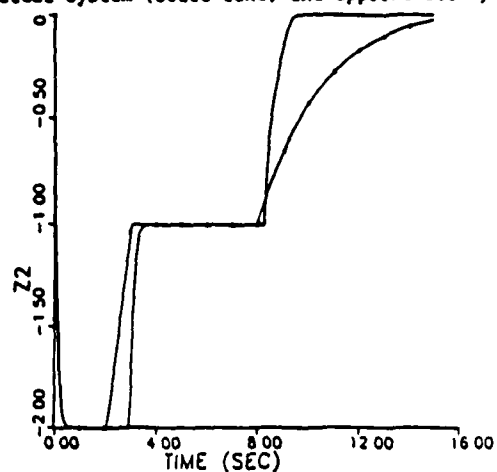


Fig. 7. Response of z_2 to initial condition for actual system (solid line) and approximate system

APPENDIX X

SINGULAR PERTURBATION IN PIECEWISE LINEAR SYSTEMS

SINGULAR PERTURBATION IN PIECEWISE-LINEAR SYSTEMS *

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

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

AD-A194 685

APPROXIMATIONS AND IMPLEMENTATIONS OF NONLINEAR
FILTERING SCHEMES(U) GEORGIA INST OF TECH ATLANTA
SCHOOL OF ELECTRICAL ENGINEERING A H HADDAD ET AL

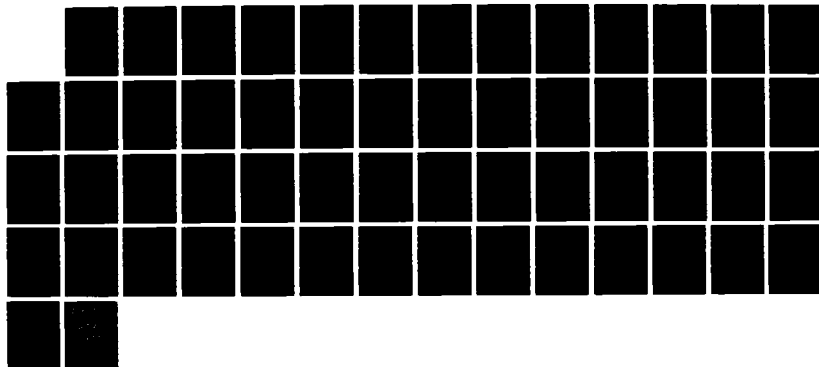
4/4

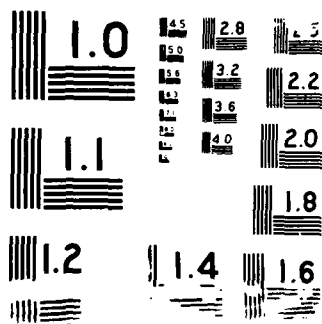
UNCLASSIFIED

FEB 88 AFATL-TR-87-73 F08635-84-C-0273

F/G 12/3

NL





1. Introduction

Systems inherently possessing both slow and fast dynamics are found commonly in electrical, mechanical and aerospace applications. 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 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. Using 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. Such systems are found in 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. These systems may exhibit both fast and slow dynamics, so it is desirable to separate the model into slow and fast subsystems.

Problem formulation

The system considered in this paper may be represented in the following form:

$$\dot{x} = f_1(x, z) \quad x(t_0) = x_0 \quad (1)$$

$$\mu \dot{z} = f_2(x, z) \quad z(t_0) = z_0 \quad (2)$$

where f_1 and f_2 are continuous piecewise-linear functions, where $\mu > 0$ is a

small parameter and where $x \in R^p$ and $z \in R^r$. The functions are linear in specific regions of the phase space (R^{p+r}) where a region is typically defined as an intersection of halfspaces. For example, equations (1) and (2) are represented in the i^{th} region by the following "linear" system:

$$\dot{x} = A_{11}^i x + A_{12}^i z + w_1^i \quad (3)$$

$$\mu \dot{z} = A_{21}^i x + A_{22}^i z + w_2^i \quad (4)$$

For the purposes of this paper, the i^{th} region is defined by the set $S_i = \{(x, z): d_{i-1} < K_x x + K_z z \leq d_i\}$ 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. The reason for this 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 on the order of $O(\mu)$. 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 develops a reduced-order model to approximate 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 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 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:

$$\frac{d\hat{z}}{d\tau} = \hat{f}(\hat{z}) \quad (5)$$

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 regions where the function is linear are found by projecting the regions from R^{p+r} of the original problem into the manifold $x = x_0$. For example, the i^{th} region in R^r is the set $R_i = \{z: d_{i-1} < K_x x_0 + K_z z \leq d_i\}$. A degenerate case where $K_z = 0$ results in the existence of only one region in R^r so that $\hat{f}$ is linear. (Note that the equilibrium point of the degenerate case is easily found assuming that the system is stable). A stable equilibrium point is the initial quasi-steady-state value, $z_s(t_0)$, of $z(t)$.

The equilibrium point(s) of system (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, see for example [3-9]. Fujisawa and Kuh show in [4] that a continuous piecewise-linear function satisfies a Lipschitz 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 $k=1,2,\dots,r$, the determinants of the $k \times k$ matrices

$$J^1_k, J^2_k, \dots, J^r_k$$

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 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 $w^i = A_{21}^i x_0 + w^i$ $\forall i$, and denote the

iteration number on z_s , y , and λ by superscripts.

- 0) initialize by letting $z_s = z_0$ and $i=1$
- 1) solve $y^i = A_{22}^j z_s^i + w^j$, where region R_j contains z_s
- 2) solve $z = z_s^i - (A_{22}^j)^{-1} y^i$
- 3) if z lies in region R_j then $z_s = z$ and stop
- 4) 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 $j=j-1$
- 5) solve $\lambda^i = (K_z z_s + K_x x_0 - d) / K_z (A_{22}^j)^{-1} y^i$
 (assuming that the denominator is not zero)
- 6) solve $z_s^{i+1} = z_s^i + \lambda^i (A_{22}^j)^{-1} y^i$
- 7) 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 many singular perturbation cases, the equilibrium point z_s can be written as an explicit function of x so the fast model is typically given in terms of $\hat{z}_f$. In this application, z_s must be found implicitly. 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

$$\begin{aligned} \frac{d\hat{z}}{d\tau} &= A_{21}^i x_0 + A_{22}^i \hat{z} + w_2^i & \hat{z}(0) &= z_0 \\ \hat{z}_f(\tau) &= \hat{z}(\tau) - z_s(t_0) \end{aligned} \quad (7)$$

where the i^{th} region is defined by the set $R_i = \{\hat{z}: d_{i-1} < K_x x_0 + K_z \hat{z} \leq d_i\}$.

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 phase 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

$$\frac{d\hat{z}}{d\tau} = A\hat{z} + Bu \quad (8)$$

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 are due to the approximation of x by x_0 , introducing errors of $O(\mu)$ into equation (7) and into the definition of the regions. Substituting $x = x_0 + O(\mu)$ into (7) and into R_i yields the system

$$\frac{d\bar{z}}{d\tau} = A_{21}^i (x_0 + O(\mu)) + A_{22}^i \bar{z} + w_2^i \quad \bar{z}(0) = z_0 \quad (9)$$

$$R_i = \{\bar{z}: d_{i-1} + O(\mu) < K_x x_0 + K_z \bar{z} \leq d_i + O(\mu)\}$$

where $\bar{z}$ 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 $\bar{z}(\tau') = \hat{z}(\tau') + O(\mu)$ then $\bar{z}(\tau'') = \hat{z}(\tau'') + O(\mu)$ for $\tau'' > \tau'$ as long as both $\bar{z}$ 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 $\bar{z}$ or $\hat{z}$ crosses into another region, the other must also cross into that region. The resulting error in the approximation remains on the order of $O(\mu)$. 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 is on the order of $O(\mu)$.

Theorem 2: Let the vector field near a boundary at $d_i = K_z \bar{z} + K_x x_0 + O(\mu)$ be given by

$$f(\bar{z}) = A_{21}^i (x_0 + O(\mu)) + A_{22}^i \bar{z} + w_2^i. \quad (10)$$

Assume that $f(\bar{z})$ does not vanish near the boundary. If $f(\bar{z})$ 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 $\bar{z}$ crosses the d_i boundary at τ' and $\hat{z}$ has not crossed any boundary. Prior to crossing $\bar{z} = \hat{z} + O(\mu)$. 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

$$K_z (A_{21}^i (x_0 + O(\mu)) + A_{22}^i \bar{z} + w_2^i) = O(\mu^0). \quad (11)$$

It follows that

$$K_z (A_{21}^i x_0 + A_{22}^i \hat{z} + w_2^i) = O(\mu^0) \quad (12)$$

Define $\hat{s}$ and $\bar{s}$ by

$$\hat{s} = K_z \hat{z} - d_i' \quad (13)$$

$$\bar{s} = K_z \bar{z} - d_i' + O(\mu) \quad (14)$$

where $d_i' = d_i - K_x x_0$. Assume $\bar{s}, \hat{s} > 0$. For $\bar{z}$ to cross the boundary,

$\frac{d\bar{s}}{d\tau} < 0$ where $\frac{d\bar{s}}{d\tau}$ 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, $\bar{s}=0$ so

that $K_z \bar{z} - d_i' = O(\mu)$. It follows that $\hat{s}(\tau') = K_z \bar{z} - d_i' + O(\mu) = O(\mu)$.

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 $\bar{z}$ 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 $\bar{z}(\tau_0)$ lie in region R_i . Then the solution of (7) for $\tau > \tau'$ is

$$\begin{aligned} \hat{z}(\tau) = & \phi(\tau, \tau_0) \hat{z}(\tau_0) + \int_{\tau_0}^{\tau} \phi(\tau, \sigma) (\Delta A \hat{z} + A_{21}^j x_0 + w_2^j) d\sigma \\ & + \int_{\tau_0}^{\tau'} \phi(\tau, \sigma) (A_{21}^i x_0 + w_2^i) d\sigma \end{aligned} \quad (15)$$

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

$$\begin{aligned} \hat{z}(\tau) = & \phi(\tau, \tau_0) \hat{z}(\tau_0) + \int_{\tau''}^{\tau} \phi(\tau, \sigma) (\Delta A \hat{z} + A_{21}^j x_0 + w_2^j) d\sigma \\ & + \int_{\tau_0}^{\tau''} \phi(\tau, \sigma) (A_{21}^i x_0 + w_2^i) d\sigma + O(\mu) \end{aligned} \quad (16)$$

Similarly, the solution to equation (9) is found to be

$$\begin{aligned} \bar{z}(\tau) = & \phi(\tau, \tau_0) \bar{z}(\tau_0) + \int_{\tau''}^{\tau} \phi(\tau, \sigma) (\Delta A \bar{z} + A_{21}^j x_0 + w_2^j) d\sigma \\ & + \int_{\tau_0}^{\tau''} \phi(\tau, \sigma) (A_{21}^i x_0 + w_2^i) d\sigma + O(\mu) \end{aligned} \quad (17)$$

Hence, $\bar{z}(\tau) = \hat{z}(\tau) + O(\mu)$.

Theorem 3: Let the vector field near a boundary at $d_i = K_z \hat{z} + K_x x_0$ be given by

$$f(\hat{z}) = A_{21}^i x_0 + A_{22}^i \hat{z} + w_2^i. \quad (18)$$

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 (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 $\bar{z}$, then $\bar{z}$ must cross within a time of $O(\mu)$. The time delay in crossing affects the error in the approximation only by $O(\mu)$.

Using the results of Theorems 2 and 3 it is seen that the errors in the approximation are on the order of $O(\mu)$. The restriction given in Section 1 that the regions of linearity be parallel is used to avoid the problem of "corners". It is not known at this time how good the approximation is if either $\hat{z}$ or $\bar{z}$ crosses a boundary at the intersection of two boundary hyperplanes.

3. Outer Solution

A reduced-order model for system (1)-(2) is developed below with approximation errors on the order of $O(\mu)$ 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 . z_s is the equilibrium point of system (7) where 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.) z_s can be determined using the

Katzenelson algorithm described in Section 2.1 where the current value of x is substituted for x_0 and the algorithm is initialized with z_s^1 equaling the previous value of z_s . Due to continuity, a small change in x results in a small change in z_s so that in almost all cases while time-integrating the system only steps 1)-3) are needed to find a new z_s . Continuity of z_s as a function of x is shown in the proof of the following theorem.

Theorem 4: Let $f: \mathbb{R}^r \rightarrow \mathbb{R}^r$ be a continuous piecewise-linear mapping defined in the i^{th} region by

$$f(z) = A_{21}^i x + A_{22}^i z + w_2^i \quad (19)$$

If f is homogeneous then the equilibrium point z_s of (20) is given by a continuous function of x .

Proof: Since f is homogeneous, a unique solution for z_s exists for any x . Let x_1 be given resulting in $z_s = z_{s,1}$. Let S_i denote the region of $(x_1, z_{s,1})$ in $\mathbb{R}^{p+r}$.

Suppose $(x_1, z_{s,1})$ lies in the interior of region S_i . Then $z_{s,1}$ can be written as

$$z_{s,1} = -(A_{22}^i)^{-1}(A_{21}^i x_1 + w_2^i) \quad (20)$$

Choose x_2 close to x_1 resulting in $z_s = z_{s,2}$ such that $(x_2, z_{s,2})$ lies in region S_i . Hence,

$$z_{s,2} = -(A_{22}^i)^{-1}(A_{21}^i x_2 + w_2^i) \quad (21)$$

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 $\|x_1 - x\| < \delta$. For the region S_i defined by the set $\{(x, z_s): d_{i-1} < K_x x + K_z z_s \leq d_i\}$ a δ is given by

$$\delta = \min \left[\frac{d_{i-1} - Mx}{\|M\|}, \frac{d_i - Mx}{\|M\|} \right]$$

where $M = K_X - K_Z(A_{22}^i)^{-1}A_{21}^i$.

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 half-plane in region S_i . If $(x_2, z_{S,2})$ lies in region S_{i+1} , then

$$z_{S,2} = -(A_{22}^{i+1})^{-1}(A_{21}^{i+1} x_2 + w_2^{i+1}) \quad (22)$$

A consequence of the continuity of f is that

$$-(A_{22}^i)^{-1}(A_{21}^i x_1 + w_2^i) + (A_{22}^{i+1})^{-1}(A_{21}^{i+1} x_1 + w_2^{i+1}) = 0 \quad (23)$$

Adding equation (23) to equation (22), subtracting the result from (20) 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 Lipschitz condition in the open half-plane 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:

$$\dot{x}_S = A_{11}^i x_S + A_{12}^i z_S + w_1^i \quad x_S(t_0) = x_0 \quad (24)$$

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 on the order of $O(\mu)$.

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

$$\dot{x} = A_{11}x + A_{12}z - B_1u \quad (25)$$

$$\mu\dot{z} = A_{21}x + A_{22}z - B_2u \quad (26)$$

$$u = \begin{cases} -1 & \text{if } K_x x + K_z z < -1 \\ K_x x + K_z z & \text{if } |K_x x + K_z z| \leq 1 \\ 1 & \text{if } K_x x + K_z z > 1 \end{cases}$$

where $\mu=0.1$. The parameter matrices are given as follows:

$$A_{11} = \begin{bmatrix} -3 & 0.4 \\ 0 & 0 \end{bmatrix} \quad A_{12} = \begin{bmatrix} 0 & 0 \\ 0.345 & 0 \end{bmatrix} \quad B_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} 0 & -0.524 \\ 0 & 0 \end{bmatrix} \quad A_{22} = \begin{bmatrix} -0.465 & 0.262 \\ 0 & -1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$K_x = [1 \ 0.861] \quad K_z = [1.220 \ 0.310]$$

The initial conditions are given as $x(0) = z(0) = [2 \ 3.]'$.

This system is put into the piecewise-linear form by substituting the expressions for u into equations (25) and (26). The three regions are defined as $S_1 = \{(x, z) : K_x x + K_z z < -1\}$, $S_2 = \{(x, z) : |K_x x + K_z z| \leq 1\}$ and $S_3 = \{(x, z) : K_x x + K_z z > 1\}$. The initial condition is located in S_3 .

The reduced-order models given in the form of equations (7) and (24) 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 order of $O(\mu)$. The computation time for the approximation was roughly half as long as 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, 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, Mar. 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.

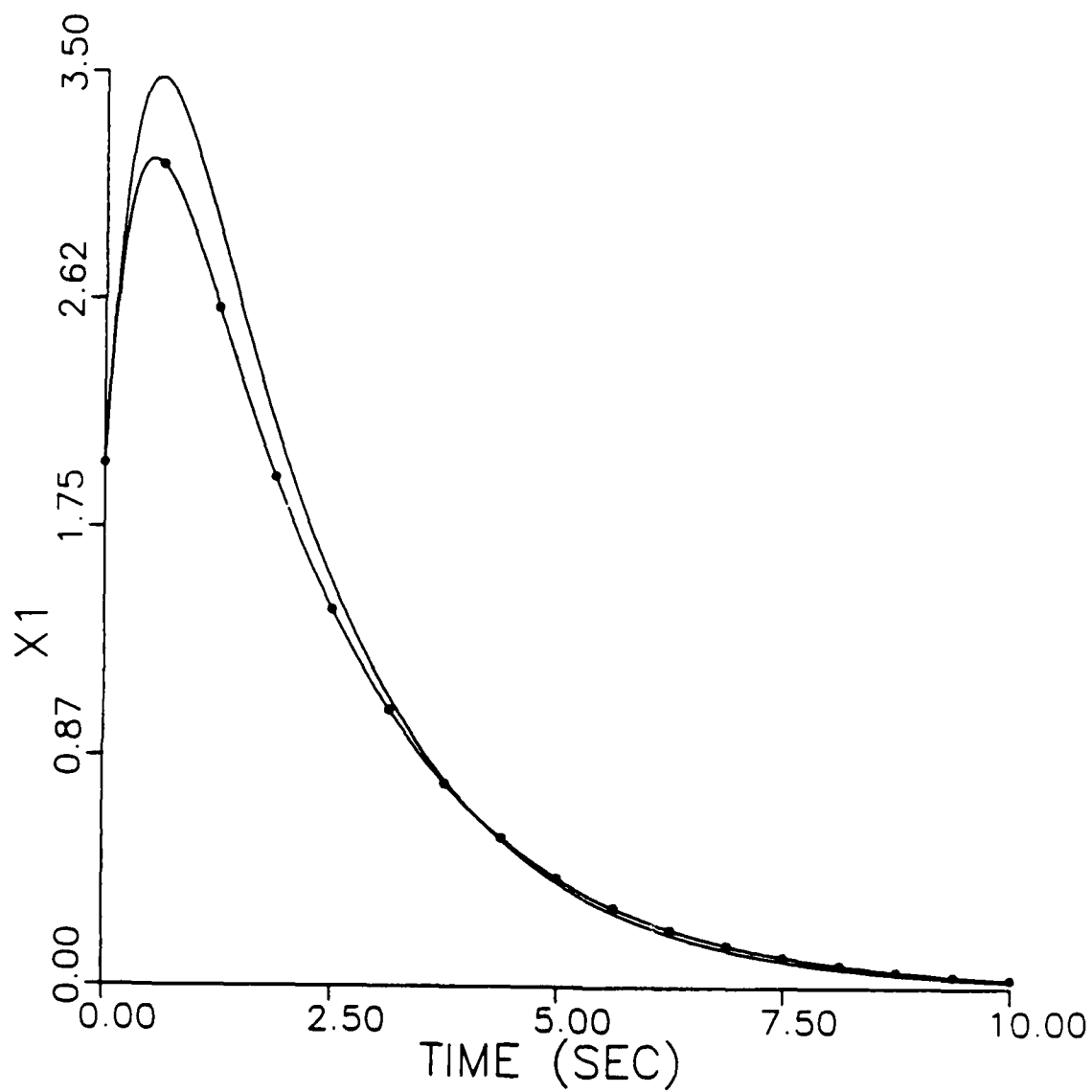


Figure 1: Response of x_1 to initial condition for actual system (solid line) and approximate system

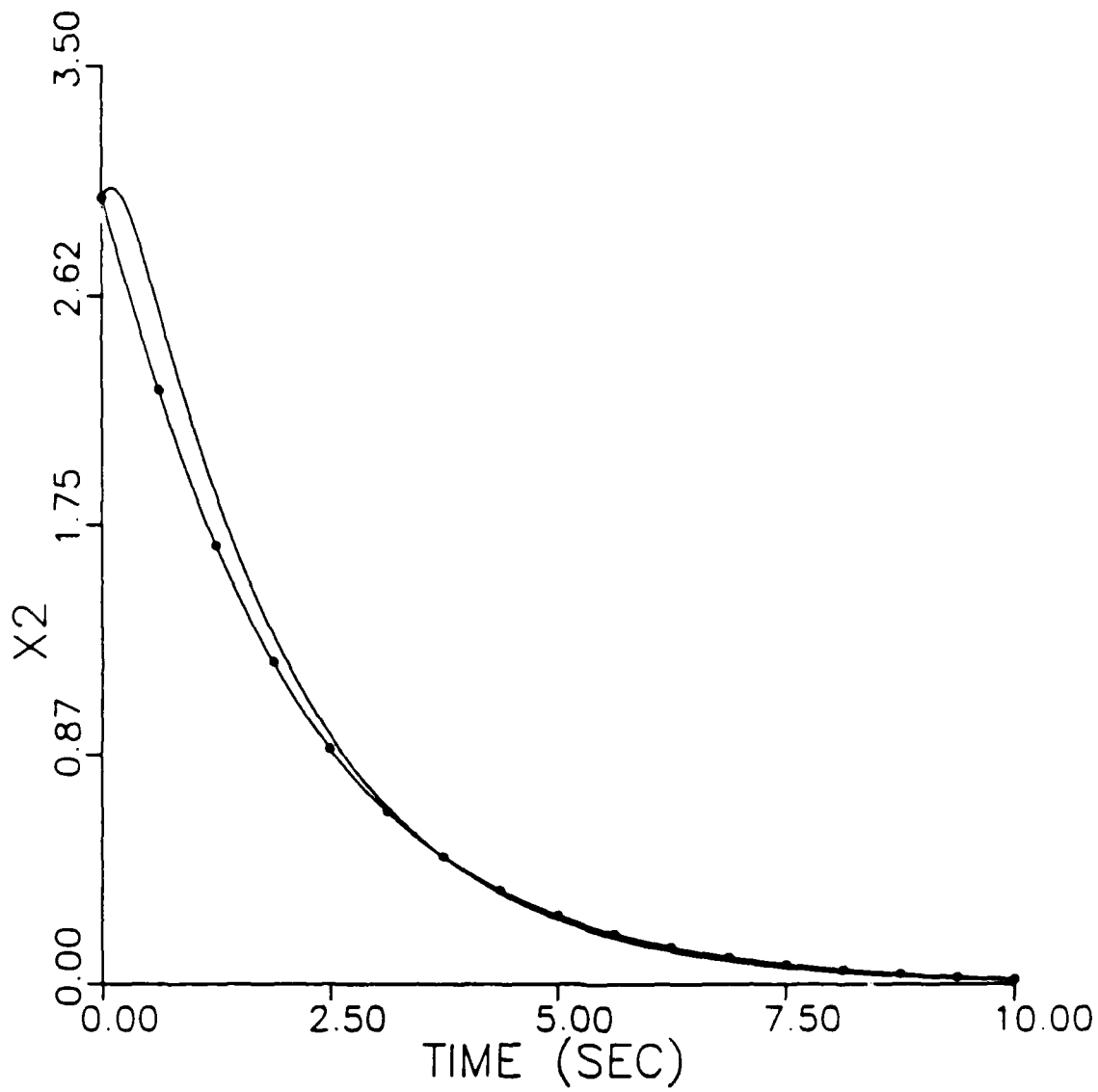


Figure 2: Response of x_2 to initial condition for actual system (solid line) and approximate system

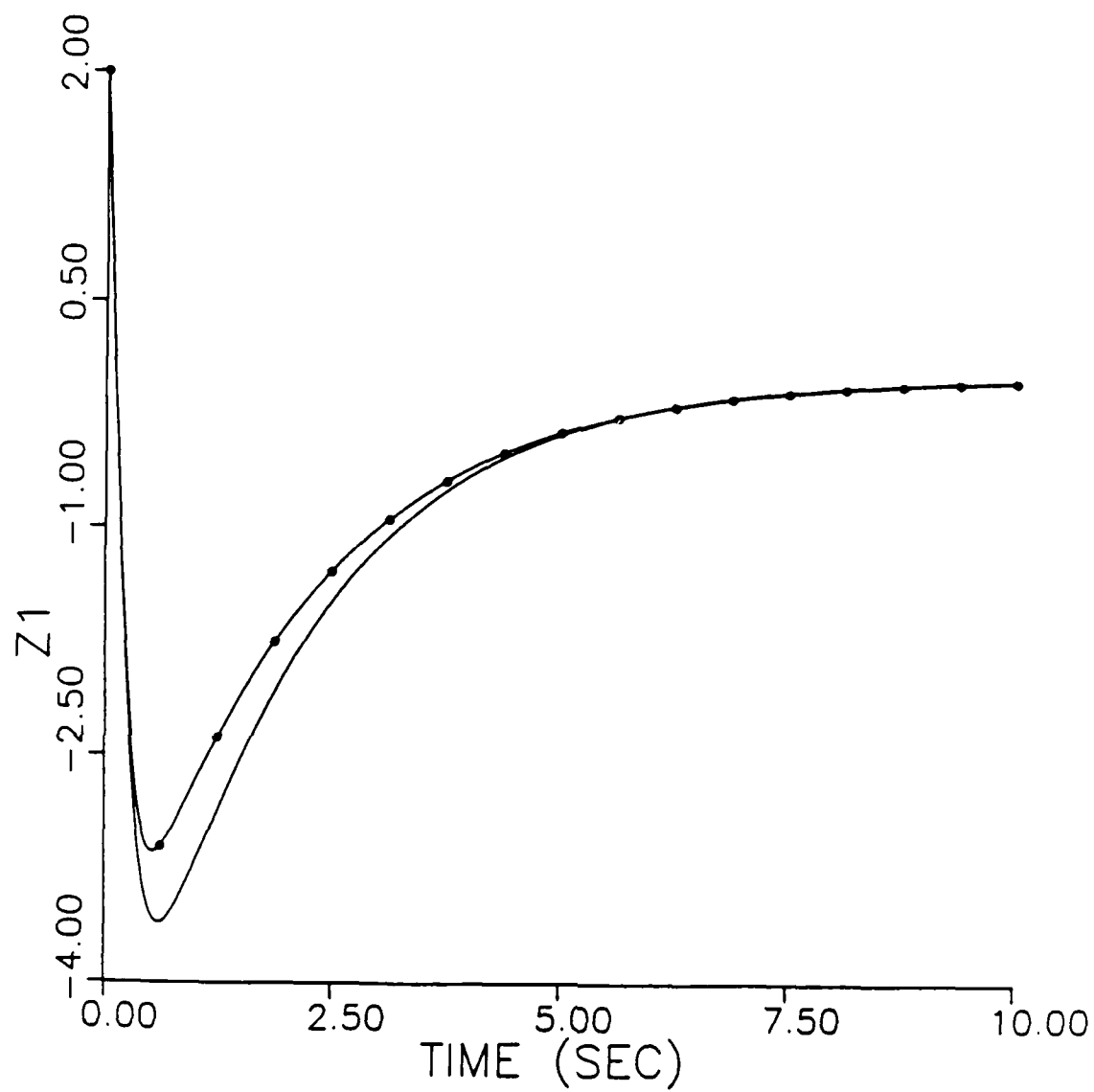


Figure 3: Response of z_1 to initial condition for actual system (solid line) and approximate system

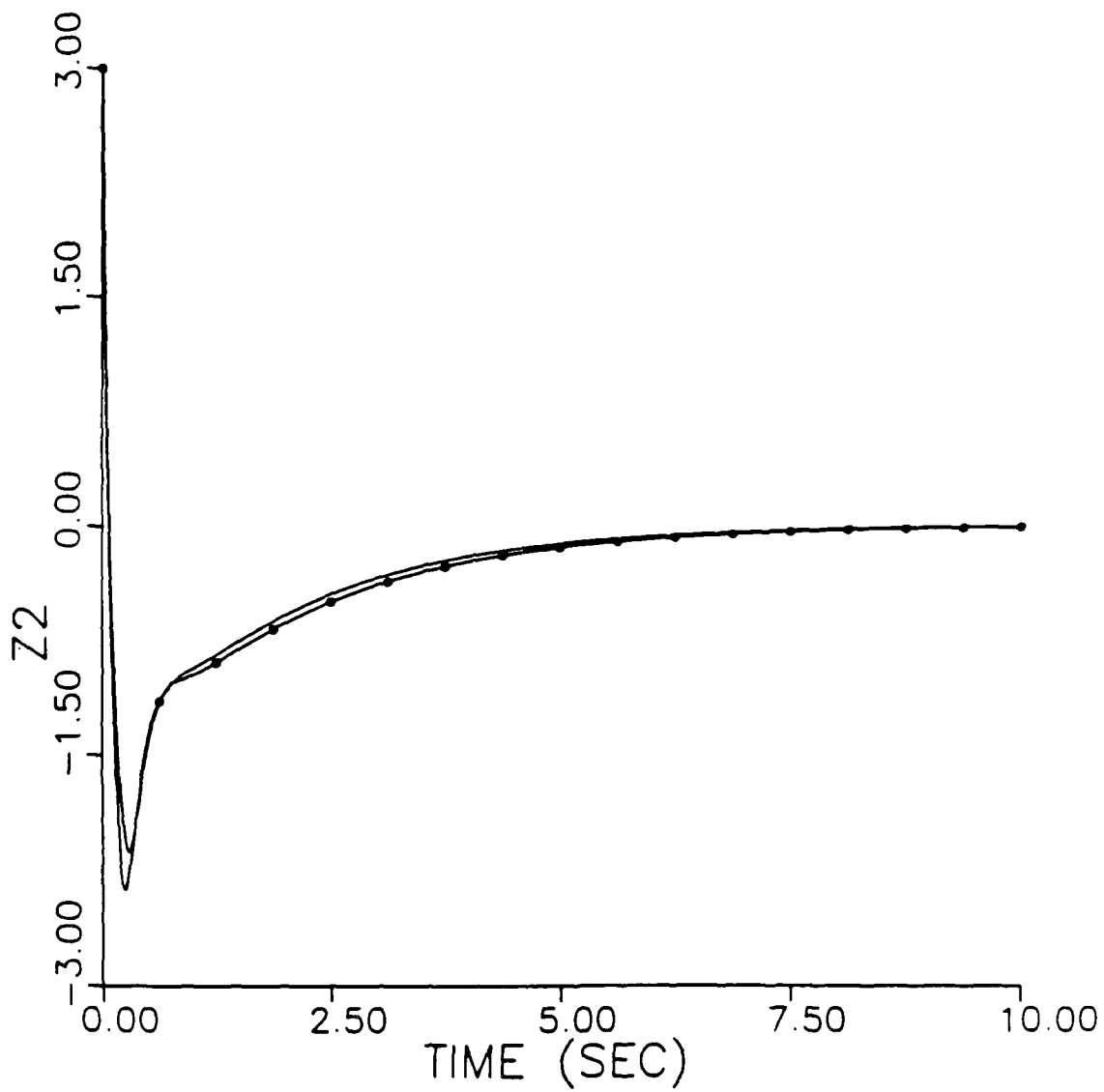


Figure 4: Response of z_2 to initial condition for actual system (solid line) and approximate system

APPENDIX Y

ON THE CONTROLLABILITY AND OBSERVABILITY OF HYBRID SYSTEMS

ON THE CONTROLLABILITY AND OBSERVABILITY OF HYBRID SYSTEMS¹

Jelal Ezzine and A. H. Haddad

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

ABSTRACT:

In this paper a special class of linear piece-wise constant time-varying systems is introduced. 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. Our main results are: a necessary and sufficient algebraic condition, a very simple algebraic criterion, and a computationally appealing algebraic sufficient test for controllability and observability. Moreover we give a simple sufficient stability condition.

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.

1. INTRODUCTION AND PROBLEM FORMULATION:

In this paper we examine the controllability, observability and related issues of hybrid systems. Hybrid systems in this paper are a special class of piece-wise constant time-varying systems. The set of constant realizations among which the model is switching is finite. Systems of this type can be used to model synchronously switched linear systems [1], networks with periodically varying switches [2] and fault-prone systems [3]. 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 to test for their stability, and most interestingly the possibility to derive algebraic controllability/observability test similar to the celebrated one found in the theory of linear time-invariant systems. This is possible because of 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 from which they bring their variable structure property that will be of a great help in controlling, observing and stabilizing them.

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

$$dx/dt = A(r(t))x(t) + B(r(t))u(t) \quad (1.1)$$

$$y(t) = C(r(t))x(t); \quad (1.2)$$

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 deterministic scalar sequence taking values in the finite index set $N = \{1, 2, \dots, N\}$.

This type 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.

The latter problem has received considerable attention from the control community but many important generalizations remain to be worked out. Chizeck et al [3] calls such a control problem the Jump Linear Quadratic (JLQ) problem since they view it as an extension 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 latter problem.

Let S_M denote any sequence of length M of the values taken by $r(t)$, and let δt_i denote 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

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

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 $M \gg N$ in the assumption on S_M will not affect the results. The $(M=N)$ -assumption makes the notation less cumbersome.

The i th form is the realization $\Sigma_i = (A_i, B_i, C_i)$ associated with the i th form index (i.e., $r(t)=i$), with $i \in N$.

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 example is used for illustration purposes. Section 7 concludes the paper and points to possible future research directions.

2. STABILITY:

Even though Hybrid systems are time-varying systems it is possible to obtain necessary and sufficient asymptotic stability conditions. To gain some

familiarity with these systems we start by studying the stability of periodic hybrid systems. To this end we recall a theorem by Willems [5] that gives a necessary and sufficient conditions under which piece-wise constant periodic systems will be uniformly asymptotically stable.

THEOREM [5]:

The null solution of the periodic hybrid system (piecewise constant periodic system) (1) is uniformly asymptotically stable iff the eigenvalues of the matrix

$$\phi(t_0+T, t_0) = \prod_{i=N}^1 \text{Exp}(A_i(\delta t_i)) \quad (2)$$

have magnitudes less than 1. It is unstable if at least one eigenvalue of this matrix has magnitude greater than 1.

Basically what the theorem is saying is that for the system to be asymptotically stable its transition matrix over one period of time has to be a contraction.

It is obvious how to modify this theorem to derive a similar one for hybrid systems which are not necessarily periodic. However the theorem will be difficult to use, since one has to compute matrices similar to (2) $N!$ times 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. The result is a simpler condition that is only sufficient.

The logarithmic norm (also known as the logarithmic derivative, the measure of a matrix) was introduced in 1958 separately by Dahlquist [6] and Lozinskij [7] 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

$$\mu(A) = \lim_{h \rightarrow 0^+} (\|I + hA\| - 1)/h. \quad (3.1)$$

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

$$\mu(A) = \max\{\mu : \mu \in \lambda((A+A^*)/2)\}. \quad (3.2)$$

Then the following inequality is true:

$$\|\text{Exp}(At)\| \leq \text{Exp}(\mu(A)t) \quad (4)$$

One very important property of the logarithmic norm follows from the fact that it may be shown to be the smallest element of

$$S = \{x : \|\text{Exp}(At)\| \leq \text{Exp}(xt), t \geq 0\}. \quad (5)$$

Therefore it gives an optimal bound on the exponential behavior of $\|\text{Exp}(At)\|$ for $t \geq 0$. It may be concluded immediately that

$$\sup_{t \geq 0} \|\text{Exp}(At)\| = 1 \text{ iff } \mu(A) \leq 0. \quad (6)$$

In the case where A is normal square matrix (i.e., $A^* = A$), then

$$\|\text{Exp}(At)\| = \text{Exp}(\alpha(A)t) = \text{Exp}(\mu(A)t), \quad (7)$$

where $\alpha(A)$ is the spectral radius, i.e. the maximal real part of the eigenvalues of A .

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

THEOREM 1:

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

$$\sum_i \mu(A_i) \delta t_i < 0, \quad \delta t_i \equiv t_i - t_{i-1}, i \in \mathbb{N}. \quad (8)$$

The proof is a simple application of the logarithmic norm to Willems's theorem. It is important to know that the above theorem is stated not only for periodic hybrid systems but it applies to the more general 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 because 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 $t_f \geq t_0 + T$ 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 iff the

observability matrix

$$\begin{bmatrix} O_1 \\ O_2 \text{Exp}(A_1(\delta t_1)) \\ \vdots \\ O_N \text{Exp}(A_{N-1} \delta t_{N-1}) \dots \text{Exp}(A_1(\delta t_1)) \end{bmatrix} \quad (9)$$

has full rank, where O_i is the observability of the i th form, $i \in N$.

Proof:

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

$$y(t) = C_i \text{Exp}(A_i(t-t_i)) \prod_{j=i-1}^1 \text{Exp}(A_j(\delta t_j)) x(t_0) \quad (10)$$

taking $n-1$ derivatives of $y(t)$ and arranging them in a matrix yields

$$Y_i(t) = O_i \text{Exp}(A_i(t-t_i)) \prod_{j=i-1}^1 \text{Exp}(A_j(\delta t_j)) x(t_0) \quad (11)$$

where O_i is the observability matrix of the i th form. Now repeating the same procedure for all $i \in N$ and stacking the Y_i 's starting by Y_1 gives the following equation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} O_1 \text{Exp}(A_1(t-t_0)) \\ O_2 \text{Exp}(A_2(t-t_1)) \text{Exp}(A_1 \delta t_1) \\ \vdots \\ O_N \text{Exp}(A_N(t-t_{N-1})) \dots \text{Exp}(A_1 \delta t_1) \end{bmatrix} x(t_0). \quad (12)$$

Before proceeding any further we would like to note that we have a "free" parameter t for every Y_i . The t parameter is free because we can choose it any way we want in the appropriate interval. As it will be seen in this proof picking $t=t_k$, $k=0, 1, \dots, N-1$, for every Y_k yields a simpler observability/controllability criteria. That is, less computation is needed to apply the test. Nevertheless, picking it otherwise will be of use as discussed in the sequel.

For notational convenience (12) can be written in the following compact way

$$Y = O x(0) \quad (13)$$

and the question is whether we can find $x(0)$. Now if the $N \times n$ matrix O has rank less than n , then there exists a linear combination of the matrix O columns adding to zero, i.e.

$$0 = O x = \sum_{i=1}^n x_i \text{col}_i O. \quad (14)$$

Therefore, the condition that O has full rank is necessary for observability.

To prove that it is sufficient we first multiply (13) on both sides by O' to obtain

$$O'Y = O'O x(0). \quad (15)$$

Now if O has rank n , it is equivalent [Kailath] to say that $O'O$ is nonsingular, which means that $x(0)$ can be directly obtained from (15) as

$$x(0) = (O'O)^{-1}O'Y. \quad (16)$$

This is the only solution of (15). Moreover, it is also the only solution of (13). To wit assume that x_1 and x_2 are two different solutions of (13), then we shall have

$$O[x_1 - x_2] = 0$$

which means that some combination of the columns of O is zero, which contradicts the assumption that O has full rank. Therefore the theorem is proved.

4. CONTROLLABILITY:

At this point the dual algebraic controllability test is introduced. First an analog 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 any new 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. Computing the controllability grammian and using the fact that our system is piece-wise constant yields the following theorem.

THEOREM 3:

A periodic hybrid system of N forms is controllable iff

$$W(t_0, t_0+T) = \sum_{i=1}^N \int_{t_{i-1}}^{t_i} \Phi_i(t, t_0) B_i B_i^T \Phi_i^T(t, t_0) dt \quad (17)$$

has full rank.

COROLLARY:

A periodic hybrid system is completely controllable iff it is controllable.

Proof:

See Remark (2.18) in [14], then use the above theorem.

Before proceeding any further, a necessary and sufficient condition for a

periodic hybrid system to be uniformly completely controllable is given. This result will be of importance when stabilizability of such systems is in question.

THEOREM 4:

A periodic hybrid system is uniformly completely controllable iff it is completely controllable.

Proof:

If the periodic system is completely controllable, there must exist a finite $s \geq T$ such that $W(0,s) \geq \epsilon I > 0$. Therefore the result is proved by using a result from Silverman et al (Lemma 1) [10] and Remark (2.18) in Kalman et al [14].

Having used the usual test we are ready to present an algebraic controllability test very similar to the one used in linear time-invariant theory. The following criterion applies for periodic hybrid systems. An analog criterion for hybrid systems will be introduced later in this paper.

THEOREM 5:

A periodic hybrid system of N forms is controllable iff the controllability matrix

$$[C_N, \text{Exp}(A_N(\delta t_N))C_{N-1}, \dots, \text{Exp}(A_N(\delta t_{N-1}))\dots\text{Exp}(A_2(\delta t_2))C_1] \quad (18)$$

has full rank, where C_i is the usual controllability matrix of the i th form, $i \in N$.

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

$$[C_N, \text{Exp}(A_N(\delta t_N))\{C_{N-1}, \dots\{C_4, \text{Exp}(A_3(\delta t_3))\{C_2, \text{Exp}(A_2(\delta t_2))C_1\}\}\}. \quad (19)$$

This way one does not have to compute all of the matrices needed to express (18) and compute its rank. That is the rank is checked sequentially and (19) is augmented appropriately until full rank is achieved. If not, the system is not controllable. The same observation applies to the observability criterion.

Besides the above algebraic criteria for controllability and observability, we are ready to introduce two more tests. The first test is a very 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

$$\text{rank}[C_1, C_2, \dots, C_N] \equiv \text{rank } \tilde{C} = n. \quad (20)$$

Where C_i is the controllability matrix for the i th form, $i \in N$.

Proof:

We write for the state at s , when at time t_0 the system is in zero state,

$$x(s) = \int_{t_0}^s \Phi(s, \tau) B(\tau) d\tau. \quad (21)$$

Using the fact that the system is piece-wise constant and the linearity property of the integral operator yields for $s=t_N$

$$\begin{aligned} x(t_N) = & \text{Exp}(A_N \delta t_N) \dots \text{Exp}(A_2 \delta t_2) \int_{t_0}^{t_1} \text{Exp}(A_1(t_1 - \tau)) B_1 u(\tau) d\tau + \dots \\ & + \text{Exp}(A_N \delta t_N) \int_{t_{N-2}}^{t_{N-1}} \text{Exp}(A_{N-1}(t_{N-1} - \tau)) B_{N-1} u(\tau) d\tau \\ & + \int_{t_{N-1}}^{t_N} \text{Exp}(A_N(t_N - \tau)) B_N u(\tau) d\tau. \end{aligned} \quad (22)$$

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 $\tilde{C}$ given in the theorem. Moreover, it is easy to see that

$$\text{rank } \tilde{C} \leq \text{rank } C \leq n$$

an inequality that dictates that full rankness of $\tilde{C}$ 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 if we did not lose it 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 way 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

$$\text{rank}[B_1, \text{Exp}(A_1 \delta t_1)B_2, \dots, \text{Exp}(A_N \delta t_N) \dots \text{Exp}(A_1 \delta t_1)B_N] \equiv \text{rank } \tilde{C} = n. \quad (23)$$

Proof:

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

$$\hat{C}(s_1, s_2, \dots, s_N) \hat{C}'(s_1, s_2, \dots, s_N) = \sum_{k=1}^N \Phi(s_k, t_0) B_k B_k' \Phi'(s_k, t_0). \quad (24)$$

Where $s_k \in [t_k, t_{k-1}]$. Then

$$\begin{aligned} W(t_0, t_N) &= \int_{t_0}^{t_N} \Phi(s, t_0) B(s) B'(s) \Phi(s, t_0) ds \\ &\geq \sum_{k=1}^N \int_{s_k}^{s_k + \sigma} \Phi(s, t_0) B(s) B'(s) \Phi(s, t_0) ds \\ &= \sigma \hat{C}(s_1, s_2, \dots, s_N, t_0) \hat{C}'(s_1, s_2, \dots, s_N, t_0) + o(\sigma), \end{aligned} \quad (25)$$

for σ sufficiently small.

If we assume that $\hat{C}$ has full rank then for σ small enough (25) is positive definite. But then (25) implies that $W(t_N, t_0) > 0$ and the theorem is proved.

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 the above results apply to hybrid systems (i.e., not only the periodic ones) without modification. Therefore we will state the most important results and leave the rest to the interested reader.

THEOREM 8:

A hybrid system is controllable iff theorem 5 holds for all possible $N!$ permutations of the form-index set N .

THEOREM 9:

In case theorem 7 holds for the $N!$ permutations of the index-set N then the hybrid system is controllable.

It is obvious that theorem 6 applies for general hybrid systems too. Moreover, we think that the theorem is very "close" to being a sufficient condition too. 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 drastically. That is if, for example, C_1 and C_2 have algebraic complementary range spaces (i.e, $\text{range}(C_1)$ is perpendicular to $\text{range}(C_2)$) then $\text{range}(\text{Exp}(AT)C_1)$ will almost always remain an algebraic complement but not necessarily perpendicular to $\text{range}(C_2)$. As a matter of fact, Mariton [12] 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 off-the-shelf techniques to control/stabilize hybrid systems.

6.1 Stabilizability:

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 $\hat{\Phi}(.,.)$ of the closed loop system can be made such that $\|\hat{\Phi}(t,t_0)\| \leq a(t_0)\text{Exp}\{-\delta(t,t_0)\}$ 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 $\|\hat{\Phi}(t_2,t_1)\| \leq a\text{Exp}\{-m(t_2-t_1)\}$ 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.

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 recipe 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.

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 iff there exists two positive constant numbers c_1 and c_2 such that

$$\|\Phi(t, t_0)\| \leq c_1 \text{Exp}(-c_2(t-t_0)) \quad (26)$$

for all $t \geq 0$. Therefore using theorem 1 leads to the following design criterion

$$\sum_i \mu(A_i) \delta t_i \leq k_1 - k_2 T \quad (27)$$

where $k_1 = \ln(c_1)$ and T the period of the hybrid system. The k_i 's, $i=1, 2$, are the design parameters that the designer chooses according to his specifications to make the upper bound of the transition matrix of the system and consequently the time response of the plant to be as desired. This is possible because every form is controllable, therefore the closed-loop poles of each form can be assigned arbitrarily. Consequently, (27) can be always obtained via state feedback since every form is observable. It is important to note that this design procedure applies to periodic hybrid systems and hybrid systems as well. It is encouraging to remember that the minimality condition for every form is not necessary to achieve such design.

Before closing this section, we would like to mention another way to control uniformly completely controllable hybrid systems. This technique² is due to Kalman [9]. Kalman showed that by using the mathematical concept of the generalized inverse of a matrix introduced by Penrose it is possible to define a suitable control that will accomplish the desired transfer. Moreover, he was

² It is interesting to note that this technique never made it into standard optimal control text books.

able to prove that the proposed control is the minimum energy control required to accomplish the transfer.

6.2 Example:

The example is a 2-form periodic hybrid system. Form number one and two are respectively described with the following dynamics

$$\begin{aligned}\Sigma_1: \quad dx_1/dt &= x_1 + u \\ dx_2/dt &= x_2,\end{aligned}$$

$$\begin{aligned}\Sigma_2: \quad dx_1/dt &= x_1 \\ dx_2/dt &= x_2 + u.\end{aligned}$$

It is clear that both forms are unstable. Since both forms are diagonal their transition matrices are simple to compute and the transition matrix of the hybrid system is found to be unstable too.

Our goal is to stabilize this hybrid system. It is obvious that both forms are uncontrollable but the hybrid system is controllable. The controllability of the system is easily checked by any of the controllability tests introduced in this paper.

To stabilize the system we use Kalman's technique [14] to control the controllable subspace of each form. Starting at time zero Σ_1 is on and remains so for T_1 time units. It is easy to check that

$$u(t) = 2\{\text{Exp}(-t)\}/\{1-\text{Exp}(-2T_1)\}x_1(0), \quad t \in [0, T_1]$$

is an optimum open-loop control action that will take $x_1(0)$ to zero in T_1 time units. A similar control action will take $x_2(T_1)$ to zero in T_2 time units.

Therefore, steering the system to the origine was accomplished in one period. This control action resembles "dead-beat" control.

7. CONCLUSION:

In this paper a special class of linear piece-wise constant time-varying systems was introduced. These systems are called hybrid systems because the set of linear time-invariant systems among which the systems are switching is finite.

Because hybrid systems share several features with linear time-invariant systems it was possible to derive the following results: 1. A necessary and sufficient stability condition and a simple sufficient criterion. 2. Algebraic necessary and sufficient controllability/observability tests analog to the usual tests. 3. A very interesting necessary controllability/observability condition which is "almost" sufficient along with a simple sufficient condition.

The necessary controllability/observability condition is a flat block matrix composed by 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 research problem.

State feedback via switching or nonswitching gains is an interesting topic that needs investigation. Nonswitching gains are very useful since they eliminate the need for form-detection.

Much more has to be done 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 $\Sigma_i = (A_i, B_i, C_i)$ 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 [4]. Viewing a hybrid system as an iterative process sheds some light on some complicated issues such as the stability of such systems.

Moreover, if we discretize our continuous-time hybrid system with samples happening at the discontinuities we come up with what we call the induced discrete-time hybrid system. Using the induced discrete-time model one can use the discrete-time LQ-theory to control/stabilize such systems. This remark implies that we probably only need to study discrete-time hybrid systems. At this point we would like to mention the work of Ludyk [15] where the author tries to solve the problem of eigenvalue assignment for time-varying discrete-time systems following the path of Wollovich [16]. Applying these techniques might give us more understanding of the control/stabilization of hybrid systems.

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

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] J. N. TSITSIKLIS, "On the Stability of Asynchronous Iterative Processes," Pro. 25th. IEEE Conf. Decision and Control, Athen, Greece, pp. 1617-1621, dec. 1986.
- [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 Info. and System Sci. Series, T. Kailath Ed., Prentice-Hall, 1980.
- [9] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, Controllability of linear dynamical systems, Contribution to Differential Equations, Vol.1, No.2, 1962, pp. 189-213.
- [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. MARITON, "Controllability, Stability and Pole Allocation for Jump Linear Systems," Proc. 25th. IEEE Conf. Decision and Control, Athens,

Greece, pp. 2193-2194, Dec. 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] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, Topics in Mathematical System Theory, Mc Graw-Hill, New York, 1969.
- [15] G. LUDYK, Time-variant discrete-time systems, Advances in Control Systems and Signal Processing, Ed. I. HARTMAN, Vol.3, Braunschweig, Wiesbaden, 1982.
- [16] W. A. WOLLOVICH, "On the stabilization of controllable systems," IEEE Trans. on Aut. Cont., Vol. AC-13, pp. 569-572, 1968.

APPENDIX Z

OPTIMAL AND SUBOPTIMAL FILTERING FOR LINEAR SYSTEMS DRIVEN BY
SELF-EXCITED POISSON PROCESSES

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:

$$dx_t = a_t x_t dt + b_t dn_t \quad (1)$$

$$y_t = \frac{dz_t}{dt} = c_t x_t + \frac{dw_t}{dt} \quad (2)$$

where n_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

of η_t is a memoryless function of the state, $\mu(x_t)$. The process w_t is a Brownian motion with diffusion V_t .

The objective is to estimate x_t given the history of the observation process, either y_s or z_s , for $s < t$. In Section 2, an expression for the minimum mean-squared error (MMSE) estimate is derived, and shown to be impractical. Good suboptimal approximations to the MMSE estimate are desirable, 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 $Z_t \triangleq \{z_s; s < t\}$, 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 , $t > t_0$, be the semi-martingale defined by

$$dQ_t = R_t dt + dM_t \quad t > t_0 \quad (3)$$

where M_t is a martingale with respect to a growing family of σ -fields F_t , $t > t_0$, and where R_t is a process adapted to F . Let z_t , $t > t_0$, be the semi-martingale process

$$dz_t = h_t dt + dw_t \quad t > t_0 \quad (4)$$

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 $t > t_0$. Define Z_t as the growing family of σ -fields generated by the process z_t . For an arbitrary process ξ_t , define $\hat{\xi}_t \triangleq E(\xi_t | Z_t)$. Then $\hat{Q}_t$ satisfies the dynamic equation

$$d\hat{Q}_t = \hat{R}_t dt + [\hat{Q}_t \hat{h}_t - \hat{Q}_t \hat{h}_t] V_t^{-1} [dz_t - \hat{h}_t dt] \quad (5)$$

The filtering theorem will be applied to $Q_t = e^{ivx_t}$, 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].

Differential Rule [2]: For an appropriately smooth function $Q(x_t)$ and for x_t defined in (1), the rule is

$$dQ(x_t) = a_t x_t \left(\frac{\partial Q(x_t)}{\partial x_t} \right) dt + \int_U [Q(x_t + b_t u) - Q(x_t)] K(dt, du) \quad (6)$$

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 $A \subseteq U$.

Equation (6) may be put in the form of (3) by letting

$$dM_t = \int_U [Q(x_t + b_t u) - Q(x_t)] [K(dt, du) - \mu(x_t) p_U(u) dt du] \quad (7)$$

and taking $R_t dt$ as the remainder. The substitution of R_t into (5) yields

$$\begin{aligned} d\widehat{e^{ivx_t}} &= \widehat{(iva_t x_t e^{ivx_t} + e^{ivx_t} [e^{ivb_t u} - 1] \mu(x_t))} dt \\ &\quad + \widehat{[c_t x_t e^{ivx_t} - c_t x_t e^{jvx_t}] V_t^{-1} [dz_t - c_t x_t dt]} . \end{aligned} \quad (8)$$

Let $\theta_t = b_t u$ (recall u is the mark variable) and $p_{\theta}(\cdot)$ 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

$$dp_{t|t}(x) = Lp_{t|t}(x) dt + V_t^{-1} c_t [x - \hat{x}_t] p_{t|t}(x) [dz_t - c_t \hat{x}_t dt] \quad (9)$$

where L is the linear operator given by

$$Lp(x) = -\frac{\partial}{\partial x} [a_t x p(x)] + (p_{\theta_t}(x) * [\mu(x) p(x)]) - \mu(x) p(x) \quad (10)$$

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:

$$P_{n,t} = E[(x_t - \hat{x}_t)^n | Z_t] \quad n = 1, 2, \dots \quad (11)$$

$$d\hat{x}_t = a_t \hat{x}_t dt + b_t E(u) \widehat{\mu(x_t)} dt + v_t^{-1} c_t P_{2,t} [dz_t - c_t \hat{x}_t dt]$$

$$\begin{aligned} dP_{n,t} = & na_t P_{n,t} dt + \sum_{k=1}^n \binom{n}{k} b_t^k E(u^k) \widehat{(x_t - \hat{x}_t)^{n-k} \mu(x_t)} dt - nb_t E(u) \widehat{\mu(x_t)} P_{n-1,t} dt \\ & + v_t^{-1} c_t [P_{n+1,t} - n P_{2,t} P_{n-1,t}] [dz_t - c_t \hat{x}_t dt] \\ & + n v_t^{-1} c_t^2 P_{2,t} \left[\frac{n-1}{2} P_{2,t} P_{n-2,t} - P_{n,t} \right] dt \quad n = 2, 3, \dots \end{aligned} \quad (12)$$

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 η_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 η_t and the jump times $\tau_{N_T} = [\tau_1, \tau_2, \dots, \tau_{N_T}]$ 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 $\tilde{\tau}_M$ satisfy

$$(\tilde{N}_T, \tilde{\tau}_M) = \arg \left(\begin{array}{c} \text{Max} \\ 0 < N^* < M \\ \tau_M^* \in R_+^M \end{array} \left[\ln P_{N_T, \tau_M} | Y_T, N_T < M(N^*, \tau_M^*) \right] \right) \quad (13)$$

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 τ_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:

$$\begin{aligned} & \ln \frac{1}{2V_T} \int_0^T [2y_t - \sum_{i=0}^{N^*} h(t, \tau_i^*)] [\sum_{j=0}^{N^*} h(t, \tau_j^*)] dt \\ & + \ln P_{\tau_M | N_T, N_T < M}(\tau_M^* | N_T = N^*, N_T < M) + \ln \Pr[N_T = N^* | N_T < M] \end{aligned} \quad (14)$$

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, $h(t, \tau_j^*)$ is defined as the unforced response due to a known initial condition x_0 .

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

$$\begin{aligned} & P_{\tau_M | N_T, N_T < M}(\tau_M^* | N_T = N^*, N_T < M) = \\ & \left\{ \begin{array}{ll} \frac{P_{\tau_M | N_T < M}(\tau_M^* | N_T < M)}{\Pr[\tau_{N^*} < T, \tau_{N^*+1} > T | N^* < M]} & \text{for } 0 < \tau_1^* < \dots < \tau_{N^*}^* < T \\ & \text{and } T < \tau_{N^*+1}^* < \tau_M^* \\ 0 & \text{otherwise} \end{array} \right. \quad (15) \end{aligned}$$

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:

$$p_{\tau_M | N_T < M}(\tau_M^* | N_T < M) = \int_T^\infty p_{\tau_{M+1} | \tau_{M+1}^* > T}(\tau_{M+1}^* | \tau_{M+1} > T) d\tau_{M+1}^* \quad (16)$$

It is noted that the region of support of the integrand is over the "wedge" $0 < \tau_1 < \dots < \tau_M < \tau_{M+1}$ minus the half space $\tau_{M+1} < T$. Therefore, (16) can be rewritten as:

$$p_{\tau_M | N_T < M}(\tau_M^* | N_T < M) = \int_{\text{Max}(\tau_M^*, T)}^\infty \frac{p_{\tau_{M+1}}(\tau_{M+1}^*)}{\text{Pr}(\tau_{M+1} > T)} d\tau_{M+1}^* \quad (17)$$

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:

$$p_{\tau_{M+1}}(\tau_{M+1}^*) = \begin{cases} \prod_{i=1}^{M+1} \frac{\partial}{\partial \tau_i^*} \exp \int_{\tau_{i-1}^*}^{\tau_i^*} -\mu[\bar{x}_t(\tau_{i-1}^*)] dt & \text{for } 0 < \tau_1^* < \dots < \tau_{M+1}^* \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

with

$$\bar{x}_t(\tau_{i-1}^*) = x_0 e^{at} u_1(t) + \sum_{j=1}^{i-1} b \exp[a(t-\tau_j^*)] u_1(t - \tau_j^*)$$

where u_1 is the unit step function. In words, $\bar{x}_t(\tau_{i-1}^*)$ is the value of the state assuming that η_t has had jumps only at times $\tau_1^*, \dots, \tau_{i-1}^*$. Let $\bar{x}_t(\tau_0^*)$ denote the unforced value of the state.

Substitution of (18) into (17) is straightforward due to the product form of (18) and yields

$$p_{\tau_M | N_T < M}(\tau_M^* | N_T < M) = \frac{p_{\tau_M}(\tau_M^*)}{\text{Pr}(\tau_M > T)} \exp \int_{\tau_M^*}^{\text{Max}(\tau_M^*, T)} -\mu[\bar{x}_t(\tau_M^*)] dt \quad (19)$$

where it has been assumed that there exists some $\alpha > 0$ such that $\mu[x] > \alpha$ for all x , thus making

$$\exp \int_{\tau_M^*}^\infty -\mu[\bar{x}_t] dt = 0.$$

It is noted that $p_{\tau_M}(\cdot)$ is defined by replacing $M + 1$ by M in (18). Evaluation of the derivatives yields:

$$p_{\tau_M}(\tau_M^*) = \begin{cases} \prod_{i=1}^M \mu[\bar{x}_{\tau_i^*}(\tau_{i-1}^*)] \exp \int_{\tau_{i-1}^*}^{\tau_i^*} -\mu[\bar{x}_t(\tau_{i-1}^*)] dt & \text{for } 0 < \tau_1^* < \dots < \tau_M^* \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

The combination of equations (14), (15), (19), and (20) results in a new MAP equation:

$$(\tilde{N}_T, \tilde{\tau}_M) = \arg \left\{ \begin{array}{l} \text{Max} \\ 0 < N^* < M \left[\begin{array}{l} \text{Max} \\ \tau_1^* < \dots < \tau_{N^*}^* < T \\ T < \tau_{N^*+1}^* < \dots < \tau_M^* \end{array} \right. \left. \begin{array}{l} \frac{1}{2V_T} \int_0^T [2y_t - \bar{x}_t(\tau_M^*)][\bar{x}_t(\tau_M^*)] dt \\ + \ln \left(\prod_{i=1}^M \mu[\bar{x}_{\tau_i^*}(\tau_{i-1}^*)] \right) - \int_0^{\text{Max}(\tau_M^*, T)} \mu[\bar{x}_t(\tau_M^*)] dt \end{array} \right] \end{array} \right\} \quad (21)$$

where the maximization is to be performed in two steps, first over the τ_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 τ_L as $[\tau_{N_F+1}^*, \dots, \tau_{N_F+L}^*]$, and redefine $\bar{x}_t(\tau_L^*)$ as the state assuming that jumps have occurred only at the finalized times and at the proposed times τ_L^* . The modified (approximate) version of (21) is:

$$(\tilde{N}_{K\Delta}, \tilde{\tau}_L) = \arg \left\{ \begin{array}{l} \text{Max} \\ \tilde{N}_F < N^* < \tilde{N}_F + L \end{array} \left[\begin{array}{l} \text{Max} \\ \tau_{N_F}^* < \tau_{N_F+1}^* < \dots < \tau_{N^*}^* < K\Delta \\ K\Delta < \tau_{N^*+1}^* < \dots < \tau_{N_F+L}^* \end{array} \right. \right. \\ \left. \left. \frac{1}{2V_T} \int_{(K-L)\Delta}^{K\Delta} [2y_t - \bar{x}(\tau_L^*)][\bar{x}(\tau_L^*)] dt \right. \right. \\ \left. \left. + \ln \left(\int_{i=\tilde{N}_F+1}^{\tilde{N}_F+L} \mu[\bar{x}_{\tau_i^*}(\tau_{i-1}^*)] \right) - \int_{(K-L)\Delta}^{\text{Max}(\tau_{N_F+L}^*, K\Delta)} \mu[\bar{x}_t(\tau_L^*)] dt \right] \right\} \quad (22)$$

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 $\tau_{N_F}^*$ 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 ($\mu[x_t] = \mu_0$) presents no difficulty, because the joint a priori density function for the jump times after $K\Delta$ has its maximum at $\tau_{N^*+1}^* = \tau_{N^*+2}^* = \dots = \tau_{N_F+L}^* = K\Delta$. 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 $\Pr[\eta_{t+\Delta} - \eta_t > 1]$ 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 $t \approx 2$ 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.

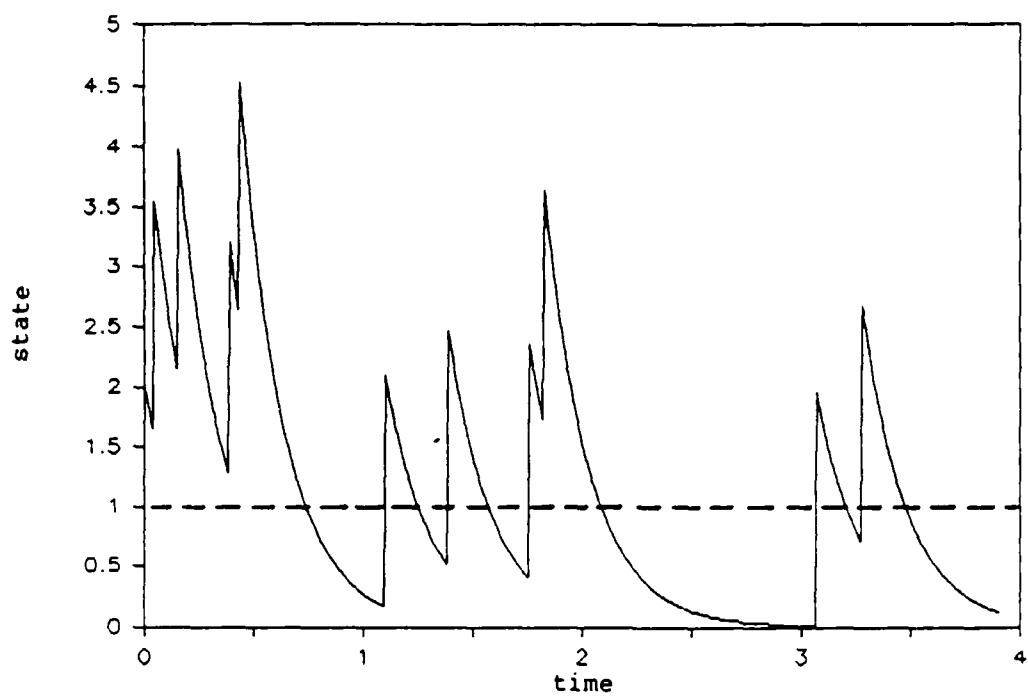


Figure 1. Example state trajectory

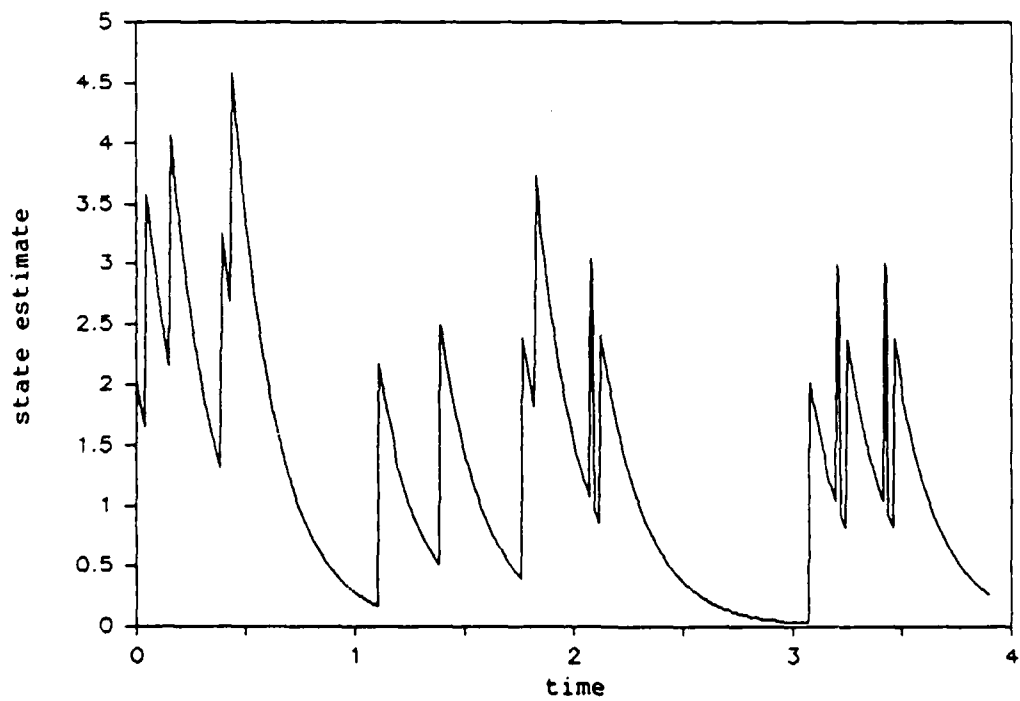


Figure 2. Estimate output

END

DATE

FILMED

8-88

DTIC