Lecture Notes in
Control and
Information Sciences

Edited by M. Thoma and A. Wyner

IIASA 105

Ch. I. Byrnes
A. Kurzhanski (Eds.)

Modelling and Adaptive Control

Proceedings of the IIASA Conference
Sopron, Hungary, July 1986

Springer-Verlag

Lecture Notes in
Control and
Information Sciences

Edited by M. Thoma and A. Wyner

IIASA 105

Ch. I. Byrnes
A. Kurzhanski (Eds.)

Modelling and Adaptive Control

Proceedings of the IIASA Conference
Sopron, Hungary, July 1986

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo

Series Editors
M. Thoma · A. Wyner

Advisory Board
L. D. Davisson · A. G. J. MacFarlane · H. Kwakernaak
J. L. Massey · Ya Z. Tsypkin · A. J. Viterbi

Editors
Christopher Ian Byrnes
Dept. of Electrical and Computer Eng.
Arizona State University
Tempe, Arizona 85287
USA

Alexander B. Kurzhanski
System and Decision Sciences Program
IIASA
2361 Laxenburg
Austria

ISBN 3-540-19019-8 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-19019-8 Springer-Verlag New York Berlin Heidelberg

Library of Congress Cataloging in Publication Data
Modelling and adaptive control.
(Lecture notes in control and information sciences ; 105)
1. Adaptive control systems--Mathematical models--Congresses.
I. Byrnes, Christopher I., II. Kurzhanski, A. B.
III. International Institute for Applied Systems Analysis.
IV. Series.
TJ217.M55 1988 629.8'36 88-4653
ISBN 0-387-19019 (U.S.)

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law of September 9, 1965, in its version of June 24, 1985, and a copyright fee must always be paid. Violations fall under the prosecution act of the German Copyright Law.

© Springer-Verlag Berlin, Heidelberg 1988
Printed in Germany

Offsetprinting: Mercedes-Druck, Berlin
Binding: B. Helm, Berlin
2161/3020-543210

PREFACE

System modelling and the control of dynamical systems in the face of uncertainty are two important topics in the study of system dynamics, which is currently a major component of the research program in the Department of Systems and Decision Sciences at the International Institute for Applied Systems Analysis.

In July 1986 an SDS IIASA workshop on Modelling and Adaptive Control at Sopron, Hungary, was attended by prominent control theorists and practitioners from both the East and West. One of the main purposes of this workshop was to give an overview of both traditional and recent approaches to the twin theories of modelling and control which ultimately must incorporate some degree of uncertainty. The broad spectrum of processes for which solutions of some of these problems were proposed was itself a testament to the vitality of research on these fundamental issues. In particular, these proceedings contain new methods for the modelling and control of discrete event systems, linear systems, nonlinear dynamics and stochastic processes.

We believe that this workshop has also achieved one of the goals at IIASA, which is to promote and encourage cooperation between the scientists of East and West.

It is our pleasure to thank Harold Kushner, George Leitman and Pravin Varaiya for helping us organize this workshop as well as the indispensable support provided by the Hungarian National Member Organization to IIASA.

C. I. Byrnes
Dept. of Electrical and Computer
Engineering
Arizona State University

A. B. Kurzhanski
Chairman
Systems and Decision Sciences
International Institute for
Applied Systems Analysis

CONTENTS

Asymptotically Efficient Rules in Multiarmed Bandit Problems

V. Anantharam and P. Varaiya
Department of Electrical Engineering and Computer Sciences
and Electronics Research Laboratory
University of California, Berkeley CA 94720

ABSTRACT

Variations of the multiarmed bandit problem are introduced and a sequence of results leading to the work of Lai and Robbins and its extentions is summarized. The guiding concern is to determine the optimal tradeoff between taking actions that maximize immediate rewards based on current information about unknown system parameters and making experiments that may reduce immediate rewards but improve parameter estimates.

1. Setup

We begin with an abstract description and then give two examples. We are given N discrete-time real-valued stochastic processes

$$X^1 \colon X^1(1), X^1(2), \cdots$$

$$\cdots$$

$$X^N \colon X^N(1), X^N(2), \cdots .$$

The essential assumption is that these processes are *independent*. For historical reasons these processes are also called *arms* or *jobs*.

A fixed number m, $1 \le m \le N$, is specified. At each time t we must select m different arms. Let $T^j(t)$ be the number of times that arm j was selected during the interval $1, \cdots, t$; and let $U(t) \subset \{1, \cdots, N\}$ be the m arms that are selected at time t. Then at time t we receive the reward

$$Y(t) = \sum_{j \in U(t)} X^j[T^j(t)], \tag{1}$$

and the information available before making the next decision at time $t+1$ is given by the

σ-field

$$I(t) = \sigma\{X^j(s) \mid s = 1, \cdots, T^j(t);\ j = 1, \cdots, N\}.$$

Our aim is to select the arms so as to maximize the sequence of rewards. Because the rewards are random we are never sure which arm gives maximum rewards. Hence in selecting the arms we must consider both short term and long term gain. Short term considerations lead us to select those arms which yield large *expected* immediate rewards, but then we won't learn anything more about the arms we did not select. Therefore, we may select arms with lower expected immediate reward simply to obtain better information about those arms. In other words, in our selection we must weigh making immediate rewards against gaining information that will be valuable in the future.

We consider two criteria for evaluating different selection rules. The ***discounted reward*** criterion associates to a rule Φ the number

$$J_\beta(\Phi) = \sum_{t=1}^{\infty} \beta^t EY(t), \tag{2}$$

and the ***average reward*** criterion associates the number

$$J(\Phi) = \liminf \frac{1}{N} \sum_{t=1}^{N} EY(t). \tag{3}$$

In (2) $0 < \beta < 1$ is a fixed discount factor. The optimal selection rule will incorporate the best tradeoff between immediate and future gains referred to above. How this is done is discussed later after we consider two examples.

2. Examples

The first example is the original bandit problem. Each arm X^j is a sequence of iid variables $X^j(1), X^j(2), \cdots$ with probability density $f(x, \vartheta_j)\nu(dx)$ relative to some common measure ν on R. Let $\mu(\vartheta) = \int x f(x, \vartheta)\, \nu(dx)$ be the mean. The parameters ϑ_j that characterize the arms are not known. However, if $\vartheta_1, \cdots, \vartheta_N$ were known ahead of time, then the best rule (for both (2), (3)) is always to select the m arms with largest mean values $\mu(\vartheta_j)$.

At each time t the mean value $\mu(\vartheta_j)$ can be estimated for example by its sample mean

$$\mu^j(t) = \frac{1}{T^j(t)} \sum_{s=1}^{T^j(t)} X^j(s).$$

If arm j is not selected at time t, $T^j(t)$ is unchanged, and so its sample mean stays the same. So a good rule must balance selecting arms with larger sample means against arms with lower sample means in order to improve the estimates of their means.

In the second example we are given N jobs which must be processed on m machines. The jth job requires a random amount Q^j of processing time and it costs c_j per unit time that the job is not finished. The problem is to schedule the N jobs on the m machines to minimize the total waiting cost. This can be cast in our abstract form by associating to the

jth job the 'reward' process

$$X^j(t) = -c_j\, 1(t \leq Q^j),$$

where $1(\cdot)$ is the indicator function. If a scheduling rule Φ finishes job j at the (random) time S^j, say, then the expected discounted waiting cost is

$$\sum_{j=1}^{N} c_j E \sum_{t=1}^{S^j} \beta^t .$$

To minimize the cost one wants to select jobs with large c_j and small Q^j. Q^j is not known in advance, but can be estimated using the fact that if j is not finished by time t then we observe the event $\{Q^j > T^j(t)\}$.

Many applications of bandit problems are discussed in Gittins [2]. For an application in microeconomics see Weitzmann [8]. In these references $m = 1$ for which a fairly complete result is now available as we discuss next.

3. The case $m = 1$

For $m = 1$ and the discounted reward criterion there is a striking result due to Gittins and Jones [3]. We present it in the more general form given by Varaiya, Walrand and Buyukkoc [7].

Fix an arm X (we omit the superscript j):

$$X: X(1), X(2), \cdots$$

and let $F^X(s) = \sigma\{X(1), \cdots, X(s)\}$ be the information available after this arm is selected s times. The *Gittins index* of X at time s is defined as

$$\gamma(s) = \sup_{\tau > s+1} \frac{E\{\sum_{t=s+1}^{\tau} \beta^t X(t) \mid F^X(s)\}}{E\{\sum_{t=s+1}^{\tau} \beta^t \mid F^X(s)\}}, \tag{4}$$

where τ ranges over all stopping times of the family $\{F^X(s)\}$. γ can be interpreted as the maximum rate of expected discounted reward per unit expected discounted time.

Now consider the problem of maximizing (2). By the *index rule* we mean the following procedure for selecting arms: At each time t calculate the current Gittins index of each arm j, namely, $\gamma^j(T^j(t))$, and at $t+1$ select the arm with the largest current index.

Theorem 1. The index rule is optimal.

The index of an arm does not depend on the other arms. That is what makes this result important: it converts an N-dimensional problem into N one-dimensional problems.

The index $\gamma(s)$ of arm X in (4) summarizes the optimal tradeoff between selecting X for immediate rewards versus selecting it for information that will be valuable in the future. Note that in (4), τ is a random stopping time, so that it allows us to continue or to stop selecting X based on what we learn. For example, consider the arm: $X(1) \equiv 0$ and

$$X(k) = \begin{cases} 1, \ k \geq 2, \ \text{with prob. } 0.5 \\ 0, \ k \geq 2, \ \text{with prob. } 0.5 \end{cases} \cdot$$

Then,

$$\gamma(0) = \sup_{\tau > 1} \frac{E\{\sum_1^{\tau} \beta^t X(t)\}}{E\{\sum_1^{\tau} \beta^t\}} = \frac{\beta}{2-\beta^2} \tag{5}$$

is achieved by the stopping rule

$$\tau = \begin{cases} \infty, \ \text{if } X(2) = 1 \\ 2, \ \text{if } X(2) = 0 \end{cases} \cdot$$

which selects the arm twice and then continues indefinitely or stops accordingly as $X(2) = 1$ or 0. Observe from (5) that as $\beta \to 0$ the value of knowing $X(2)$ decreases (or the cost of experimentation increases), while as $\beta \to 1$ this value increases.

Whittle [9] provides an alternative interpretation of the index by considering the two-armed bandit problem: one arm is X and the other arm, Λ, gives constant reward λ at each time instant:

$$X: X(1), X(2), \cdots \qquad \Lambda: \lambda, \lambda, \cdots$$

Clearly, if λ is very small ($\lambda \to -\infty$), it will be optimal to select X at $t = 1$, whereas if λ is very large it will be optimal to select Λ at $t = 1$. There is some intermediate value of λ at which the optimal rule is indifferent to selecting X or Λ at $t = 1$. This value of λ is in fact $\gamma(0)$ and gives an interpretation of the index as a "reservation price."

While the index rule result "solves" the bandit problem for $m = 1$, calculating the index (4) may itself be a difficult optimal stopping time problem.

It may be conjectured that the optimal rule for $m \geq 1$ is simply to select the arms with the m largest Gittins indexes. Unfortunately, this is false as can be seen from the following simple deterministic example with $N = 3$ and $m = 2$:

$$X^1: 1, 0, 0, 0 \cdots; \quad X^2: 1, 1, 0, 0 \cdots; \quad X^3: 1, 1, 1, 0 \cdots$$

Now $\gamma^1(0) = \gamma^2(0) = \gamma^3(0) = 1$ so the conjecture says that it does not matter which two arms are selected at $t = 1$. In fact the optimal strategy must select X^3 and either X^1 or X^2. The example can be strengthened by taking $X^3 = 1-\varepsilon, 1-\varepsilon, 1-\varepsilon, \cdots$ with $\varepsilon > 0$ small. Then $\gamma^3(0) = 1-\varepsilon$ is strictly smaller than $\gamma^1(0)$ and $\gamma^2(0)$, so according to the conjecture we should select X^1 and X^2 at $t = 1$ leading to the instantaneous reward sequence

$$2, 2-\varepsilon, 1-\varepsilon, 1-\varepsilon, 0 \cdots \tag{6}$$

whereas if X^3 and X^1 are selected at $t = 1$ we would have the sequence

$$2-\varepsilon, 2-\varepsilon, 1-\varepsilon, 0 \cdots \tag{7}$$

whose discounted value exceeds that of (6) by $\beta^2 - \varepsilon - \beta^3(1-\varepsilon)$ which is positive for $\beta < 1$ (and ε small). However, this difference disappears as $\beta \to 1$ suggesting that although a straightforward extension of the index rule for $m > 1$ may not be optimum for the discounted reward criterion, there may be an extension for the average reward criterion (3).

4. The average reward criterion

We return now to the bandit problem as formulated in the first example of §2. The *total* expected reward up to time t obtained by a rule Φ is

$$\sum_{j=1}^{N} \mu(\vartheta_j) E T^j(t).$$

For any parameter configuration $C = (\vartheta_1, \cdots, \vartheta_N) \in R^N$ let σ be a permutation of $\{1, \cdots, N\}$ so that

$$\mu[\vartheta_{\sigma(1)}] \geq \cdots \geq \mu[\vartheta_{\sigma(N)}]. \tag{8}$$

If C were known from the beginning, the maximum total expected reward up to t would be

$$\sum_{j=1}^{m} \mu[\vartheta_{\sigma(j)}]t,$$

so we may define the *regret* of Φ at t as

$$R(t, C, \Phi) = \sum_{j=1}^{m} \mu[\vartheta_{\sigma(j)}]t - \sum_{j=1}^{N} \mu[\vartheta_j] E T^j(t), \tag{9}$$

and we want to find a rule to

$$\underset{\Phi}{\text{Minimize}}\ R(t, C, \Phi) \quad \text{for all } t \text{ and } C. \tag{10}$$

It is evident that there will not exist Φ that minimizes the regret "for all t and C". If such Φ exists it must achieve identically zero regret because the rule that always selects a fixed set of m arms gives zero regret for configurations C for which those m arms have the largest means, but for every other configuration this rule is quite bad since it will have regret *proportional* to t. This suggests that in order to exclude such non-learning or non-adaptive rules from consideration we should modify (10) keeping "for all C" while relaxing the condition "for all t". One way of doing this is to replace (10) with an expected average reward over time in a Bayesian setting.

In such a Bayesian setting we suppose given a prior distribution $P_j(d\vartheta_j)$ for ϑ_j and we try to minimize

$$J(\Phi) = \limsup \frac{1}{t} \int R(t, \vartheta_1, \cdots, \vartheta_N, \Phi) P_1(d\vartheta_1) \cdots P_N(d\vartheta_N). \tag{11}$$

It is an easy matter to construct near-optimal rules for this problem. Note that (under some simple restrictions on the density $f(x, \vartheta)$) the mean $\mu(\vartheta_j)$ can be accurately estimated by the sample mean, i.e., for $\delta > 0$, there exists $T < \infty$ such that for every j

$$P_j \{ | \mu(\vartheta_j) - \frac{1}{T} \sum_{t=1}^{T} X^j(t) | < \delta \} > 1 - \delta. \tag{12}$$

Now consider the following two-phased rule Φ_δ: In the first or estimation phase the rule selects each of the N arms at least T times, and in the second phase the rule selects the m arms with the largest sample means at the end of the estimation phase. >From (12) it follows that this rule is near-optimal since

$$J(\Phi_\delta) \to 0 \text{ as } \delta \to 0.$$

Although Φ_δ is better than a fixed, non-adaptive rule, there are good reasons for not considering it close to optimal. First, observe that no matter how small $\delta > 0$ is, with positive probability the m arms selected will not have the largest mean values for a set of configurations C, and for each of these configurations the regret will grow proportional to t. Second, as will be seen below, there *do* exist rules Φ for which $\frac{1}{t} R(t, C, \Phi) \to 0$ for *every* C. Such a rule is qualitatively superior to Φ_δ, and leads us to conclude that the Bayesian loss (11) while it excludes non-adaptive rules, it does not adequately discriminate among adaptive rules.

This discussion suggests that we should impose the *adaptation* requirement on admissible rules:

$$\lim\sup \frac{1}{t} R(t, C, \Phi) = 0, \text{ for all } C. \tag{13}$$

This is a significant restriction. For example, it excludes all rules that, like Φ_δ, stop learning after a predetermined finite time. Indeed, it is not at all clear whether there exist rules satisfying (13). Furthermore, if there is such a rule Φ, then any rule Φ' that selects the same arms as Φ except over $n(t)$ time instants during $1, \cdots, t$ with $\frac{n(t)}{t} \to 0$ will also satisfy (13). This brings us to finally to the problem of distinguishing among arms that satisfy (13) and to the work of Lai and Robbins.

5. Asymptotically efficient adaptive rules

In a remarkable study [4-6] Lai and Robbins posed and answered the question of asymptotically efficient adaptive rules. Their work deals with the case $m = 1$. We summarize here its extension to $m > 1$ by Anantharam [1].

Recall that an arm is described by iid rewards with distribution $f(x, \vartheta)\nu(dx)$ and mean $\mu(\vartheta) = \int x f(x, \vartheta)\nu(dx)$. For a configuration $C = (\vartheta_1, \cdots, \vartheta_N)$ let σ be a permutation so that (8) holds. Let $0 \le l < m \le n \le N$ be such that

$$\mu[\vartheta_{\sigma(11)}] \ge \cdots \ge \mu[\vartheta_{\sigma(l)}] > \mu[\vartheta_{\sigma(l+1)}] = \cdots = \mu[\vartheta_{\sigma(m)}] =$$

$$= \cdots = \mu[\vartheta_{\sigma(n)}] > \mu[\vartheta_{\sigma(n+1)}] \ge \cdots \ge \mu[\vartheta_{\sigma(N)}].$$

We call $\sigma(1), \cdots, \sigma(l)$ the *best* arms, $\sigma(l+1), \cdots, \sigma(m)$ the *border* arms, and

$\sigma(n+1), \cdots, \sigma(N)$ the *worst* arms. [If $\mu[\vartheta_{\sigma(m)}] > \mu[\vartheta_{\sigma(m+1)}]$, then $\sigma(m)$ is simultaneously a best and border arm.]

A selection rule Φ is said to be *uniformly good* if $R(t, C, \Phi) = o(t^{\alpha})$ for every $\alpha > 0$ and every C. >From (9) it follows that Φ is uniformly good iff

$$E[t - T^j(t)] = o(t^{\alpha}) \text{ for every best arm } j,$$

$$E[T^j(t)] = o(t^{\alpha}) \text{ for every worst arm } j,$$

for every $\alpha > 0$ and every C.

The Kullback-Liebler number

$$I(\vartheta, \lambda) = \int \log \frac{f(x, \vartheta)}{f(x, \lambda)} f(x, \vartheta) \nu(dx)$$

is a well-known measure of dissimilarity between two distributions. In general $0 \leq I(\vartheta, \lambda) \leq \infty$. Define conditions A1-A4.

A1. $\mu(\vartheta)$ is strictly increasing in ϑ.

A2. $0 < I(\vartheta, \lambda) < \infty$, for $\lambda > \vartheta$.

A3. $I(\vartheta, \lambda)$ is continuous in $\lambda > \vartheta$ for fixed ϑ.

A4. For all λ, and all $\delta > 0$, there exists λ' with $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$.

Theorem 2. Suppose A1-A4 hold. Let Φ be any uniformly good rule and $C = (\vartheta_1, \cdots, \vartheta_N)$ be any configuration. Then

$$\liminf \frac{R(t, C, \Phi)}{\log t} \geq \sum_{j \text{ is worst}} \frac{[\mu(\vartheta_{\sigma(m)}) - \mu(\vartheta_j)]}{I(\vartheta_j, \vartheta_{\sigma(m)})}. \tag{14}$$

Thus every uniformly good rule must select each worst arm j at least $[I(\vartheta_j, \vartheta_{\sigma(m)})]^{-1} \log t$ times during $1, \cdots, t$. This number decreases as the "information distance" $I(\vartheta_j, \vartheta_{\sigma(m)})$ between arm j and the arm $\sigma(m)$ with the mth largest mean increases. [Remark: Unlike the mean $\mu(\lambda)$, the information distance $I(\vartheta, \lambda)$ need not increase with λ; however, that assumption is needed in Theorem 3.]

As example, in the Gaussian case, $f(x, \vartheta)\nu(dx) = N(\vartheta, \sigma^2)$ so $\mu(\vartheta) = \vartheta$. Then $I(\vartheta, \lambda) = (\vartheta - \lambda)^2 / 2\sigma^2$ and we get

$$\liminf \frac{R(t, C, \Phi)}{\log t} \geq \sum_{j \text{ is worst}} \frac{2\sigma^2}{\vartheta_{\sigma(m)} - \vartheta_j}.$$

Say that a rule Φ is **asymptotically efficient** if its regret achieves the lower bound (14) for every C.

The crucial feature in constructing an asymptotically efficient rule is this. At time t we have $T^j(t)$ observations of arm j from which we can estimate its mean. At $t+1$ we must decide whether to select the m arms whose estimated mean values are the largest -- "play the winners" rule -- or to select an apparently losing arm. The idea is to consider an apparently losing arm, say arm j, to estimate an *upper* bound for its mean value, and to compare that estimate with the estimate of the least best of the apparent winners.

We now describe a rule that is asymptotically efficient under the additional conditions A5, A6.

A5. $\log f(x, \vartheta)$ is concave in ϑ for each fixed x.

A6. $\int x^2 f(x, \vartheta)\nu(dx) < \infty$ for each ϑ.

Assumption A5 implies that $I(\vartheta, \lambda)$ is convex in λ, and since I is minimized at $\lambda = \vartheta$, it is increasing in λ for $\lambda > \vartheta$.

Let $X(1), X(2), \cdots$ be the sequence of rewards from an arm. Let $h: (0, \infty) \to (0, \infty)$ be a fixed continuous function with $\int h(s)ds = 1$, and let

$$W(a, \vartheta) = \int_0^\infty \prod_{b=1}^{a} \frac{f(X(b), \vartheta - s)}{f(X(b), \vartheta)} h(s)ds.$$

[A5 implies that $W(a, \vartheta)$ is increasing in ϑ.]

For $K > 0$, let

$$U(a, X(1), \cdots, X(a), K) = \inf \{\vartheta \mid W(a, \vartheta) > K\},$$

and, lastly, for a fixed $p > 1$, let

$$g(t, a, X(1), \cdots, X(a)) = \mu[U(a, X(1), \cdots, X(a), t(\log t)^p].$$

Now consider the following rule:

1. In the first N steps select each arm m times in order to establish an initial estimate.

2. Fix $0 < \delta < 1/N^2$. At any time t say that arm j is *well-sampled* if $T^j(t) > \delta t$. Then there are at least m well-sampled arms when $t > N$. At each t, from among the well-sampled arms choose the m *leaders* ranked by the sample mean $\mu^j(t)$ for arm j:

$$\mu^j(t) = \frac{X^j(1) + \cdots + X^j(T^j(t))}{T^j(t)}.$$

Now consider the decision at $t+1$. Consider the arm j for which $t+1 = j \bmod N$, and estimate its upper bound

$$\bar{\mu}^j(t) = g[t, T^j(t), X^j(1), \cdots, X^j(T^j(t))].$$

(a) If arm j is already one of the m leaders at time t, then select the m leaders at $t+1$.

(b) If arm j is not one of the leaders at t, and if its upper bound $\bar{\mu}^j(t) < \mu^k(t)$ for every m leader k, then again select the m leaders at $t+1$.

(c) If arm j is not one of the leaders at t, and if $\bar{\mu}^j \geq \mu^k(t)$ where k is a leader with the least mean estimate, then at $t+1$ select the (m-1) leaders other than k and arm j.

Note that at each time $(m-1)$ well-sampled arms with the largest estimated means are always selected.

Theorem 3. Suppose A1-A6 hold. Then this rule is asymptotically efficient.

6. Final remarks

Theorems 2 and 3 also hold without the "denseness" condition A4. They have been extended to the important case where the arms are finite Markov chains with stationary transition probability matrix depending upon one unknown parameter, see [1]. For several families of distributions, including Bernoulli, Poisson, Gaussian and double exponential, the statistics $g(t, a)$ can be calculated recursively, see [6].

Condition A5 is essential in the proof of Theorem 3. It would seem, however, that asymptotically efficient rules should exist under the condition that $I(\vartheta, \lambda)$ is increasing in λ for $\lambda > \vartheta$.

7. Acknowledgements

Research supported in part by the Joint Services Electronics Program Contract F49620-84-C-0057. This paper was written while Varaiya was visiting the International Institute of Applied Systems Analysis (IIASA). The kind hospitality of IIASA is acknowledged with pleasure.

8. References

[1] Anantharam, V., Ph.D Dissertation, Univ. of California, Berkeley, 1986.

[2] Gittins, J.C., "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc.*, vol. 41, 1979, 148-177.

[3] Gittins, J.C. and D.M. Jones, "A dynamic allocation index for the sequential design of experiments," in Gani, J., K. Sarkadi and I. Vince, Eds., *Progress in Statistics, Euro. Meet. Statist.*, vol. 1, New York, North-Holland, 1972, 241-266.

[4] Lai, T.L., "Some thoughts on stochastic adaptive control," *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, Dec. 1984, 51-56.

[5] Lai, T.L. and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, 1985, 4-22.

[6] Lai, T.L. and H. Robbins, "Asymptotically efficient allocation of treatments in sequential experiments," in Santner, T.J. and A.C. Tamhane (eds) *Design of Experiments*, New York, Marcel Dekker, 1985, 127-142.

[7] Varaiya, P., J.C. Walrand and C. Buyukkoc, "Extensions of the multiarmed bandit problem," *IEEE Trans. Automat. Contr.*, vol. AC-30, May 1985, 426-439.

[8] Weitzman, M.L., "Optimal search for the best alternative," *Econometrica*, vol. 47, 1979, 641-654.

[9] Whittle, P., "Multi-armed bandits and the Gittins index," *J. Roy. Statist. Soc.*, vol. 42, 1980, 143-149.

ON SOME NEW TYPES OF MATHEMATICAL MODELS OF COMPLEX SYSTEMS

Asarin Ye.A., Kozyakin V.S., Krasnosel'skii M.A., Kuznetsov N.A.
and Pokrovskii A.V.
Institute of Control Sciences
Moscow, USSR

The paper is aimed at consideration of two new models whose study has just begun.

1. Desynchronized linear models

Consider a system W consisting of independently operating parts $W^1, \dots, W^k$ referred hereinafter to as system components. Subsequent definitions and constructions will be referred to such situations when the components states are described by vectors of some dimensions. Below we shall restrict ourselves with the case when the state of each component is described by a scalar.

Assume the variable state of component W^j is a function $x_j(t) \quad (t \geq 0)$ which may vary its values jumpwise in some instants of time

$$0 < T_1^j < T_2^j < \dots < T_n^j < \dots , \tag{1}$$

where

$$\lim_{n \to \infty} T_n^j = \infty \quad (j = 1, \dots, k). \tag{2}$$

It is assumed that

$$x_j(t) = const \qquad (T^j_{n-1} < t \le T^j_n). \qquad (3)$$

To describe the functioning of component W^j one should specify the rule according to which the states following switching time instants are defined. Let us regard this rule to be described by the equality

$$x_j(T^j_n) = a_{j1}\, x_1(S^{j1}_{n-1}) + a_{j2}\, x_2(S^{j2}_{n-1}) + \dots + a_{jk}\, x_k(S^{jk}_{n-1}). \qquad (4)$$

The time instants $S^{j\tau}_{n-1}$ will be referred to as the component states measurement times.

It is quite natural to assume that

$$S^{j1}_n, S^{j2}_n, \dots, S^{jk}_n \le T^j_n \qquad (j=1,2,\dots k) \qquad (5)$$

and

$$\lim_{n\to\infty} S^{j\tau}_n = \infty \qquad (j,\tau=1,2,\dots,k). \qquad (6)$$

Formula (4) means that component W^j is capable of measuring its own state and the states of all other components and then use these measurements to calculate the correction in order to update its own (only its own) state. As evidence by this formula, the complex system under consideration is linear.

Formula (4) could have been assumed to simultaneously include measurements of states of one or several components at different instants of time. However this generalization would be only formal.

If the following equalities are true

$$S^{j\tau}_n = T^j_n = T_n \qquad (j,\tau=1,\dots,k;\ n=1,2,\dots), \qquad (7)$$

then system (4) may be written as

$$x(T_n) = A\,x(T_{n-1}) \quad (n = 2, 3, \dots), \tag{8}$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} \dots a_{1k} \\ \dots & \dots \\ a_{k1} & a_{k2} \dots a_{kk} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}. \tag{9}$$

The first problem arising in the use of complex systems models described by equations (4) is that of the asymptotic stability. In the case of a synchronized system (8) this problem is solved most easily: the asymptotic stability takes place if and only if

$$\sigma(A) < 1), \tag{10}$$

where $\sigma(A)$ is the spectral radius of matrix A, i.e. the largest of its absolute eigenvalues.

Our reasonings did not imply any conflicts between the system components whatsoever. Moreover all components are assumed to have a common goal which is to be attained by joint efforts. Such joint efforts are reflected in an appropriate choice of those coefficients in (4) which allow manipulation with respect to the unovoidable constraints imposed by the components structure and the available communication channels between the components.

At first sight, it seems reasonable to synchronize the system, i.e. aloways try to make equalities (7) true since this allows the use of simple mathematics to analyze the asymptotic stability. However one should bear in mind that it is desynchronization that is the easiest way to attain stability for some systems - indeed, system (8) may be unstable white its simplest desynchronizations (4) are stable.

Besides, in many cases there is no way to achieve synchronization since the system components are, in principle, separated (enterprises consisting of separate shops, systems of separately moving objects, movable objects and dispatch services, etc.) employ independent computers of different performance, and their possible updates are determined by different technical, financial and other potentials. One cannot speak of synchronization when the switching times are apriori unknown, planned updates for individual components shipped, etc.

A few words on the terminology used. If system (4) is of the form

$$x_j(nh+\varphi_j)=a_{j1}x_1[(n-1)h+\varphi_j]+\dots+a_{jk}x_k[(n-1)h+\varphi_j], \quad (11)$$

where $h>0$, and not all phase differences φ_j of components switchings are identical, then it will be called a phase desynchronized system. If its form is

$$x_j(nh_j+\varphi_j)=a_{j1}x_1[(n-1)h_j+\varphi_j]+\dots+a_{jk}x_k[(n-1)h_j+\varphi_j], \quad (12)$$

where not all periods h_j of switching the components are identical, it will be called frequency desynchronized.

2. Asymptotic stability of desynchronized systems

We have first given consideration to phase desyhchronized systems of two scalar state components. The degree of desynchronization could be as small as required, therefore there was a feeling that we might neglect desynchronization in the analysis of asymptotic stability of the system. Four classes of systems are possible: stable systems loosing their stability under desynchronizations; stable systems maintaining stability under arbitrary desynchronizations; unstable systems capable of gaining stability under desynchronizations of certain types, and finally, unstable systems maintaining their instability under all types of desynchronization.

As a simple example, consider system (11) with two scalar components. This system is synchronized with $\varphi_1 = \varphi_2$, and phase desynchronized with $\varphi_1 \neq \varphi_2$ ($0 < \varphi_1, \varphi_2 < h$). Let $a_{11} = a_{21} = a_{22} = -0,5$. If $a_{12} = 1$ the synchronized system is asymptotically stable while the desynchronized system is unstable. If $a_{12} = -0,6$ the synchronized system is unstable while the desynchronized system is asymptotically stable. With $a_{12} = 0$ both systems feature asymptotic stability and with $a_{12} = 2$ both are unstable.

As noted above, the stability theory for desynchronized systems has been insufficiently, developed. However some observations are already available. Some of them was reported in [1 - 5] and other papers while other results have been obtained quite recently. Below some easily formulated assertions will be given.

It is easy to prove that system (4) is asymptotically stable with any sequences of moments of updating and observation of all its components if the following condition is satisfied:

$$\sigma(|A|) < 1, \tag{13}$$

where $\sigma(|A|)$ is the spectral radius of the matrix

$$|A| = \begin{pmatrix} |a_{11}| & |a_{12}| \ldots |a_{1k}| \\ \cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot \\ |a_{k1}| & |a_{k2}| \ldots |a_{kk}| \end{pmatrix}. \tag{14}$$

Since all the elements of matrix (14) are nonnegative condition (13) may often be checked without any calculations of eigenvalues of matrix (14): condition (13) is satisfied iff for some vector $u_0 \in R^k$ with positive coordinates the strict coordinatewise inequality in true:

$$|A| u_0 < u_0. \tag{15}$$

Systems (4) with scalar components and nonnegative elements a_{ij} are insensitive to all kinds of desynchronizations - both the synchronized and desynchronized systems are either asymptotically stable, or unstable.

An important class is formed by desynchronized systems with the symmetric matrix A. If a synchronized system with matrix A is asymptotically stable the same is true of any desynchronized system with the same matrix. In a sense, desynchronized systems with the symmetric matrix A feature a greater degree of stability than synchronized systems with the same matrix. Thus, if no two components are allowed to switch simultaneously the sufficient condition for asymptotic stability of a desynchronized system states that the eigenvalues of matrix A should be less than 1 while its diagonal elements a_{ii}, more than - 1. The necessary and sufficient condition $\sigma(A) < 1$ of asymptotic stability for a synchronized system imposes a greater deal of constraints.

A more detailed analysis was carried out for two component phase desynchronized systems

$$\begin{aligned} x(n h_1 + \varphi_1) &= A_{11} x[(n-1) h_1 + \varphi_1] + A_{12} y[(n-1) h_1 + \varphi_1], \\ y(n h_2 + \varphi_2) &= A_{21} x[(n-1) h_2 + \varphi_2] + A_{22} y[(n-1) h_2 + \varphi_2]. \end{aligned} \tag{16}$$

Here the states x and y are vectors of some dimensions, and A_{ij} are matrices of the corresponding orders. Major results in the study of asymptotic stability of systems (16) were obtained by our young colleague A.F. Kleptsyn. At each step of Kleptsyn's algorithm a five-tuple of matrices $C_1(s), \ldots, C_5(s)$ is generated. These matrices are used to obtain the value of $\lambda(s)$. If $\lambda(s) < 1$, system (16) is asymptotically stable. Otherwise if $\lambda(s) \geqslant 1$ some explicit rules are employed to construct a new five-tuple of matrices yielding the value of $\lambda(s+1)$, and the validity of the inequality $\lambda(s+1) < 1$ is checked.

The algorithm suggested by A.F. Kleptsyn has an interesting feature. It is unable to detect unstability of system (16) which makes it rather "distressing" for a researcher. However in case system (16) is asymptotically stable this fact is revealed by the algorithm at some step.

If the update times are unknown it is more reasonable to apply probabilistic techniques to investigate the desynchronized systems behaviour.

Consider system (4) of the form

$$x_j(T_n^j) = a_{j1}\, x_1(T_{n-1}^j) + \ldots + a_{jk}\, x_k(T_{n-1}^j). \tag{17}$$

Assume each sequence T_n^j $(n = 1, 2, \ldots)$ is a simplest random flow of events with intesity $\lambda_j > 0$, the flows with different j being independent of one other. Random events flow T_n is called simplest with intensity $\lambda > 0$ if the values of $T_n - T_{n-1}$ and are independent and distributed identically with the density

$$p(t) = \lambda e^{-\lambda t} \quad (t > 0). \tag{18}$$

The above conditions, in particular, imply that $T_n^j \to \infty$ with $n \to \infty$ with probability one, and that $T_n^j \neq T_m^\tau$ with $|j-\tau|+|n-m| > 0$ with the same probability.

Introduce the k-th order square matrices

$$A^\tau = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{\tau 1} & a_{\tau 2} & \dots & a_{\tau\tau} & \dots & a_{\tau k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & & 1 \end{pmatrix}, (\tau = 1, 2, \dots, k). \quad (9)$$

Put each matrix (19) into correspondence with its Kroneker square B^τ, i.e. the k^2 - order square matrix determined by the formula

$$B^\tau = \begin{pmatrix} A^\tau & 0 & \dots & 0 & \dots & 0 \\ 0 & A^\tau & \dots & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{\tau 1} A^\tau & a_{\tau 2} A^\tau & \dots & a_{\tau\tau} A^\tau & \dots & a_{\tau k} A^\tau \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix} (\tau = 1, 2, \dots, k) \quad (20)$$

and, finally, construct a matrix

$$C = \frac{\lambda_1 B^1 + \dots + \lambda_k B^k}{\lambda_1 + \dots + \lambda_k}. \quad (21)$$

It turns out that inequality

$$\sigma(C) < \sigma_* < 1 \quad (22)$$

provides stability of the desynchronized system (17) in terms of probability. In other words, it follows from (22) that the probability $P\{\|x(t)\| > \varepsilon\}$ of the inequality $\|x(t)\| > \varepsilon$ tends to zero with $t \to \infty$ uniformly with respect to the initial states $x(0)$ from the unity sphere. True is the following estimate:

$$P\{\|x(t)\|>\varepsilon\} < \frac{d}{\varepsilon^2}\, e^{-(\lambda_1+\ldots+\lambda_k)[1-\sigma_*]\, t} \quad (\|x(0)\|\leq 1), \tag{23}$$

which characterizes the rate with which the trajectories of desynchronized system (17) converge to zero. Coefficient d in (23) is a function of merely the matrix (21) and the value of σ_* .

Note, furthermore, that inequality (22) is not necessary for the stable system (17).

In the authors' opinion , further development of the desynchronized system theory is of great interest.

3. Limit hysteresis nonlinearities

The concluding part of the paper is deveoted to an almost untouched mathematical operation associated with the known Bogoluybov - Krylov principle of averaging 6 on the one hand, and mathematical models of systems with hysteriesis, on the other.

In its classical form the averaging principle refers to systems described by the equations of the form

$$\frac{dx}{dt} = \varepsilon F(t,x), \quad x(0)=x_0 \quad (x\in R^n) \tag{24}$$

with a small parameter $\varepsilon > 0$ and a time oscillating (for instance, periodic) function $F(t,x)$. The averaging principle is in replacing (24) with the autonomous equation

$$\frac{dy}{dt} = \varepsilon F_1(y), \quad y(0)=x_0, \tag{25}$$

where

$$F_1(y) = \lim_{T\to\infty} \frac{1}{T}\int_0^T F(t,y)\,dt. \qquad (26)$$

The basic Bogolyubov's theorem includes a weakly limiting condition due to the fact that the solution of problem (24) with a small ε within a finite interval of variation of the slow variable $\tau = \varepsilon t$ displays little difference fromlthat of problem (25).

Now a few words on hysteresis nonlinearities. Phenomenologic models of hysteresis are sometimes associated with concepts of a multivalued function and, rather rardy, with hysteresis loops reflecting the system's responce on a periodic external action. As a rule, both such representations are insufficient.

More sophisticated phenomenological models of such nonlinearities as lost motion and rests, general hysterous and models suggested by Ishlinskii, Mizes and Trocks, Traizakh and Giltai, etc. and much more complete and allow consideration of a sufficiently wide classes of external actions. Such classes may, for instance, include certain sets of piecewise-monotonic continuous functions.

The next step in developing the hysteresis nonlinearities models implies treating them as systems with natural state spaces and input-state/input-output operators. The transition from initial phenomenological representation (similar in ideology to the transition from integral sums to integrals, but differing from this transition in realization technique) allows one to regard the above operators to be defined at the corresponding complete functional spaces and to feature some useful properties. Realization of the system approach to describing hysteresis nonlinearities has required substantial difficulties to be overcome – see [7].

A general description of these and other forms of hysteresis nonlinearities looks as follows. A hysteresis nonlinearity is a physically realizable deterministic system W with continuous input $u(t)$, states $\omega(t)$ and outputs $\xi(t)$ Extended states of the system $\{u, \omega\}$ form the set $\Omega(W)$. Here we consider only those nonlinearities whose properties are time independent and whose functioning laws are independent of the reference point and time scale. Under these conditions and with the given initial extended state of the system

$$M_0 = \{u(t_0), \omega(t_0)\} \in \Omega(W) \tag{27}$$

its input $u(t)$ determines the law of the system state variation

$$\omega(t) = W[t_0, \omega(t_0)]u(t) \quad (t \geq t_0) \tag{28}$$

and the law of its output variation

$$\xi(t) = \Xi[t_0, \omega(t_0)]u(t) \quad (t \geq t_0). \tag{29}$$

As an example, consider a stop and A. Ishlinskii's model.

For a stop $\mathcal{U}(h)$ with a 2h span the set $\Omega[\mathcal{U}(h)]$ is formed by a strip $|\omega| \leq h$ which may be conveniently presented (see Fig. 1) as a number of sectors of straight lines of the form $\omega = u + c$. Provided the input $u(t)$ is monotonic, the variable state

$$M(t) = \{u(t), \omega(t)\} \in \Omega[\mathcal{U}(h)] \quad (t \geq t_0) \tag{30}$$

describes a part of an open polygon of an increased thickness in Fig. 1. A transition to piecewise monotonic continuous inputs is carried out with the help of a semigroup identity which follows from the deterministic nature of the system. The limit transition allows consideration of arbitrary continuous inputs. In the case of a stop, the output coinsides with the state. Therefore operators (28) and (29) may be denoted by a common notation

$$\omega(t)=\mathcal{U}[t_0,\omega(t_0);h]u(t) \quad (t\geq t_0). \tag{31}$$

The state of the continuum family of stops $\mathcal{U}(h)$ $(h\geq 0)$ will be a function $\omega(h)$ such that $\omega(0)=0$ and

$$|\omega(h_1)-\omega(h_2)|\leq |h_1-h_2| \quad (h_1,h_2\geq 0). \tag{32}$$

The extended state $\{u,\omega(h)\}$ by curve $\tilde{\omega}$ parametrically specified by the equations

$$\gamma_1=u-h-\omega(h),\ \gamma_2=u+h-\omega(h). \tag{33}$$

In Fig. 2 curve $\tilde{\omega}$ is shown by a double line.

The continuum family of stops is transformed into the Ishlinskii nonlinearity if operators (28) and (29) are defined as the equalities

$$\omega(h;t)=\mathcal{U}[t_0,\omega(h;t_0);h]u(t) \quad (t\geq t_0) \tag{34}$$

and

$$\xi(t)=\int_0^{h_*}\mathcal{U}[t_0,\omega(h;t_0);h]u(t)d\mu(h) \quad (t\geq t_0), \tag{35}$$

where $\mu(h)$ is some finite measure given on the interval $[0,h_*]$.

Let us now turn to differential equations with hysteresis nonlinearities which are described by the above model. In case of the equations with a small parameter we obtain a system consisting of the following differential equation

$$\frac{dx}{dt} = \varepsilon f(t, x, \xi), \quad x(0) = x_0 \tag{36}$$

and the operator equation

$$\xi(t) = \Xi[0, \omega_0] g[t, x(t)], \{g(0, x_0), \omega_0\} \in \Omega(W). \tag{37}$$

Assume functions $f(t, x, \xi)$ and $g(t, x)$ to be oscillating (for instance, periodically) in time t. An od-hoc averaging procedure (see, for instance, [8]) may, in certain cases, reduce problem (36) - (37) to problem (25). For such cases an analog of Bogolyubov's theorem was found which gives ground for the averaging.

However generally an averaging procedure leads to equations significally differing from (25):

$$\frac{dx}{dt} = \varepsilon \Phi(\omega_0) x(t), \quad x(0) = x_0, \tag{38}$$

whose righthand part consists of the Volterra nonlinear operator

$$\Phi(\omega_0) x(s) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f\{\tau, x(s), W_*[0, \omega_0] g(\tau, x(s))]\} d\tau, \tag{39}$$

where $W_*[0, \omega_0]$ is the limit hysteresis nonlinearity whose importance is emphasized by the authors.

The concept of the limit hysteresis nonlinearity is applied to the functions $u(t, s)$ of two variables for which $\{u(0, 0), \omega_0\} \in \Omega(W)$, by means of the equality

$$W_*[0, \omega_0] u(t, s) = \lim_{\delta \to 0} \Xi[0, \omega_0] u[t, \varphi(t, s; \delta)], \tag{40}$$

where

$$\varphi(t, s; \delta) = \begin{cases} \delta t & \text{with } 0 \leq t \leq \delta^{-1} s, \\ s & \text{with } t \geq \delta^{-1} s. \end{cases} \tag{41}$$

The investigation of limit hysteresis nonlinearities has just begun; some important results were obtained by T. Gilman an A. Vladimirov. However a number of situations were found when finding of the values of operator $W_*[0, \omega_0]$ is reduced to simple manipulations. As an example let us describe the procedure of obtaining the limit hysteresis nonlinearity operator magnitudes corresponding to A. Ishlinskii's transformations.

Assume that with $0 \leq \sigma \leq s$

$$\gamma_1(\sigma) = \inf_{-\infty < t < \infty,\ \sigma \leq \mu \leq s} u(t, \mu),$$

$$\gamma_2(\sigma) = \sup_{-\infty < t < \infty,\ \sigma \leq \mu \leq s} u(t, \mu) \qquad (42)$$

and plot curve Γ in plane $\{\gamma_1, \gamma_2\}$ with a fixed s (shown by a line of an increased thichness in Fig. 2). This curve is specified by its parametric equations

$$\gamma_1 = \gamma_1(\sigma), \ \gamma_2 = \gamma_2(\sigma) \quad (0 \leq \sigma \leq s). \qquad (43)$$

The end points of curve Γ are

$$N(0) = \{\gamma_1(0), \gamma_2(0)\} \text{ and } N(s) = \{\gamma_1(s), \gamma_2(s)\}. \qquad (44)$$

Using curve $\tilde{\omega}_0$ describing the initial state of the Ishlinskii transformation (shown in Fig. 2 by a double line) and curve Γ we have to obtain some new extended state $\{u(0, s), \omega_s(h)\}$ in the form of a curve $\tilde{\omega}_s$ in plane $\{\gamma_1, \gamma_2\}$ in order to find the magnitude of operator (40).

Curve Γ and the section connecting point $N(s)$ and point $N^* = \{u(0,s), u(0,s)\}$ will be a part of curve $\tilde{\omega}_s$. Then we must include into this curve the part of curve $\tilde{\omega}_0$ which lies neither to the right of nor lower than point $N(0)$.

Let the terminal point of this section be denoted as N^{**}. To complete the construction me just have to include into curve $\tilde{\omega}_S$ the section which connects points $N(0)$ and N^{**}. Curve $\tilde{\omega}_S$ is shown yby a dotted line in Fig. 2.

It turns out that in a t-periodic input $u(t,s)$ the limit hysteresis nonlinearity magnitude corresponding to the Ishlinskii transformation is determined by the equality

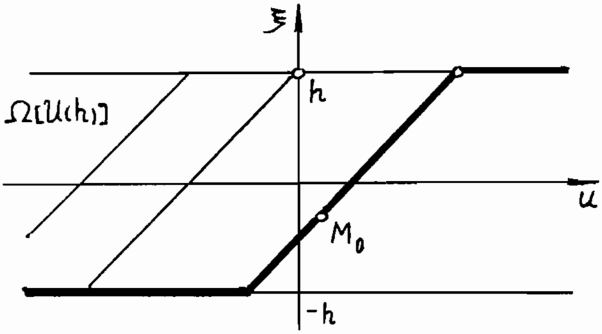
$$W_*[0,\omega_0(h)]u(t,s)=P\int_0^{h_*}\mathcal{U}[0,\omega_S(h);h]u(t;s)\,d\mu(h), \tag{45}$$

where P is an operator putting each function $z(t)$ periodic at large t's into correspondence with the function $Pz(t)$ periodic along the entire numerical axis and coinsiding with $z(t)$ at large t's. A similar formula holds in the case of almost t-periodic functions $u(t,s)$.

Note in conclusion that N. Bogolyubov's theorem on the averaging principle maintains its significance when turning from problem (36) - (37) to problem (38) featuring operator (39). This fact shows a reason for the importance of limit hysteresis nonlinearities.

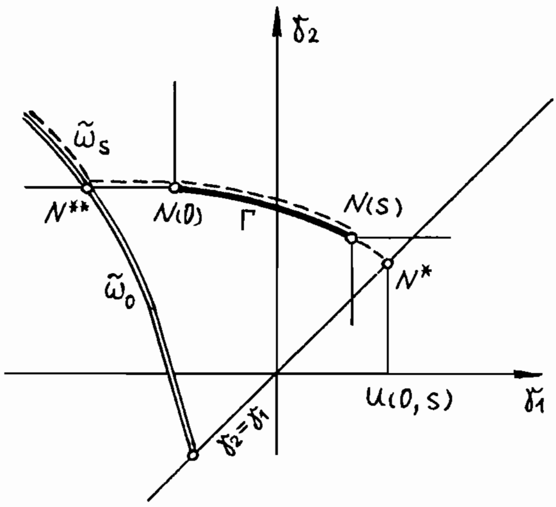
References

1. Kleptsyn A.F., Kozyakin V.S., Krasnosel'skii M.A. and Kuznetsov N.A. On the effect of small desynchronization on stability of complex systems. Parts I-III. Avtomatika i Telemekhanika, 1983, No.7, pp.44-51; 1984, No.3, pp.42-47; 1984, No.8, pp.63-67 (in Russian).
2. Kleptsyn A.F., Kozyakin V.S., Krasnosel'skii M.A. and Kuznetsov N.A. Stability of desynchronized systems. Doklady AN SSSR, 1984, vol.274, No.5, pp.1053-1056.
3. Kleptsyn A.F., Krasnosel'skii M.A., Kuznetsov N.A., Kozjakin V.S. Desynchronization of linear systems.- Mathematics and Computers in Simulation, 1984, XXVI, p.423-431.
4. Kleptsyn A.F. On stability of special type complex desynchronized systems. Avtomatika i Telemekhanika, 1985, No.4, pp.169-172. (in Russian).
5. Krasnosel'skii A.M. On stability of desynchronized multicomponet systems. Avtomatika i Telemekhanika, 1985, No.11, pp.170-171 (in Russian).
6. Bogolyubov N.N., Mitripol'skii Yu.A. Asymptotic methods in the theory of nonlinear oscillations. Nauka, Moscow, 1974 (in Russian).
7. Krasnosel'skii M.A., Pokrovskii A.V. Systems with hysteresis. Nauka, Moscow, 1983 (in Russian).
8. Pisarenko G.S. Oscillations of mechanical systems with regard to imperfect elastisity of the material. AN Ukr.SSR Publ., Kiev, 1970 (in Russian).



ξ
h
Ω[U(h)]
u
M_0
-h

Fig. I



γ_2
$\tilde{\omega}_s$
N^{**}
$N(0)$
Γ
$N(s)$
N^*
$\tilde{\omega}_0$
$U(0,s)$
γ_1
$\gamma_2=\gamma_1$

Fig. 2

VIABILITY TUBES

Jean Pierre Aubin

CEREMADE, Universite de Paris-Dauphine

International Institute for Applied Systems Analysis

Abstract

We define viability tubes and invariant tubes of a differential inclusion, we study some asymptotic properties and we characterize them by showing that the indicator functions of their graphs are solutions to the contingent Hamilton-Jacobi equation. We provide some examples of viability tubes.

Contents

Introduction

Let X be a finite dimonsional vector space and $F : [o\ \infty[\times X \to X$ a set-valued map which associates with any state $x \in X$ and any time t the subset $F(t\ ,x)$ of velocities of the system. The evolution of the system is governed by the differential inclusion

(*) $x'(t) \in F(t\ ,x(t))\ ,x(t_0) = x_0$

We consider now "tubes", i.e., set-valued maps $t \to \underline{P}(t)$ from $[o\ ,\infty[$ to X. We say that a trajectory $t \to x(t) \in X$ is "viable" (in the tube $\underline{P}$) if

(**) $\forall t \geq 0\ ,x(t) \in \underline{P}(t)$

A tube $\underline{P}$ enjoys the viability property if and only if, for all $t_0 \geq 0$ and $x_0 \in \underline{P}(t_0)$, there exists at least a solution $x(\cdot)$ to the differential inclusion (*) which is viable.

Remark

A simple-valued tube $t \to \{x(t)\}$ enjoys the viability property if and only if $x(\cdot)$ is a solution to the differential inclusion (*). So it is legitime to regard a tube having the viability property as a "multivalued solution" to the differential inclusion (1).

The knowledge of a tube enjoying the viability property allows to infer some informations upon the asymptotic behaviour of some solutions to the differential inclusion (1), as we do with Lijapunov functions. They also share the same disadvantages: the dynamics F being given, how do we construct the tubes of F?

We shall begin by characterizing such tubes as "viability tubes". For that purpose, we need an adequate concept of derivative of set-valued map, the "contingent derivative" defined as follows:

If $x \in \underline{P}(t)$, v belongs to $D\underline{P}(t\ ,x)$ (1) if $\liminf_{h \to 0+} d(v\ ,\frac{\underline{P}(t+h)-x}{h}) = 0$

Viability tubes are those tubes satisfying

(***) $\forall t \geq 0\ ,\forall x \in \underline{P}(t)\ ,F(t\ ,x) \cap D\underline{P}(t\ ,x)(1) \not\subset \phi$

We can regard (***) has a "differential equation for tubes".

We prove in the second section that the "limit" when $t \to \infty$ of a viability tube $\underline{P}(t)$ (namely, the Kuratowski limsup) is a viability domain: hence targets of a differential inclusion are necessarily viability domains. We construct in the fourth section the largest viability tube "converging" to a given target. We also provide a surjectivity criterion which is useful for solving such problems.

We can characterize viability tubes $\underline{P}(t)$ by the indicator functions $V_{\underline{P}}$ of their graphs, defined by: $V_{\underline{P}}(t\ ,x) : =0$ if $x \in \underline{P}(t)$, $+\infty$ if not. We thus observe that $\underline{P}$ is a viability tube if and only if $V_{\underline{P}}$ is a solution to the "contingent Hamilton-Jacobi equation".

$$\inf_{v \in F(t\ ,x)} D_+ V(t\ ,x)(1\ ,v) = 0$$

where

$$D_+ V(t\ ,x)(1\ ,v) : = \liminf_{\substack{h \to 0+ \\ v' \to v}} \frac{V(t+h\ ,x+hv') - V(t\ ,x)}{h}$$

is the contingent epiderivative of V at $(t\ ,x)$ in the direction $(1\ ,v)$.

We then investigate tubes enjoying a dual property, the ***invariance property***: for all $t_0 \geq o$ and $x_0 \in \underline{P}(t_0)$, ***all*** solutions to the differential inclusion are viable.

We justify in section 7 the claim that viability tube and invariant tube are in some convenient sense "dual". When $F(t\ ,x) :=A(t)x$ is "set-valued linear operator" (called a closed convex process), we can define its "transpose". Therefore, we associate with the "linear differential inclusion"

$$x'(t) \in A(t)x(t)$$

its "adjoint" differential inclusion

$$-p'(t) \in A(t)^{*}p(t)$$

We show that if a tube $t \to R(t)$, the values of which are closed convex cones, enjoys the invariance property (for the original system), its polar tube $t \to R(t)^{+}$, where $R(t)^{+}$ is the positive polar cone to $R(t)$, is a viability tube of the adjoint differential inclusion.

We end this exposition of viability tubes with two families of examples. In section 8, we investigate "finite horizon" tubes of the form

$$\underline{P}(t) := \varphi(t\ ,G\ ,D)$$

where $\varphi(o\ ,C\ ,D) = C$ and $\varphi(T\ ,C\ ,D) = D$, which "carry" a subset C to a subset D. In the last section, we consider tubes derived from "potential functions" in the following way

$$\underline{P}(t) := \{x \mid V(t\ ,x(t) - c(t)) \leq w(t)\}$$

where $c(t)$ and $w(t)$ are given functions.

1. Viability Tubes

Let X be a finite dimensional vector space. We consider a set-valued map $F : [o\ ,T] \times X \to X$ which associates with every $(t\ ,x)$ the subset $F(t\ ,x)$ of velocities of the system at time t when its state is $x \in X$. We shall study the *differential inclusion.*

$$\begin{cases}(i)\ \ x'(t) \in F(t\ ,x(t))\ \text{ for almost all }\ t \in [t_0\ ,T[\\ (ii)\ \ x'(t) \in F(T\ ,x(t))\ \text{ for almost all }\ t \geq T\ (\text{if } T < +\infty)\\ (iii)\ \ x(t_0) = x_0\end{cases} \tag{1.1}$$

It will be convenient to regard a set-valued map $\underline{P}$ from $[o\ ,T]$ to X as a "tube".

Definition 1

We say that a tube $\underline{P}$ enjoys the "viability property" if and only if for all $t_0 \in [o\ ,T]$, $x_0 \in \underline{P}\ (t_0)$, there exists a solution $x(\cdot)$ to (1) which is "viable" in the sense that

$$\begin{cases}(i)\forall t \in [t_0\ ,T]\ ,x(t) \in \underline{P}(t)\\ (ii)\ \text{if } T < +\infty\ ,\forall t \geq T\ ,x(t) \in \underline{P}(T)\end{cases} \tag{1.2}$$

A subset K has the "viability property" if and only if the "constant tube" $t \to \underline{P}(t) :=K$ does enjoy it.

For time independent systems, we know how to characterize closed subsets K which enjoy the viability property (see Haddad [1981], Aubin-Cellina [1984]). For that purpose, we introduce the "contingent cone" $T_K(x)$ to K at x, the closed cone of vectors $v \in X$ such that

$$\liminf_{h \to 0+} \frac{d(x + hv\ ,K)}{h} = 0$$

A subset K is said to be a "viability domain" of a set-valued map $F : X \to X$ if and only if

$$\forall x \in K \, , F(x) \cap T_K(x) \neq \phi$$

When F is upper semicontinuous with compact convex images, such that $\| F(x) \| \leq a \, (\| x \| + 1)$, Haddad's viability theorem states that a closed subset K enjoys the viability property if and only if it is a viability domain.

Our first task is to characterize tubes enjoying the viability property thanks to its "contingent derivative" (see Aubin [1981], Aubin-Ekeland [1984]). We recall that

$$v \in D\underline{P}(t \, , x)(\tau) \Longleftrightarrow \liminf_{\substack{h \to o+ \\ \tau' \to \tau}} d\left[v \, , \frac{\underline{P}(t + \tau' h) - x}{h} \right] = 0 \tag{1.3}$$

We observe that it is enough to know this contingent derivative in the only directions 1, 0 and -1. In particular, we note that

$$\left\{ \begin{array}{l} (i) D\underline{P}(t \, , x) \, (1) = \{ v \in X \mid \liminf\limits_{\substack{h \to o+ \\ \tau' \to 1}} d\left[v \, , \dfrac{\underline{P}(t + \tau' h) - x}{h} \right] = 0 \\ (ii) T_{\underline{P}(t)}(x) \subset D\underline{P}(t \, , x)(o) \end{array} \right. \tag{1.4}$$

(Equality in (1.4) (i) holds when $\underline{P}$ is Lipschitzian in a neighborhood of x).

We observe that the graph of $D\underline{P}(t \, , x)$ is the contingent cone to the graph of $\underline{P}$ at $(t \, , x)$.

Definition 2

A tube $\underline{P} : [o \, , T] \to X$ is called a "viability tube" of a set-valued map $F : [o \, , T] \times X \to X$ if its graph is contained in the domain of F and if

$$\left\{ \begin{array}{l} (i) \;\; \forall t \in [o \, , T[\; \forall x \in \underline{P}(t) \, , F(t \, , x) \cap D\underline{P}(t \, , x)(1) \neq \phi \\ (ii) \;\; \text{if } T < \infty \; \forall x \in \underline{P}(T) \, , F(T \, , x) \cap D\underline{P}(T \, , x)(o) \neq \phi \end{array} \right. \tag{1.5}$$

A tube is said to be "closed" if and only if its graph is closed. Haddad's viability theorem for autonomous systems and other results imply easily the following:

THEOREM 1

Assume that the set-valued map $F : [o \, , \infty[\times X \to X$ satisfies:

$$\left\{ \begin{array}{l} (i) \;\; F \text{ upper semi-continuous with closed convex values} \\ (ii) \;\; F(t \, , x) \| \leq a \, (\| x \| + 1) \end{array} \right. \tag{1.6}$$

(a) a necessary and sufficient condition for a closed tube to enjoy the viability property if and only if $\underline{P}(\bullet)$ is a viability tube.

(b) There exists a largest closed viability tube contained in the domain of F.

(c) If $\underline{P}_n$ is a sequence of closed viability tubes, then the tube $\underline{P}$ defined by the Kuratowski upper limit

$$Graph \; (\underline{P}) : = \limsup_{n \to \infty} \text{Graph}(\underline{P}_n) \tag{1.7}$$

is also a (closed) viability tube.

Proof

We introduce the set-valued map G from Graph $(\underline{P})$ to $\mathbf{R}_+ \times \mathbf{R}^n$ defined by

$$G(s\ ,\ x) : = \begin{cases} \{1\} \times F(s\ ,\ x) & \text{if } s \in [o\ ,\ T[\\ [o\ ,\ 1] \times F(T\ ,\ x) & \text{if } s = T \\ \{o\} \times F(T\ ,\ x) & \text{if } s > T \end{cases}$$

We observe that $(s\,(\bullet)\ ,\ x\,(\bullet))$ is a solution to the differential inclusion

(i) $(s\,'\,(t)\ ,\ x\,'(t)) \in G(s\,(t)\ ,\ x\,(t))$

(ii) $(s\,(t_0)\ ,\ x\,(t_0)) = (t_0\ ,\ x_0)$

if and only if x is a solution to the differential inclusion (1). We also note that the tube $\underline{P}$ has the viability property if and only if its graph enjoys the viability property for G and that $\underline{P}$ is a viability tube if and only if its graph is a viability domain of G. It thus remains to translate the time independent results.

2. Asymptotic properties of viability tubes

Theorem 2

Consider a set-valued map F from X to X, which is assumed to be upper semicontinuous, convex compact valued and satisfies

$$\| F(x) \| \leq a(\| x \| + 1) \text{ for all } x \in Dom(F)$$

Then the Kuratowski upper limit

$$C : = \limsup_{t \to \infty} \underline{P}(t)$$

is a viability domain of F.

Proof

We shall prove that C enjoys the viability property. Let ξ belong to C. Then $\xi = \lim \xi_n$ where $\xi_n \in \underline{P}(t_n)$. We consider the solutions x_n to the differential inclusion.

$$x_n\,'(t) \in F(x_n\,(t))\ ,\ x_n\,(t_n) = \xi_n$$

which are viable in the sense that

$$\forall t \geq t_n\ ,\ x_n\,(t) \in \underline{P}(t)$$

The function y_n defined by $y_n\,(t) : = x_n\,(t + t_n)$ are solutions to

$$y_n\,'(t) \in F(y_n\,(t))\ ,\ y_n\,(o) = \xi_n$$

The assumptions of Theorem 2 imply that these solutions remain in a compact subset of $\mathbf{C}(o\ ,\ \infty\ ; X)$. Therefore, a subsequence (again denoted) in convergence to y, which is a solution to

$$y\,'(t) \in F(y\,(t))\ ,\ y\,(o) = \xi\ .$$

Furthermore, this solution is viable in C since for all $t \geq o\ ,\ y\,(t)$ is the limit of a subsequence of $y_n\,(t) = x_n\,(t + t_n) \in \underline{P}(t + t_n)$ and thus belongs to C.

3. The target problem

A closed viability domain C of F being given regarded as a "target", find the largest closed viability tube $\underline{P}_c$ ending at C in the sense that

$$\underline{P}_c(T) = C \qquad \text{if } T < +\infty$$

or

$$\limsup_{t\to\infty} \underline{P}_c(t) = C \qquad \text{if } T = +\infty$$

Knowing such a tube $\underline{P}_c$, we thus deduce that starting at time o from $K := \underline{P}_c(o)$, a solution to the differential inclusion $x' \in F(x)$ must bring this initial state to the target.

Proposition 1

The assumptions are those of Theorem 2. We can associate with any closed viability domain C of F a largest viability tube $\underline{P}_c$ ending at C. This tube is closed if we assume, for instance, that for any compact subset K, the set S of solutions to

$$x'(t) \in F(x(t)),\ x(o) \in K$$

is compact in the Banach space $B(o\ ,\ \infty\ ; X)$ of bounded functions.

Proof

(a) The solution is obvious when $T < +\infty$: We take

$$\underline{P}_c(t) := \{x(t) \mid x' \in F(x),\ x(T) \in C\}.$$

It has the viability property: if $(t\ ,\ \xi)$ belongs to the graph of $\underline{P}_c$, there exists a solution x to the differential inclusion $x' \in F(x)$ such that $x(t) = \xi$ and $x(T) \in C$ and $x(s)$ belongs to $\underline{P}_c(s)$ for all $s \geq t$ by the very definition of $\underline{P}_c$. Hence it is viability tube ending at C. It is the largest one: if $\underline{P}$ is any viability tube, then, for all $(t\ ,\ \xi) \in Graph(\underline{P})$, there exists, thanks to the viability theorem, a solution x to $x' \in F(x)$ such that $x(s) \in \underline{P}(s)$ for all $s \geq t$. Since $x(T) \in \underline{P}(T) \subset C$, so that ξ belongs to $\underline{P}_C(t)$.

The graph of $\underline{P}_C$ is closed : if $\xi_n \in \underline{P}_c(t_n)$ and if $(t_n\ ,\ \xi_n)$ converges to $(t\ ,\ \xi)$, we see that $(t\ ,\ \xi)$ belongs to the graph of $\underline{P}_c$. For there exists a sequence of solutions x_n to $x_n' \in F(x_n)$ satisfying $x_n(t_n) = \xi_n$ and $x_n(T) \in C$. Since these solutions remain in a compact subset of $C(o\ ,\ T\ ; X)$, a subsequence (again denoted) x_n converges uniformly to a solution x to the differential inclusion $x' \in F(x)$ which satisfies $x(t) = \xi$ and $x(t) = \lim_{n\to\infty} x_n(t) \in C$

We also observe that

$$\underline{P}_C(t) = \{y(T-t) \mid y' \in -F(y),\ y(o) \in C\}$$

Those two subsets do coincide because x is a solution to $x' \in F(x)$ if and only if the function y defined by $y(t) := x(T-t)$ is a solution to $y' \in -F(y)$ such that $y(o) = x(T)$.

(b) Consider now the case when $T = \infty$ and denote by L the set-valued map associating with any continuous function $x(\cdot) \in C(o\ ,\ \infty\ ,\ X)$ its limit set

$$L(x) := \limsup_{t\to\infty} \{x(t)\} = \bigcap_{T\geq o} cl(x([T\ ,\ \infty[))$$

The same arguments as those in the finite horizon case imply that the tube $\underline{P}_C$ defined by

$$\underline{P}_C(t) := \{x(t) \mid x' \in F(x),\ L(x) \subset C\}$$

is the largest viability tube "converging" to C. We have to show that it is closed. As in the finite horizon case, we consider a sequence $(t_n , x_n) \in Graph P_C$ which converges to (t , x) and solutions x_n to

$$x'_n(t) \in F(x_n(t)) , x_n(t_n) = \xi_n \text{ and } L(x_n) \subset C$$

Since the ξ_n's belong to a compact K, the last assumption we made implies that the solutions $x_n(\cdot)$ lie in a compact subset of $B(o , \infty , X)$. A subsequence (again denoted) $x_n(\cdot)$ converges uniformly on $[o , \infty[$ to a solution $x(\cdot)$ to $x' \in F(x) , x(t) = \xi$. We deduce that its limit set $L(x)$ is contained in C from the fact that the set-valued map L is lower semicontinuous: for if y belongs to $L(x)$ and if a sequence x_n converges uniformly to y, then there exists $y_n \in L(x_n) \subset C$ which converges to y, and which thus belongs to C, which is assumed to be closed. The lower semicontinuous of L follows from:

Lemma 1 Let $B(o , \infty , X)$ be the Banach space of bounded continuous functions. The set-valued map L is lower semicontinuous from $B(o , \infty ; X)$ to X.

Proof of Lemma 1

Let $\xi \in L(x)$ and $x_n \in B(o , \infty , X)$ converge uniformly to x on $[o , \infty[$. There exists $t_k \to \infty$ such that $x(t_k)$ converges to ξ. Further, for all $\varepsilon > o$, there exists N such that $\| x_n(t_k) - x(t) \| \leq \varepsilon$ for all $n \geq N$. Hence $\|x_n(t_k) - \xi \| \leq \varepsilon$ for all t_k large enough. Since the dimension of X is finite, the subsequence $x_n(t_k)$ converges to an element ξ_n which belongs to $L(x_n)$ and thus, $\| \xi_n - \xi \| \leq 2 \varepsilon$ for all $n \geq N$. Hence L is lower semicontinuous.

4. A surjectivity criterion for set-valued maps

We propose now a criterion which allows to decide whether a compact convex subset C lies in the target of a differential inclusion. It belongs to the class of surjectivity theorems for "outward maps" (see Aubin-Ekeland, [1984]). The idea is the following. We consider a set-valued map R (the reachable map in our framework) from a subset K of a Hilbert space X to another Hilbert space Y. We want to solve the following problem:

$$\text{For every } y \in C \text{ , find } x \in K \text{ such that } y \in R(x) \tag{4.1}$$

(i.e. we can reach any element of the target C from K).

Assume that we know how to solve this problem for a "nicer" set-valued map Q from K to Y (say, a map with compact convex graph).

$$\text{For every } y \in C \text{ , find } x \text{ such that } y \in Q(x) \tag{4.2}$$

The next theorem states how a relation linking R and Q (R is "outward with respect to" Q) allow to deduce the surjectivity of R from the surjectivity of Q.

Theorem 3

We assume that the graph of Q is convex and compact and that R is upper semicontinuous with convex values. We set

$$K := Dom\ Q , \quad C := Im\ Q \tag{4.3}$$

If R is "outward with respect to" Q in the sense that

$$\forall x \in K , \forall y \in Q(x) , y \in R(x) + T_c(y) , \tag{4.4}$$

then R is surjective from K to C (in the sense that $C \subset R(K)$).

Proof

It is a simple consequence of Theorem 6-4.12 p.343 of Aubin-Ekeland [1984]. We replace X by $X \times Y$, K by Graph Q (which is convex compact), A by the projection π_y from $X \times Y$ to Y and R by the set-valued map G from $X \times Y$ to Y defined by:

$$G(x\ ,y) : =R(x) - y_0 \text{ where } y_0 \text{ is given in } C. \tag{4.5}$$

The outwardness condition implies that the tangential condition is satisfied. $O \in -y + R(x) + T_C(y)$ and, since $y_0 - y$ belongs to $T_C(y)$ (because $y_0 \in C$), then $O \in -y_0 + R(x) + T_C(y) = G(x\ ,y) + T_C(y)$.

We observe that

$$T_C(y) = T_{\mathrm{Im}Q}(y) = T_{\pi_y(Graph\ Q)}(\pi_y(x\ ,y)) = cl(\pi_y\ T_{Graph\ Q}(x\ ,y) - G(x\ ,y))$$

Theorem 6.4.12 implies the existence of $(\bar{x}\ ,\bar{y})$ in Graph Q, a solution to the inclusion $o \in G(\bar{x}\ ,\bar{y})$, i.e., to the inclusion $Y_0 \in R(\bar{x})$.

Remark

The dual version of the "outwardness condition" is the following:

$$\forall q \in N_C(y)\ ,\ \forall x \in A^{-1}(y)\ ,\ <q\ ,y> \leq \sigma(R(x)\ ,q) \tag{4.5}$$

where

$$\sigma(R(x)\ ,q) : =sup\{<q\ ,z> \mid z \in R(x)\}$$

is the support function of $R(x)$.

Remark By using the concept of T-selectionable maps introduced by Haddad-Lasry [1983] (see also Aubin-Cellina [1984], p. 235), we can extend the above theorem to the case when R is T-selectionable instead of being convex-valued. We obtain:

Theorem 4.

We assume that the graph of Q is convex and compact and that R is T-selectionable. If R is "strongly outward with respect to" Q in the sense that

$$\forall x \in K\ ,\ \forall y \in Q(x)\ ,\ R(x) \subset y - T_C(y) \tag{4.6}$$

than R is surjective from K to C.

Remark

Other sufficient conditions can be proposed to guarantee the surjectivity of R. For instance, "inwardness" condition

$$-C \subset \bigcap_{x \in K} (R(x) + T_C(Q(x))) \tag{4.7}$$

implies the surjectivity condition when R is upper semicontinuous with convex valued and "strong inwardness" condition

$$C - R(x) \subset \bigcap_{y \in Q(x)} T_C(y) \tag{4.8}$$

implies the surjectivity condition when R is only T-selectionable. We use the same methods applied to the set-valued map $H(x\ ,y) : =R(x) - y_0$.

5. Contingent Hamilton-Jacobi Equations

We may regard condition (1.5)(i) involved in the definition of viability tubes as a "set-valued differential inclusion" the solutions to which are "viability tubes" and condition (1.5)(ii) as a "final" condition.

Actually, conditions (1.5) defining "viability tubes" is a multivalued version of the Hamilton-Jacobi equation in the following sense.

We characterize a tube $\mathcal{P}$ by the indicator function $V_{\mathcal{P}}$ of its graph defined by

$$V_{\mathcal{P}}(t\ ,\ x) : = \begin{cases} o & \text{if } x \in \mathcal{P}(t) \\ +\infty & \text{if not} \end{cases} \tag{5.1}$$

The contingent epiderivative $D_+ V(t\ ,\ x)$ of a function V from $\mathbf{R} \times X$ to $\mathbf{R} \cup \{+\infty\}$ at $(t\ ,\ x)$ in the direction $(alpha\ ,\ v)$ is defined by

$$D_+ V(t\ ,\ x)(a\ ,\ v) : = \liminf_{\substack{h \to 0+ \\ w \to v \\ \beta \to a}} \frac{V(t + \beta h\ ,\ x + hw) - V(t\ ,\ x)}{h} \tag{5.2}$$

The epigraph of $D_+ V(t\ ,\ x)$ is the contingent cone to the epigraph of V at $(t\ ,\ x\ ,\ V(t\ ,\ x))$. Hence, conditions (1.5) can be translated in the following way:

Proposition 2

A tube $\mathcal{P}$ is a viability tube if and only if the indicator function $V_{\mathcal{P}}$ of its graph is a solution to the "contingent Hamilton-Jacobi" equation.

$$\inf_{v\ \in F(t\ ,\ x)} D_+ V_{\mathcal{P}}\ (t\ ,\ x)(1\ ,\ v) = 0 \tag{5.3}$$

satisfying the final condition (when $T < \infty$):

$$\inf_{v\ \in F(t\ ,\ x)} D_+ V_{\mathcal{P}}\ (T\ ,\ x)(0\ ,\ v) = 0 \tag{5.4}$$

Remark

When the function V is differentiable, equation (5.3) can be written in the form

$$\frac{\partial V}{\partial t} + \inf_{v\ \in F(t\ ,\ x)} \sum_{i=1}^{n} \frac{\partial V}{\partial x_i}(t\ ,\ x) v_i = 0$$

We recognize the classical Hamilton-Jacobi equation (see Aubin-Cellina [1984], Chapter 6). A thorough study of contingent Hamilton-Jacobi equations (for Lipschitz maps $F(t\ ,\ x)$) is carried out in Frankowska [1986]), where relations with viscosity solutions introduced by Crandall-Lions P.L. [ ] (see also Lions P.L. [1982]) and generalized Hamilton-Jacobi equations (Clarke-Vinter [1983], Rockafellar [to appear])

6. Invariant tubes

We distinguish between viability tubes and invariant tubes in the same way as viability domains and invariant domains.

Definition 3.

We say that a tube $\mathcal{P}$ enjoys the invariance property if and only if for all t_0 and $x_0 \in \mathcal{P}(t_0)$, all the solutions to the differential inclusion

$$x'(t) \in F(t\ ,\ x(t)) \tag{6.1}$$

are viable in the tube $\underline{P}$.

We say that $\underline{P}$ is an "invariant tube" if

$$\begin{cases}(i)\ \forall t \in [O\ ,\ T[\ ,\ \forall x \in \underline{P}(t)\ ,F(t\ ,x) \subset D\underline{P}(t\ ,x)(1) \\ (ii)\ \text{if } T < +\infty\ , \quad \forall x \in \underline{P}(T), F(t\ ,x) \subset D\underline{P}(t\ ,x)(0)\end{cases} \tag{6.2}$$

We obtain the following theorem.

Theorem 5 Assume that $F : [o\ ,\ T\ [\ \times \Omega \to \mathbf{R}^n$ is Lipschitz with respect to x in the sense that

$$\exists\ k(\cdot) \in L^1(o\ ,\ T)\ |\ F(t\ ,x) \subset F(t\ ,y) + k(t) \| x - y \| B \tag{6.3}$$

(B is a unit ball). Let $t \to \underline{P}(t) \subset \Omega$ be a closed tube: If $\underline{P}$ is invariant, then it enjoys the invariance property.

Proof

The theorem follows from the following lemma, an extension to a result from Aubin-Clarke [1977].

Lemma 2

Let $\underline{P}$ be a closed tube and $\Pi_{\underline{P}(t)}(y)$ denote the set of best approximations of y by elements of $\underline{P}(t)$.

$$\begin{cases}\liminf\limits_{h \to 0+} \quad \dfrac{d(y + hv\ ,\underline{P}(t + h)) - d(y\ ,\underline{P}(t))}{h} \\ \leq \inf\limits_{x \in \Pi_{\underline{P}(t)}(y)} \quad d(y\ ,\ D\underline{P}(t\ ,x)(1)))\end{cases} \tag{6.4}$$

Then, with any solution to the differential inclusion $x' \in F(t\ ,x)$, we associate the function $g(t) := d(x(t)\ ,\underline{P}(t))$

let us choose $y(t) \in T_{\underline{P}(t)}(x(t))$. Inequalities

$$\frac{g(t + h) - g(t)}{h}$$

$$= \frac{d(x(t) + hx'(t) + ho(h)\ ,\underline{P}(t + h)) - d(x(t)\ ,\underline{P}(t))}{h}$$

$$\leq \| o(h) \| + \frac{d(x(t) + hx'(t)\ ,\underline{P}(t + h)) - d(x(t)\ ,\underline{P}(t))}{h}$$

$$\leq d(x'(t)\ ,\ D\underline{P}(t\ ,y(t)(1))$$

$$\leq d(x'(t)\ ,\ F(t\ ,y(t)))$$

$$\leq \sup_{v \in F(t\ ,x(t))} \quad d(v\ ,F(t\ ,y(t)))$$

$$\leq k(t) \| y(t) - x\ (t) \| = k(t)\, d(x(t)\ ,\underline{P}(t))$$

$$= k(t)\, g(t)$$

imply that g(t) is a solution to the differential inequality.

$$D_+ g(t)(1) \leq k(t) g(t)\ ;\ g(t_0) = d(x_0\ ,\underline{P}(t_0)) = 0$$

Hence $d(x(t)\ ,\underline{P}(t)) = g(t) = 0$ for all $t \in [t_0\ ,\ T[$.

Proof of Lemma 2

(a) Let $y \in \underline{P}(t)$ and $u \in D\underline{P}(t\ ,y)(1)$ be given. We consider sequences $h_n \to 0+$ and $u_n \to u$ such that

$$\liminf_{n \to \infty} \frac{d(y + h_n\ u_n\ ,\underline{P}(t + h_n))}{h_n} = 0$$

Hence, for all $v \in X$,

$$\frac{1}{h} d(y + h_n v , \underline{P}(t + h_n)) \leq \| v - u_n \| + \frac{d(y + h_n u_n , \underline{P}(t + h_n))}{h_n}$$

which implies the desired inequality by letting h_n go to o.

(b) Let us choose $y \notin \underline{P}(t)$ and $x \in \underline{P}(t)$ such that $\| x - y \| = d(y , \underline{P}(t))$. We observe that

$$\frac{1}{h} d(y + hv , \underline{P}(t + h)) - d(y , \underline{P}(t))$$
$$\leq \frac{1}{h} (\| y - x \| + d(x + hv , \underline{P}(t + h)) - d(y , \underline{P}(t))$$
$$= \frac{1}{h} d(x + hv , \underline{P}(t + h)) .$$

Since x belongs to $\underline{P}(t)$, the desired inequality for x implies the one for y since

$$\liminf_{h \to 0+} \frac{1}{h}(d(y + hv , \underline{P}(t + h) - d(y , \underline{P}(t)))$$
$$\leq \liminf_{h \to 0+} \frac{1}{h} (d(x + hv , \underline{P}(t + h))$$
$$\leq d(v , D\underline{P}(t , x)(1))$$

Remark

This lemma implies that if

$$\forall t , \forall x \in \underline{P}(t) , F(t , x) \subset D\underline{P}(t , x) \tag{6.5}$$

and if

$$\forall t , x \to F(t,x) \text{ is lower semicontinuous,} \tag{6.6}$$

then

$$\forall t , \forall x \in \underline{P}(t) , F(t , x) \subset C\underline{P}(t , x)(1) \tag{6.7}$$

where

$$v \in C\underline{P}(t , x)(1)$$

if and only if

$$\lim_{\substack{h \to 0+ \\ y \to x \\ \underline{P}(t)}} \frac{d(y + hv , \underline{P}(t + h))}{h} = 0$$

This convergence is uniform with respect to $v \in F(t , x)$ if this subset is compact. In particular, if

$$x \to D\underline{P}(t , x)(1) \tag{6.8}$$

is lower semicontinuous then

$$D\underline{P}(t , x)(1) = C\underline{P}(t , x)(1) \tag{6.9}$$

Remark

If we assume that the condition

$$\begin{cases}\forall (t\ ,\ y) \in \mathrm{Dom}\, F\ ,\ \exists\, x\ \in \Pi_{\underline{P}(t)}(y) \text{ such that} \\ F(t\ ,\ y) \subset D\underline{P}(t\ ,\ x)(1)\end{cases} \tag{6.10}$$

then the tube $\underline{P}$ is invariant by F: this knowledge of the behavior of F outside the graph of the tube $\underline{P}$ allows to dispose of the Lipschitz assumption.

We can characterize the indicator functions of the graphs of invariant tubes in the following way.

Proposition 3

A tube $\underline{P}$ is invariant by F if and only if the indicator function of its graph is a solution to the equation

$$\sup_{v\ \in F(t\ ,\ x)} D_+ V_{\underline{P}}(t\ ,\ x)(1\ ,\ v) = 0 \tag{6.11}$$

satisfying the final condition

$$\text{if } T < +\infty\ ,\ \sup_{v\ \in F(t\ ,\ x)} D_+ V_{\underline{P}}(T\ ,\ x)(o\ ,\ v) = 0 \tag{6.12}$$

7. Duality relations between invariant and viability tubes

Let us consider the case when $F(t\ ,\ x) : =A(t)x$ is a time dependent closed convex process $A(t)$ whose domain is the whole space X. In this case, we look for tubes R the images of which are closed convex cones. We associate to the tube R its "polar tube" R^+ associating with any t the (positive) polar cone

$$R(t)^+ : = \{q\ \in X^*\ |\ \forall y\ \in R(t)\ ,\ <q\ ,\ y\ > \geq 0\} \tag{7.1}$$

We also associate with $A(t)$ its "transpose" $A(t)^*$ defined by

$$\begin{cases} P\ \in A(t)^*\ q \iff \\ \forall (x\ ,\ y) \in Graph\ A\ (t)\ ,\ <p\ ,\ x\ > \leq\ <q\ ,\ y\ > \iff \\ (-p\ ,\ q) \in (Graph\ A\ (t))^+ . \end{cases} \tag{7.2}$$

We consider the "linear" differential inclusion

$$x'(t) \in A(t)\, x(t) \tag{7.3}$$

and its "adjoint" differential inclusion

$$-p'(t) \in A(t)^*\, p(t) \tag{7.4}$$

We shall prove that the invariance of the tube R implies that its positive polar tube R^+ is a viability tube of the adjoint inclusion.

THEOREM 6

Let us assume that the domains of the closed convex processes are all equal to X and that

$$\begin{cases} (i) & \text{the lipschitz constants of } A(t) \text{ are less than} \\ & \text{or equal a function of } k(\cdot) \text{ of } L^2(o\ ,\ T) \\ (ii) & (t\ ,\ q) \to \sigma(A(t)x\ ,\ q) \text{ is lower semicontinuous} \\ & \text{for all } x\ \in X \end{cases} \tag{7.5}$$

Let R be a tube with closed convex cone values. If R enjoys the viability property for $A(t)$, then the tube R^+ is a viability tube of the adjoint differential inclusion and thus, it enjoys the viability property in the sense that $\forall t \in [o\ ,\ T]$, $\forall q \in R(t)^+$, there exists a solution q to the adjoint inclusion such that $q(t) = q$ and $q(\tau) \in R(\tau)^+$ for all $\tau \in [o\ ,\ t]$.

Proof

We have to prove that

$$\forall t \in [o\ ,\ T]\ ,\ \forall q_t \in R(t)^+$$

$$A(t)^* q_t \cap DR^+(t\ ,\ q_t)(-1) \neq \phi \tag{7.6}$$

Since the transpose $A(t)^* q$ is upper semicontinuous with compact convex images, Theorem 1.1 will imply that R^+ enjoys the viability property. Let $S \subset H^1(o\ ,\ T\ ;\ X)$ be the set of solutions to the differential inclusion $x'(t) \in A(t)\,x(t)$. We denote by γ_τ the linear operator from $H^1\ (o\ ,\ T\ ;\ X)$ to X associating with every x its value $\gamma_\tau x\ :=x(\tau)$ at $\tau \in [o\ ,\ T]$.

To say that R enjoys the invariance property means that for all $o \leq s \leq t \leq T$,

$$\gamma_t\ (S \cap \gamma_s^{-1} R(s)) \subset R(t)$$

By polarity, we deduce that

$$R(t)^+ \subset (\gamma_t(S \cap \gamma_s^{-1}\ (R(s)))^+ = \gamma_t^{*-1}\ (S \cap \gamma_s^{-1}(R(s)))]^+$$

We deduce from Frankowska [1986a] that

$$(S \cap \gamma_s^{-1} R(s))^+ = S^+ + \gamma_s^* R(s)^+$$

Hence, for all $q_t \in R(t)^+$ and for all $s \leq t$, there exists $q_s \in R(s)^+$ such that $\gamma_t\ q_t - \gamma_s\ q_s$ belongs to S^+. Always by Frankowska [1986a], there exists a solution p_s to the adjoint inclusion on the interval $[s\ ,\ t]$

$$-p_s{}'(\tau) \in A(\tau)^* p_s(\tau)\ ;\ p_s(t) = q_t\ . \tag{7.7}$$

which satisfies

$$p_s(s) \in R(s)^+$$

By replacing t by s and s by o, we can extend the solution $p_s(\cdot)$ on the whole interval $[o\ ,\ t]$. We now let s converge to t. Since *Dom* $A(t) = X$, we know that

$$\sigma(A(t)^* p\ ,\ x) = -\sigma(A(t)x\ ,\ -p)$$

Hence the lower semicontinuity of $(t\ ,\ p) \to \sigma(A(t)\,x\ ,\ -p)$ implies the upper semicontinuity of $\sigma(A(t)^* p\ ,\ x)$, and thus, the upper semicontinuity of $(t\ ,\ p) \to A(t)^* p$. (See Aubin-Ekland, [1984], Theorem 3.2.10). Therefore for all $\varepsilon > 0$, there exists $\eta > 0$ such that, for all $\tau \in [t - \eta\ ,\ t]$ and $p \in q_t + \eta B$,

$$A(\tau)^* p \subset A(t)^* q_t + \varepsilon B$$

The set of solutions p_s to the adjoint inclusion being contained in a compact set of $C(o\ ,\ T\ ;\ X)$, a subsequence (again denoted) p_s converges uniformly to a solution p_o to the adjoint equation. Hence there exists $\alpha \leq \eta$ such that, for all $\tau \in [t - \alpha\ ,\ t]$, and for all s , $\| P_s(\tau) - q_t \| \leq \eta$. Therefore

$$\forall s\ ,\ \forall \tau \in [t - \alpha\ ,\ t]\ ,\ A(\tau)^* p_s(\tau) \subset A(t)^* q_t + \varepsilon B$$

By integrating (7.7) on the interval $[t - h\ ,\ t]$ with $s = t - h\ ,\ h \leq \alpha$, we deduce that

$$v_h : = \frac{p_{t-h}(t - h) - q_t}{h} = \frac{1}{h} \int_{t-h}^{t} p_{t-h}'(\tau) d\tau$$

$$\in -\frac{1}{h} \int_{t-h}^{t} A(\tau)^* p_s(\tau)\, d\tau \subset - \overline{co}\ (A(t)^* q_t + \varepsilon B)$$

$$-A(t)^* q_t + \varepsilon B$$

This subset being compact, a subsequence v_n converges to an element $v \in A(t)^* q_t$. Since $q_t + hv_n = p_{t-h}(t - h) \in R(t - h)^+$ for all $h > o$, we deduce that v belongs to $DR^+(t\ ,\ q_t)(-1)$.

8. Examples of viability tubes

Let us consider two closed subsets C and D of $\mathbf{R}^n$ and differentiable map Φ from a neighborhood of $[0\ ,\ T] \times C \times D$ to $\mathbf{R}^n$. We consider tubes of the form

$$P(t) : = \Phi(t\ ,\ C\ ,\ D) \tag{8.1}$$

Proposition 3: Let us assume that

$$\forall t \leq T\ \forall x \in P(t)\ ,\ \exists (y\ ,\ z) \in C \times D \text{ such that } \Phi(t\ ,\ y\ ,\ z) = x\ ,$$

$$\exists (u\ ,\ v) \in T_{C \times D}\ (y\ ,\ z) \text{ such that}$$

$$\begin{cases} (i)\ if\ \ t < T\ ,\ \Phi_y'(t\ ,\ y\ ,\ z)u\ +\ \Phi_z'(t\ ,\ y\ ,\ z)v\ \in F(t\ ,\ x) - \Phi_t'(t\ ,\ y\ ,\ z) \\ (ii)\ if\ \ t\ = T\ ,\ \Phi_y'(T\ ,\ y\ ,\ z) +\ \Phi_z'(T\ ,\ y\ ,\ z)v\ \in F(T\ ,\ x) \end{cases} \tag{8.2}$$

Then the set-valued map P defined by (8.1) is a viability tube of F on $[0\ ,\ T]$.

Proof: We observe that Graph (P) is the image of $[0\ ,\ T] \times C \times D$ under the map Ψ defined by $\Psi(t\ ,\ y\ ,\ z) = (t\ ,\ \Phi(t\ ,\ y\ ,\ z))$.

By Proposition 7.6.2, p. 430 of Aubin-Ekland [1984],

$$\Psi'\ (t\ ,\ y\ ,\ z)T_{[0\ ,\ T] \times C \times D}\ (t\ ,\ y\ ,\ z) \subset T_{Graph\,(P)}\ (\Psi(t\ ,\ y\ ,\ z))\ .$$

we deduce that conditions (8.2) imply property.

When C and D are closed and convex, we can characterized viability tubes of the form (8.1) through dual conditions. If K is a subset of $\mathbf{R}^n$, we denote by

$$\sigma(K\ ,\ p) : = \sup_{x \in K}\ <p\ ,\ x> \tag{8.3}$$

its support function.

Proposition 4: Let us assume that the values of F are compact and convex and that the subsets C and D are closed and convex. If for any $t \in [0\ ,\ T]\ ,\ \forall x \in P(t)$, there exists $(y\ ,\ z) \in C \times D$ satisfying $\Phi(t\ ,\ y\ ,\ z) = x$ and for all

$$p \in \Phi_y'(t\ ,\ y\ ,\ z)^{*-1} N_C(y) \cap \Phi_z'(t\ ,\ y\ ,\ z)^{*-1} N_D(z)\ ,$$

we have

$$\begin{cases} (i)\ \ \ \forall t < T\ ,\ <p\ ,\ \Phi_t'(t\ ,\ y\ ,\ z) >\ +\ \sigma(F(t\ ,\ \Phi(t\ ,\ y\ ,\ z))\ ,\ -p \geq 0 \\ (ii)\ \ \text{for } t = T\ ,\ \sigma(F(T\ ,\ \Phi\ (T\ ,\ y\ ,\ z))\ ,\ -p) \qquad\qquad \geq 0 \end{cases} \tag{8.4}$$

then the set-valued map P defined by (8.1) is a viability tube of F on $[0\ ,\ T]$.

Proof: When C and D are convex, $T_{C\times D}\ (y\ ,\ z) = T_C(y) \times T_D(z)$ so that conditions (8.2)i) and ii) can be written

$$\begin{cases}(i)\ F(t\ ,x) - \Phi_t'(t\ ,y\ ,z)) \cap (\Phi_y'(t\ ,y\ ,z)\,T_C(y) + \Phi_z'(t\ ,y\ ,z)T_D(z)) \neq \phi \\ (ii)\qquad F(T\ ,x)\qquad \cap (\Phi_y'(T\ ,y\ ,z)T_C(y) + \Phi_z'(T\ ,y\ ,z)T_D(z)) \neq \phi\end{cases} \qquad (8.5)$$

The separation theorem shows that they are equivalent to conditons (8.4).

Corollory 1: Let us assume that C and D are closed convex subsets and that the values of F are convex and compact. Let $\Phi : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ be a differentiable function satisfying either one of the following equivalent conditions:

For any $t \geq 0\ ,\ \forall,\ x\ ,\ \in P(t)$, there exist $y\ \in C\ ,\ z\ \in D$ such that $x\ = y\ +\ \Phi(t)z$ and either

$$\begin{cases}(i)\quad (F(t\ ,\ y\ +\ \Phi(t)z) - \Phi'(t)z) \cap (T_c(y) + T_D(z)) \neq \phi \text{ if } t < T \\ (ii)\qquad (F(T\ ,\ y\ +\ \Phi(T)z) \qquad \cap (T_C(y) + T_D(z)) \neq \phi \text{ if } t = T\end{cases} \qquad (8.6)$$

or

$$\begin{cases}\qquad \forall p\ \in N_C(y) \cap N_D(z)\ , \\ (i)\qquad \Phi'\ (t)\ \sigma_D(p) + \qquad \sigma(F\ (t\ ,\ y\ +\ \Phi(t)z\ ,\ -p) \geq 0 \text{ if } t < T \\ (ii)\qquad\qquad \sigma(F\ (T\ ,\ y\ +\ \Phi(T)z\ ,\ -p) \geq 0 \text{ if } t = T\end{cases} \qquad (8.7)$$

Then the set-valued map P defined by

$$P(T) : = C\ +\ \Phi(t)\ D \qquad (8.8)$$

is a viability tube of F on $[0\ ,\ T]$.

Let us consider the instance when $C\ = \{c\}$ and when 0 belongs to the interior of the closed convex subset D.

We introduce the function a_0 defined by

$$\begin{cases} a_0(t\ ,\ w) : = \\ \sup\limits_{z\ \in D}\quad \sup\limits_{\substack{p\ \in N_D(z) \\ \sigma_D(p)\ =\ 1}}\quad \inf\limits_{v\ \in F(t\ ,\ c\ +\ w\ z)}\quad <p\ ,\ v> \\ = \sup\limits_{z\ \in D}\quad \inf\limits_{v\ \in F(t\ ,\ c\ +\ w\ z)}\quad \sup\limits_{\substack{p\ \in N_D(z) \\ \sigma_D(p)\ =\ 1}}\quad <p\ ,\ v> \end{cases} \qquad (8.9)$$

(The last equation follows from the minimax theorem.)

Let us assume that there exists a continuous function $a\ : \mathbf{R}_+ \times \mathbf{R}_+ \rightarrow \mathbf{R}$, satisfying $a\,(t\ ,\ 0)\ = 0$ for all $t\ \geq 0$, such that

$$\forall (t\ ,\ w) \in \mathbf{R}_+ \times \mathbf{R}_+\ ,\ a\,(t\ ,w) \geq a_0(t\ ,\ w) \qquad (8.10)$$

Let Φ be a solution to the differential equation

$$\Phi'(t)\ = a\,(t\ ,\ \Phi\ (t))\ ,\ \Phi(0)\ = \Phi_0 \text{ given} \qquad (8.11)$$

satisfying

$$a\,(T\ ,\ \Phi(T)) = 0 \qquad (8.12)$$

Since $\sigma_D(p) > 0$ fpr all $p\ \neq 0$, we deduce that for all $z\ \in D$ and all $p\ \in N_D(z)$,

$$\Phi'(t)\ \sigma_D(p) \geq a\,(t\ ,\ \Phi(t))\ \sigma_D(p) \geq a_0(t\ ,\ \Phi(t))\ \sigma_D(p)$$

$$\geq \sigma_D(p) \sup_{v \in F(t\ ,\ c\ +\ \Phi(t)z)} < -\frac{p}{\sigma_D(p)}\ ,\ v >$$

$$= -\sigma(F(t\ ,\ c\ +\ \Phi(t)z)\ ,\ -p)$$

Hence, condition (8.7)i) is satisfied. Also

$$0 = a(T\ ,\ \Phi(T)) \geq a_0(T\ ,\ \Phi(T)) \geq \frac{-1}{\sigma_D(p)}\sigma(F(T\ ,\ c\ +\ \Phi(T)z)\ ,\ -p)$$

Then

$$P(t) : = c\ +\ \Phi(t)D \tag{8.13}$$

defines a viability tube of F.

For instance, if $D\ :\ = B$ is the unit ball, then $\sigma_B(p) = \| p \|$ and $N_B(z) = \lambda z$ for all $z\ \in S\ :\ = \{x\ |\ \| x \| = 1\}$. Hence, in this case we have

$$a_0(t\ ,\ w) : = \sup_{\| z \| = 1}\ \inf_{v\ \in F(t\ ,\ c\ +\ wz)} < v\ ,\ z > \tag{8.15}$$

In other words, the function a_0 defined by (8.9) conceals all the information needed to check whether a given subset D can generate a tube P.

Remark: When a is non-positive and satisfies $a(t\ ,\ 0) = 0$ for all $t \geq 0$, then there exists a non-negative non-increasing solution $\Phi(\cdot)$ of the differential equation (8.11).

When $T = \infty$, we infer that $\int_0^\infty a(\tau\ ,\ \Phi(\tau))d\tau$ is finite. Let us assume that for 0 all $w_* \in \mathbf{R}_+$,

$$\lim_{\substack{t \to \infty \\ w \to w_*}} a(t\ ,\ w) = a_*(w_*) \tag{8.16}$$

Then the limit Φ_* of $\Phi(t)$ when $t \to \infty$ satisfies the equation

$$a_*(\Phi_*) = 0$$

Otherwise, there would exist $\varepsilon > 0$ and T such that $a_*(\Phi_*) + \varepsilon < 0$ and for all $t\ > T\ ,\ a(t\ ,\ \Phi(t)) \leq a_*(\Phi_*) + \varepsilon$ by definition of a_*.

We deduce the contradiction

$$\Phi(t) = \Phi(T) = \int_T^t a(\tau\ ,\ \Phi(\tau))d\tau \leq (t\ -\ T)(a_*(\Phi_*)\ +\ \varepsilon)$$

when t is large enough.

Example: Let us consider the case when F does not depend upon t. We set

$$\rho_0 : = \sup_{\lambda \in \mathbf{R}}\ \inf_{w\ >\ 0} (\lambda w\ -\ a_0(w)) \tag{8.17}$$

Assume also that $\lambda_0 \in \mathbf{R}$ achieves the supremum. We can take $\psi(w) : = \lambda_0 w\ -\rho_0$.
If $\rho_0 > 0$, the function

$$\Phi_T(t) : = \begin{cases} \dfrac{\rho_0}{\lambda_0}(1\ -\exp\ (\lambda_0\ (t\ -\ T) & \text{if } \lambda_0 \neq 0 \\ -\rho_0(t\ -T) & \text{if } \lambda_0 = 0 \end{cases} \tag{8.18}$$

is such that $P(t) : = \{c + \Phi_T(t) D\}$ is a pipe of F such that $P(t) = \{c\}$. If $\rho_0 \leq 0$ and $\lambda_0 < 0$, then the functions

$$\Phi(t) : \frac{1}{\lambda_0} (\rho_0 - e^{\lambda_0 t}) \tag{8.19}$$

are such that $P(t) : = c + \Phi_C(t)D$ defines a pipe of F on $[0 , \infty[$ such that $P(t)$ decreases to the set $P_\infty : = c + \frac{\rho_0}{\lambda_0} D$.

9. Tubes derived from potential functions

Let $K \subset \mathbf{R}^n$ be the viability domain and let us consider a "potential function" V from $\mathbf{R}^n$ to $\mathbf{R}_+ \cup \{+ \infty\}$. We shall study in this section tubes of the form

$$P(t) : = \{x \in K \mid V(x) \leq w(t)\} \tag{9.1}$$

where w is a non-negative function defined on $[0 , T]$. We shall begin by providing sufficient conditions on K , V , w and F implying that set-valued maps P of the form (9.1) are viability tubes of F. We obtain in this case the following result:

Proposition 5: Let us assume that K is closed and that V is locally Lipschitz around K. Let w be a C^1- function defined on a neighborhood of V such that

$$\begin{cases} \forall t \in [0 , T[\text{ , the elements of } P(t) \text{ are not critical points of } V \text{ on } K \end{cases} \tag{9.2}$$

We posit the following condition

$$\begin{cases} \forall t \in [0 , T[\text{ , } \forall x \in K \text{ such that } V(x) = w(t) \text{ ,} \\ \exists u \in F(t , x) \cap C_K(x) \text{ such that } C_+ V(x)(u) \leq w'(t) \end{cases} \tag{9.3}$$

and

$$\begin{cases} \text{If } x \in K \text{ satisfying } V(x) = w(T) \text{ is a critical} \\ \text{point of } V \text{ on } K \text{ , then } 0 \in F(T , x) \text{ .} \end{cases} \tag{9.4}$$

Then the set-valued map P defined by (9.1) is a viability tube of F on $[0 , T]$.

We shall also study tubes of the form

$$P(t) : = \{x \in K \mid w_-(t) \leq V(x - c(t)) \leq w_+(t)\} \tag{9.5}$$

where c is a function from $[0 , T]$ to K and w_- and w_+ are non-negative functions, which define some kind of neighborhood around a function $t \to c(t)$, such as periodic trajectories of the dynamical system (1.1). They are special cases of tubes associated to p potential functions V_i by the formula

$$P(t) : = \{x \in K \mid V_i(\Phi(t , x)) \leq w_i(t) , \quad i = 1, \ldots , p)\} \tag{9.6}$$

where Φ is a smooth map from $[0 , T] \times K$ to Dom $\vec{V}$. We shall then provide sufficient conditions on Φ , $\vec{w}$ and the functions V_i for a set-valued P of this type to be a pipe for a given set-valued map F.

Theorem 7: Let us assume that K is closed, that Φ is C^1 around $[0 , T] \times K$, that $\vec{w}$ is C^1 around $[0 , T]$ and that the p potential functions V_i are locally Lipschitz on a neighborhood of $\Phi([0 ,T] \times K)$.

$$I(t\,,\,x) : = \{i\ = 1\,,\,\ldots\,,\,p\ \mid V_i(\Phi(t\,,\,x)) = w_i(t)\} \tag{9.7}$$

We assume that

$$\left\{\begin{array}{l} \forall t\ \in [0\,,\,T]\,,\ \forall x\ \in P(t)\,, \\ 0 \notin \ \mathrm{co}\,[\displaystyle\bigcup_{i\ \in I\ (t\,,\,x)} \Phi_x{}'(t\,,\,x)^*\ \partial\, V_i(\Phi(t\,,\,x))] + N_K(x) \end{array}\right. \tag{9.8}$$

and that

$$\left\{\begin{array}{l} 0\ \in F(t\,,\,x) \text{ for all } x\ \in P(T) \text{ such that there exists} \\ i\ \in I(T\,,\,x) \text{ such that } 0\ \in\ \Phi_x{}'\,(T\,,\,x)^*\ \partial V_i(\Phi\,(T\,,\,x)) + N_K(x) \end{array}\right. \tag{9.9}$$

We posit the following assumption

$$\left\{\begin{array}{ll} (i) & \forall t\ \in [0\,,\,T[\,,\ \forall x\ \in P(t)\,,\ \exists u\ \in F(t\,,\,x) \cap C_K(x) \text{ such that} \\ & \forall i\ \in I(t\,,\,x)\,,\ C_+\,V_i\ \ (\Phi(t\,,\,x))(\Phi_t{}'(t\,,\,x) + \Phi_x{}'(t\,,\,x)u) \leq w_i{}'(t) \\ (ii) & \forall x\ \in P(T)\,,\ \exists u\ \ \in F(T\,,\,x) \cap C_K(x) \text{ such that} \\ & \forall i\ \in I(T\,,\,x)\,,\ C_+\,V_i\ \ (\Phi(T\,,\,x))(\Phi_x{}'\,(T\,,\,x)u) \leq 0 \end{array}\right. \tag{9.10}$$

Then the set-valued map P defined by (9.10) is a viability tube of F.

Remark: Observe that the elements $x\ \in K$ satisfying

$$0\ \in\ \Phi_x{}'(t\,,\,x)^*\ \partial V_i(\Phi(t\,,\,x)) + N_K(x) \tag{9.9}$$

are the critical points of $x\ \to V_i(\Phi(t\,,\,x))$ on K. Assumption (9.9) states that critical points of some functions $V_i(\Phi\ (T\,,\,\bullet))$ on K are equilibria of $F(T\,,\,\bullet)$. We can say that a solution to

$$0 \in co\ \ (\bigcup_{i=1}^{n}\ \Phi_x{}'(t\,,\,x)^*\ \partial V_i(\Phi(t\,,\,x)) + N_K(x) \tag{9.11}$$

is a ***Pareto critical point*** of the functions $V_i(\Phi(t\,,\,\bullet))$, (Pareto minima do satisfy this inclusion).

Corollary 2 Let K be a closed subset, V be a C^1 function from a neighborhood of $[0\,,\,T]$ to K , w_- and w_+ be C^1 non-negative functions satisfying

$$\left\{\begin{array}{l} \forall t\ \in [0\,,\,T[\,,\ 0 \leq w_-(t) < w_-(T) = w_+(T) < w_+(t) \\ \text{and } w_-{}'(t) > 0\,,\ w_+{}' < 0 \end{array}\right. \tag{9.12}$$

We posit the following assumption:

$$\left\{\begin{array}{ll} (i) & \forall t\ \in [0\,,\,T[\,,\ \forall x \text{ such that } V(x\ - c(t)) = w_+(t)\,, \\ & \text{there exists } u\ \in F(t\,,\,x) \cap C_K(x) \text{ such that} \\ & C_+V(x\ - c(t))(u\ - c'(t)) \leq w_+{}'(t) \\ (ii) & \forall t\ \in [0\,,\,T[\,,\ \forall x \text{ such that } V(x\ - c(t)) = w_-(t)\,, \\ & \text{there exists } u\ \in F(t\,,\,x) \cap C_K(x) \text{ such that} \\ & C_-\,V(x\ - c(t))(u\ - c'(t)) \geq w_-{}'(t) \\ (iii) & \forall x \text{ such that } \ V(x\ - c(T)) = w_+(T) = w_-(T)\,,\ 0 \in F(T\,,\,x) \end{array}\right. \tag{9.13}$$

Then the set-valued map P defined by

$$P(t) := \{x \in K \; ; \; w_{-}(t) \leq V(x - c(t)) \leq w_{+}(t)\} \tag{9.14}$$

is a viability tube of F on $[0, T]$.

Proof of Theorem 2.2:

We set $\operatorname{Dom} \vec{V} = \bigcap_{i=1}^{p} \operatorname{Dom} V_i$, $\vec{V}(x) := (V_1(x), \dots, V_p(x))$ and $Ep(\vec{V}) := \{(x, w) \in \operatorname{Dom} \vec{V} \times \mathbf{R}^p \mid V_i(x) \leq w_i \text{ for } i = 1, \dots, p)\}$. Let A be the C^1 map from a neighborhood of $[0, T] \times K$ to $\mathbf{R}^n \times \mathbf{R}^p$ defined by

$$A(t, x) := (\Phi(t, x), w(t)) \tag{9.15}$$

Then we can write

$$\operatorname{Graph}(P) := \{(t, x) \in [0, T] \times K \mid A(t, x) \in Ep(\vec{V})\} \tag{9.16}$$

We then use Proposition 7.6.3, p. 440 of Aubin-Ekeland [1984]. It states that

$$\begin{cases} \{\tau \in T_{[0, T]}(t), u \in T_K(x) \mid A'(t, x)(\tau, u) \in T_{Ep(\vec{V})}(A(t, x))\} \\ \supset T_{Graph(P)}(t, x) \end{cases} \tag{9.17}$$

and that if the transversality condition

$$A'(t, x)(T_{[0, T]}(t) \times C_K(x)) - C_{Ep(\vec{V})}(A(t, x)) = \mathbf{R}^n \times \mathbf{R}^p \tag{9.18}$$

then

$$\begin{cases} \{\tau \in T_{[0, T]}(t), u \in C_K(x) \mid \\ A'(t, x)(\tau, u) \in C_{Ep(\vec{V})}(A(t, x))\} \subset C_{Graph(P)}(t, x) \end{cases} \tag{9.19}$$

Inclusion (9.17) implies that for all $t \in [0, T]$,

$$\begin{cases} DP(t, x) \\ \subset \{u \in T_K(x) \mid \forall i \in I(t, x), D_+ V_i(\Phi(t, x)(\Phi_t'(t, x) + \Phi_x'(t, x) u) \\ \leq w_i'(t)\} \end{cases} \tag{9.20}$$

since

$$A'(t, x)(\tau, u) = (\Phi_t'(t, x)\tau + \Phi_x'(t, x)(u), w'(t)\tau) \quad) \tag{9.21}$$

and since

$$\begin{cases} T_{Ep(\vec{V})}(A(t, x)) = T_{Ep(\vec{V})}(\Phi(t, x), w(t)) \\ = \{(u, \lambda) \in \mathbf{R}^n \times \mathbf{R}^p \mid \forall i \in I(t, x), \lambda_i \geq D_+ V_i(\Phi(t, x)(u)\} \end{cases} \tag{9.22}$$

In the same way, inclusion (9.19) can be rewritten in the following form

$$\begin{cases} \{u \in C_K(x) \mid \forall i \in I(t, x), C_+ V_i(\Phi(t, x))(\Phi_t'(t, x)\tau + \Phi_t'(t, x)u) \\ \leq w_i'(t)\tau\} \subset CP(t, x)(\tau) \subset DP(t, x)(\tau). \end{cases} \tag{9.23}$$

This inclusion and assumption (9.10) imply that P is a viability tube of F. It remains to check the transversality condition (9.18), which can be written in the following way:

$$\forall u_d \in \mathbf{R}^n \,,\ \forall \lambda_d \in \mathbf{R}^p \,,\ \exists\ u \in C_K(x) \,,\ \exists\ \tau \in T_{[0\,,\,T]}(t)$$

such that

$$\begin{cases} \forall i \in I(t\,,x)\,,\ w_i'(t)\tau \geq \\ C_+\ V_i(\Phi(t\,,x)(\Phi'(t\,,x)\tau + \Phi_x'(t\,,x)\,u\ - u_d) + \lambda_d \end{cases} \tag{9.24}$$

By assumption (9.8) and the separation theorem, there exists $\hat{u} \in C_K(x)$ such that

$$\forall i \in I(t\,,x)\,,\ C_+\ V_i(\Phi(t\,,x))\,(\Phi_x'(t\,,x)\hat{u}) < 0 \tag{9.25}$$

There exists η such that $C_+\ V_i(\Phi(t\,,x))\,(\Phi_x'(t\,,x)\hat{u}) + v) \leq 0$ when $v \in \eta B$. Let $\beta = 0$ if $\lambda_d \leq 0$ and

$$\beta > \lambda_d\ /\ \mid C_+ V_i(\Phi(t\,,x))\,(\Phi_x'(t\,,x)\hat{u})\mid \ \text{if}\,\lambda_d > 0\,.$$

We take $\alpha = \beta + \eta \mid \| u_d \|$. Hence, $\tau := 0$ and $u := \alpha\hat{u}$ provide a solution to (9.24).

Then this transversality condition holds true for all $t \in [0\,,T[$ and all $x \in P(t)$. When it fails to be true for some $x \in P(T)$, we then assume that such an x is an equilibrium of $F(T\,,\bullet)$.

References

Aubin, J.-P. and A. Cellina (1984) *Differential Inclusions*, Springer-Verlag (Grundlehren der Math. Wissenschaften, Vol. 264, pp. 1-342)

Aubin, J.-P. and F.H. Clarke (1977) *Monotone Invariant Solutions to differential Inclusions*, J. London Math. Soc., 16, pp. 357-366

Aubin, J.-P. and I. Ekeland, (1984) *Applied Nonlinear Analysis*, Wiley-Interscience

Aubin, J.-P., H. Frankowska and C. Olech (1986) *Controllability of convex processes*, SIAM J. of Control and Optimization

Clarke, F.H. (1983) *Optimization and Nonsmooth Analysis*, Wiley-Interscience

Clarke, F.H. and R.B. Vinter (1983) *Local Optimality Conditions and Lipschitzian Solutions to the Hamilton-Jacobi Equation*, SIAM J. of Control and Optimization, 21(6), pp. 865-870

Clarke, F.H. and R.B. Vinter (1986) *On the Relationship between the Dynamic Programming and the Maximum Principle*, Preprint CRM, Université de Montréal

Crandall, M.G. and P.L. Lions (1983) *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math Soc. 277, pp. 1-42

Crandall, M.G., L.C. Evans and P.L. Lions (1984) *Some Properties of Viscosity Solutions of Hamilton-Jacobi Equation*, Trans. Amer. Math. Soc., 282 (2), pp. 487-502

Frankowska, H. (1986a) *Contingent Hamilton-Jacobi Equations*, IIASA, WP-86—

Frankowska, H. (1986b) *Contingent Cones to Reachable Sets of Control Systems*, Preprint CRM-1381, Université de Montréal

Frankowska, H. (1986c) *Local Controllability of Control Systems with Feedback*, Preprint CRM-1364, Université de Montréal

Frankowska, H. (1986d) *Local Controllability and Infinitesimal Generators of Semigroups of Set-valued Maps,* SIAM J. of Control and Optimization

Frankowska, H. (1986e) *The Maximum Principle for the Differential Inclusions with End Point Constraints,* SIAM J. of Control and Optimization

Haddad, G. and Lasry J. -M. (1983) *Periodic Solutions of Functional Differential Inclusions and Fixed Points of S-Selectionable Correspondences,* J. Math. Anal. Appl.

Haddad, G. (1981) *Monotone Trajectories of Differential with Memory,* Israel J. Math s. 39, pp. 38-100

Kurzhanski, A.B. and T.F. Filippova (1986) *On Viable Solutions for Uncertain Systems,* IIASA, CP-86-11

Kurzhanski, A.B. (1977) *Control and Observation under Conditions of Uncertainty,* Nauka (in Russian)

Kurzhanski, A.B. (1986) *On the Analytical Description of the Viable Solutions of a Controlled System,* Uspekhi Mat. Nauk. 4

Kurzhanski, A.B. (1986) *On the Solution Sets for Uncertain Systems with Phase Constraints,* IIASA, WP-86-11

Lions, P.-L. (1982) *Generalized Solutions of Hamilton-Jacobi Equations,* Pitman

Rockafellar, R.T. (1967) *Monotone Processes of Convex and Concave Type,* Mem. of AMS # 77

Rockafellar, R.T. (1970) *Convex Analysis,* Princton University Press

Rockafellar, R.T. (1979) *La théorie des sous-gradients,* Presses de l'Université de Montréal

Heuristics for Nonlinear Control

Christopher I. Byrnes*
Department of Electrical
and Computer Engineering
Department of Mathematics
Arizona State University
Tempe, Arizona 85287
U.S.A.

Alberto Isidori
Dipartimento di Informatica
ed Sistemistica
Università di Roma
Via Eudossiana, 18
00184 Roma, Italia

* Research partially supported by AFOSR Grant No. 85-0224 and NSF Grant No. ECS85-13099.

1. Introduction

Despite ubiquitous success in the implementation of classical automatic control, there are many pressing needs for the design of more advanced, high performance, real-time command generators. For example, the needs for a significant increase in the accuracy, speed and versatility of robotic manipulators have led to a reexamination of classical (e.g., PD) controllers for DC actuators and an exploration and evaluation of the use of new and more sophisticated control schemes (see, e.g., [1]-[6]). Aside from specific needs to meet more demanding performance requirements, more versatile command generators are now required to fully realize the benefits of the new design options which have been made possible by recent hardware innovations, ranging from devices such as microprocessors to DC motors. Indeed, recent advances in DC motor technology have made the implementation of direct drive actuators for robot arms attractive and feasible: the first of two new kinds of DC motors, based on rare earth cobalt magnets, has already been used in the Carnegie-Mellon direct-drive arm in 1981 and in the MIT direct drive arm (see [7]) in 1982, while a second kind of DC motor is currently being used in the construction of a four degree of freedom robot arm at the ASU Robotics Laboratory. The advent of direct drive actuators will allow robot arm motion which is an order of magnitude faster than previous conventional arms, with end effector speeds of up to 30 feet per second and accelerations of up to 5 to 7 G's, bringing robot motion control out of the quasi static domain and into a more complex dynamic domain.

In order to develop command generators capable of real-time high performance operation in a variation of environments, it becomes necessary to be able to use analysis and design principles which apply to (at least some broad class of) nonlinear systems as well as linear systems. In this paper, we describe a research program which we have been pursuing for the past 3 years, whose goal is the development of heuristics for nonlinear control, similar in scope and spirit to classical control, to be used in the analysis and design of nonlinear feedback control systems. Thus, for example, we wish to develop nonlinear generalizations of some of the concepts familiar from frequency domain theory and to use these, in much the same way as classical control methods, to design and analyze nonlinear versions of PD control, lead-lag compensation, etc., to "shape the response" of nonlinear systems. The methodology we propose in Section 3 is based in part on providing a set of sufficiently powerful system-theoretic heuristics to permit the development of a class of simply structured control laws capable, for example, of stabilizing a nonlinear system given only a crude knowledge of the actual system parameters. In practice, these parameters would typically consist of literally thousands of transcendental functions, rendering the on-line parameter estimation of the system coefficients an extremely unattractive alternative, from the point of view of both rigorous analysis and cost-effectiveness. We also believe that the simpler the design philosophy and the simpler the controller structure, the more likely it is that the controller can be "molded" to fit a particular nonlinear application.

Thus motivated, in a recent series of papers, we have begun the development of analogues, for nonlinear systems, of familiar frequency domain concepts based on a reformulation of classical control theory in a state-space setting. Suggestions of such a reformulation can be found in the use of singular perturbation methods in the analysis of adaptive control or learning systems (see, e.g., Sec. 2.5), as well as in the differential geometric reformulation of (A,B)-invariant subspaces by Isidori et al ([8], see also [9]). Combining these methods with tools from global differential geometry, dynamical systems and PDE's we have further developed the nonlinear formulation and use of familiar concepts such as "minimum phase", "relative degree". Naturally, for linear systems our usage of this terminology agrees with the classical usage. For nonlinear systems, there also are antecedents in the literature for certain of these "frequency domain" notions; e.g., our relative degree plays a fundamental role in Hirschorn's work [10] on system invertibility, where it is called the system relative order. It also plays a crucial role in Freund's intriguing design of computer-controlled nonlinear robotic manipulator, where it is called the system dynamical order. On the other hand, perhaps one of the major technical and novel contributions in this program is the intrinsic definition of (finite) "zero dynamics." For linear systems, the zero dynamics is a canonically associated linear system with natural frequencies precisely at the system transmission zeroes. For nonlinear systems, (finite) zeroes correspond instead to a nonlinear dynamical system whose asymptotic properties determine the stability of closed-loop feedback systems. In this regard, our design philosophy retains much of the classical control intuition. Our definition of "zero dynamics" was inspired by and in the scalar case coincides with the (local) definition given in Isidori-Krener [11], see however Isidori-Moog [12] for a (local) development of the MIMO case.

To illustrate what we have in mind, in section 2 we discuss the PD control of a robotic manipulator, to which we return in Example 5.5. Section 3 contains some of the basic development of the analogues, for nonlinear systems, of certain classical control concepts such as relative degree or minimum phase properties. Section 4 gives a sketch of our program to design, e.g., stabilizing compensators on the basis of our nonlinear enhancement of root-locus methods. Finally, in section 5, we illustrate, in a series of 5 examples, our design methodology. Notable among these examples is the control of rigid satellite motion, using only two actuators, to a revolute motion about a principal axis and a rigorous analysis of PD control of a robotic manipulator, giving in effect a nonlinear version of the Ziegler-Nichols rules.

2. Nonlinear PD control of a Robot Manipulator

For the sake of illustration, we consider the problem of set-point control, i.e., stabilization about the state $(q, \dot{q}) = q_d, 0)$, for the rigid body model of a robotic manipulator.

$$M(q)q + B(q,q)\dot{q} + K(q) = T \tag{2.1}$$

In contrast to path-planning using the method of computed torque (see, e.g., [1] - [2]), which requires explicit knowledge of thousands of nonlinear terms in (2.1), one might expect that a nonlinear PD controller

$$\tau = -K_e(q-q_d) - K_d(\dot{q}) \tag{2.2}$$

could be designed using much less explicit knowledge. Indeed, one of our goals to find systematic means, perhaps using describing function methods together with the "frequency domain" methods developed in section 3, to obtain a nonlinear version of the Ziegler-Nichols rules which would determine appropriate nonlinear functions K , K in (2.2). There are, of course, several existing heuristic derivations of controllers (2.2). For example, one design cited in the robotics literature (see, e.g., [13], [14]) involves cancelling only the gravitational field K(q) and adding as a dissipative term a linear PD controller; e.g.

$$\tau = -[K(q) - (q_d - q)] - \dot{q} \tag{2.3}$$

We note that, in particular, for space or underwater applications the gravitational field can be ignored, in which case (somewhat remarkably) we would be controlling the nonlinear system (2.1) by a fixed parameter linear PD controller. The argument offered in [14] used the Lyapunov function

$$V(q,\dot{q}) = \frac{1}{2}\left(\dot{q}^t M(q)\dot{q} + (q_d - q)^T (q_d - q)\right)$$

for which one can show

$$\dot{V} \leq - \dot{q}^t \dot{q}$$

By LaSalle's invariance principle, each bounded solution tends to the largest invariant set contained in $\dot{q} = 0$, which is simply the equilibrium point

$$(q_d, \dot{q}) = (q_d, 0)$$

Elementary counterexamples show, however, that boundedness cannot be automatically guaranteed; from a dynamical systems point-of-view, it is a question of whether ∞ behaves like a saddle point for (2.1) - (2.2).

The justification ([13], [14]) offered for global stability of (2.1) - (2.2) is the somewhat heuristic belief that "all physical trajectories are bounded".

Certainly, all initial states x_0 have finite energy, but for an actual digital control implementation of (2.3) eventual unboundedness of x_t manifests itself in overflow, saturation, etc. Such closed-loop behaviour of (2.1), (2.3) has, in fact, been observed in simulations of a 2 degree of freedom direct-drive horizontal arm (where one can also neglect the gravitational terms) designed by D. W. Parish (see [15]) at the ASU Robotics Laboratory. Nonetheless, we believe the derivation and justification of (2.3) is appealing, retaining as it does classical control intuition. Our goal is to extend this intuition by regorously developing a set of classically-based heuristics for nonlinear systems, giving a basis for designing the nonlinear analogues of classical control laws (2.2) and for a rigorous analysis of resulting closed-loop behaviour. Since linear state space concepts often generalize more immediately than frequency domain concepts, in this paper we will explicitly describe how to interpret the latter for nonlinear systems. In particular, on the basis of this theory we can give a rigorous analysis of the closed-loop behaviour of feedback systems such as (2.1)-(2.2), see, e.g., Example 5.5.

3. Frequency Domain Methods for Nonlinear Systems

In this section, we illustrate our development of the analogues for nonlinear systems of those frequency domain notions so important in classical control. One of our long-term goals, about which we can say quite a bit in the scalar input - scalar output case, is to develop a design philosophy for the construction of (globally) stabilizing compensators for nonlinear systems. Rather than a dependence, say, on explicit knowledge of the Taylor coefficients, this design philosophy is based on seemingly familiar notions such as the (strong) relative degree of a nonlinear system, or knowledge that a nonlinear system is "minimum phase". And, based on such knowledge, we design classical compensators, e.g., leadlag compensators, which we show stablize the system (globally) for initial data in any given bounded open subset of state-space. Thus, these algorithms are designed to achieve set-point control of initial states with an a priori bounded "energy". This gives us a rigorous version of the often appealed to heuristic belief that "all physical trajectories are bounded".

As a first step, we formulate several definitions which are the nonlinear analogues of the linear notions of left or right half plane zeroes and of zeroes at infinity. For simplicity, these definitions are given in the scalar real analytic case. The appropriate multivariable definitions are given in Isidori-Moog ([12], these proceedings). We consider then real analytic systems evolving on a real analytic manifold M of dimension n. Thus, in local coordinates, such a system is described by

$$\dot{x} = f(x) + ug(x) \tag{3.1a}$$

$$y = h(x) \tag{3.1b}$$

Denoting the Lie derivative of a function F with respect to a vector field V by $L_V F$, we formulate

Definition 3.1. The system (3.1) has a zero at infinity of multiplicity ν_∞ if

$$L_g h(x) = L_g L_f h(x) = \dots = L_g L_f^{\nu_\infty - 2} h(x) = 0 \tag{3.2}$$

and

$$L_g L_f^{\nu_\infty - 1} h(x) \neq 0 . \tag{3.3}$$

For a linear system

$$\dot{x} = Ax + ub \tag*{(3.3a)'}$$

$$y = cx \tag*{(3.1b)'}$$

one computes $L_b L_{Ax}^r cx = cA^r b$. For this reason, we shall also call ν_∞ the relative degree of (3.1).

Definition 3.2. The system (3.1) has strong relative degree ν_∞ provided it has a zero at infinity of multiplicity ν_∞ and $L_g L_f^{\nu_\infty - 1} h$ never vanishes.

Turning to the multiplicity of "finite zeros", denote by Δ^* the maximal locally (f,g)-invariant distributin contained in ker(dh) (see [8], [9]).

Definition 3.3. The system (3.1) has finite zero dynamics of orver ν_f provided

$$\nu_f = \dim \Delta^* ,$$

where dimension is understood in the generic sense.

We note, for example, that just as in the linear case,

$$\nu_\infty + \nu_f = n.$$

Example 3.1. (A local form of systems of relative degree 1.) To say $\nu_\infty = 1$ is to say there exists $x_0 \in M$ such that $L_g h(x_0) \neq 0$. In particular, $\Delta^* = \ker(dh)$. Thus, there exists a coordinate chart $(x_1, \dots, x_n)$, centered at x_0 and defined on a neighborhood U of x_0, such that

(i) $\Delta^* + \text{span}(g) = T_X(U),\ x \in U;$

(ii) $\Delta^* = \text{span}\ \{\frac{\partial}{\partial x_1}, \ldots, \{\frac{\partial}{\partial x_{n-1}}\}$;

(iii) $\text{span}\ \{g\} = \text{span}\ \{\frac{\partial}{\partial x_n}\}.$

In these coordinates, setting $z = \frac{\partial}{\partial x_n}$, (3.1) takes the form

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

$$\dot{x}_n = f_2(z,x_n) + ug_2(z,x_n)$$

$$y = h(x_n)$$

In the light of the third equation, the second equation may be replaced by

$$\dot{y} = f_2(z,y) + ug_2(z,y)$$

where, of course, $f_2 = L_f h$ and $g_2 = L_g h$. Therefore, (3.1) can be expressed as

$$\dot{z} = f_1(z,y) \tag{3.4}$$

$$\dot{y} = L_f h(z,y) = uL_g h(z,y).$$

In this setting, the zero dynamics is the (n-1)-th order system

$$\dot{z} = f_1(z,0). \tag{3.5}$$

Remark: In the linear case, a straight forward Laplace transform argument shows that the linear system (3.5) has its spectrum precisely at the original system zeros. Thus, our definition of zeroes does not correspond to a set of complex frequencies, but rather to a dynamical system which, in the linear case, has the zero locus as its set of natural frequencies. We now proceed to give a global, coordinate free definition for $\nu_\infty \geq 1$.

Suppose x_e is an isolated equilibrium point for (3.1) and suppose (3.1) has strong relative degree ν_∞. Without loss of generality we can assume $h(x_e) = 0$. If $\nu_\infty = 1$, then $\Delta^* = \ker(dh)$ and we can consider the leaf

$$L(x_e) = h^{-1}\ (0) \tag{3.6}$$

of Δ^* containing x_e, see Figure 3.1. As in the linear case, the "zero dynamics" should correspond to the drift, $f(x)$, in (3.1) constrained to the locus (3.2.6) as Figure 3.1.

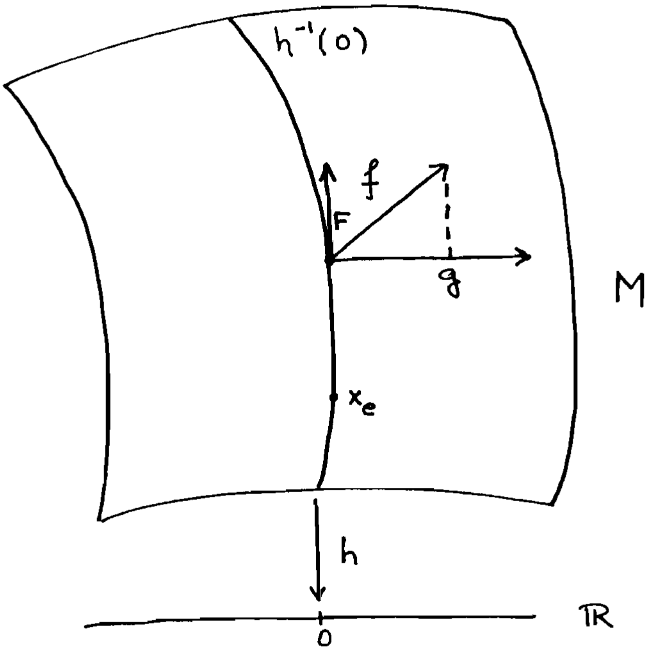


Figure 3.1 Constraining the drift term to $h^{-1}(0)$

Motivated by Euclidean geometry, we proceed formally, first defining the constrained vector field F via

$$F = f - \frac{\langle g,f\rangle}{\langle g,g\rangle} g \tag{3.7}$$

where, however, the inner product (or Riemannian metric) $\langle,\rangle$ is of course not defined, either intrinsically or extrinsically. Indeed, we will instead think of the 1-form dh as being "dual" to the vector field g since

$$\langle dh,g\rangle = L_g h \neq 0$$

and interpret (3.7) as the following definition, which does make intrinsic sense,

$$F = f - \frac{L_f h}{L_g h} g \tag{3.7)'$$

If $\nu_\infty = 1$, taking into account the definition of F and (3.4), we see that (3.5) is clearly the expression, in local coordinates, of the vector field $F|_{L(x_e)}$. of the zero dynamics. As in [17,18], we have chosen u(x) so as to constrain the dynamics (3.1a) to the locus (3.6), i.e. u(x) is chosen so as to satisfy

$$\langle dh, f(x) + u(x)g(x)\rangle = 0$$

or, if one computes, so that

$$L_f h + uL_g h = 0.$$

If $L_g h \equiv 0$, we also impose this constraint and repeat, as in the "zero dynamics algorithm" (see [12]), to obtain a constraining input, viz.

$$u(x) = -\frac{L_f^{\nu_\infty} h}{L_g L_f^{\nu-1} h}$$

For arbitrary ν_∞, then, we set

$$\beta(x) = \frac{1}{L_g L_f^{\nu_\infty - 1} h(x)} \qquad\qquad \alpha(x) = -\beta(x) L_f^{\nu_\infty} h(x)$$

and define the vector fields

$$F = f + g\alpha, \; G = g\beta$$

Thus, we may take as our definition of zero dynamics the restriction $F|_{L(x_e)}$, where

$$F = f - \frac{L_f^{\nu_\infty} h}{L_g L_f^{\nu_\infty - 1} h} g \qquad (3.7)"$$

and $L(x_e)$ is the leaf of Δ^* passing through x_e. We must, however, check that F is tangent to $L(x_e)$. For this we need some technical results from the geometric theory of (f,g)-invariant distributions:

<u>Lemma</u>. The following identities hold:

(i) $L_F^i h(x) = L_f^i h(x), \quad i = 0, \ldots, \nu_\infty - 1;$

(ii) $L_F^i h(x) = 0, \quad i = \nu_\infty, \ldots, \dim M;$

(iii) $L_G L_F^{\nu_\infty - 1} h(x) = 1$;

(iv) $[F, \Delta^*] \subset \Delta^*$; and

(v) $[G, \Delta^*] \subset \Delta^*$.

We can now verify:

<u>Lemma</u>. F is tangent to $L(x_e)$.

Proof. By (iv) of the above Lemma, F maps leaves of Δ^* to leaves of Δ^*. On the other hand, $L(x_e)$ contains an equilibrium point for F, viz, x_e, because

$$f(x_e) = 0 \text{ and } \alpha(x_e) = -\beta(x_e) < dL_f^{\nu_\infty - 1} h(x_e), f(x_e), f(x_e) > = 0$$

<u>Definition 3.4</u>. Suppose (3.1) has strong relative degree ν_∞. The zero dynamics of (3.1) is the dynamical system defined by the vector field

$$F|_{L}(x_e) \tag{3.8}$$

As for the case $\nu_\infty = 1$, for linear systems the spectrum of (3.8) coincides precisely with the zeroes of the system transfer function. Thus motivated, we define what is meant by zeroes lying in left or right half planes.

Denote by $W^s(x_0)$ and $W^u(x_0)$ the stable and unstable manifolds of x_0 for the system (3.8) and let $W^c(x_0)$ be a center manifold for (3.8). Setting

$$s = \dim W^S(x_0), \quad u = \dim W^s(x_0), \quad c = \dim W^c(x_0)$$

we will say that (3.1) has s left half plane zeros, u right half plane zeros, and c purely imaginary zeros, in analogy with the linear case. Note that

$$s + u + c = \nu_f \tag{3.9}$$

<u>Definition 3.5</u>. The system (3.1) is <u>minimum phase</u> on M, provided (3.1) has ν_f left plane zeros. The system (3.1) is <u>globally minimum phase</u> on M provided it is minimum phase and the zero dynamics (3.8) is globally asymptotically stable.

In [16,17], it was shown that the local normal form for systems of relative degree one given in Example 3.1 holds globally, under some additional minor technical hypotheses (which in fact are also necessary). From the existence of the normal form, it is possible to obtain many results concerning stabilization and control of (strong) relative degree one nonlinear systems, see e.g. [16]-[17]. The recent extensions of these methods to arbitrary relative degree reposes on the following normal form for $\nu_\infty \geq 1$, see [18] for proofs and more details:

Theorem 3.1. Suppose (3.1) has strong relative degree r, the vector fields $G, ad_F G, \ldots, ad_F^{r-1}G$ are complete, and the level sets

$$h(x) = L_f h(x) = \ldots = L_f^{r-} h(x) = 0$$

are connected. Then there is a globally defined diffeomorphism

$$T : M \simeq L(x_e) \times R^r$$

where in the new coordinate system the system (3.1) takes the normal form

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, x_2), \quad x_1 \in L(x_e) \\
\dot{x}_{2,1} &= x_{2,2} \\
&\cdot \\
&\cdot \\
&\cdot \\
\dot{x}_{2,r-1} &= x_{2,r} \qquad\qquad (3.10) \\
\dot{x}_{2,r} &= L_f^{\nu_\infty} h(x_1, x_2) + u\, L_g L_f^{\nu_\infty - 1} \\
y &= x_{2,1}
\end{aligned}$$

Remark. In (3.10) $\dot{x}_1 = f_1(x_1, x_2)$ is the expression in local coordinates for an x_2-dependent vector field on $L(x_e)$, which is now parameterized in local coordinates by x_1 and is therefore defined by the equations, $x_2 = 0$. As for the case $\nu_\infty = 1$, setting $x_2 = 0$ we obtain in (3.2) the following expression

$$\dot{x}_1 = f_1(x_1, 0) \qquad (3.11)$$

for the zero dynamics, evolving on $L(x_e)$. As in [18], if (3.1) is globally minimum phase, by Milnor's Theorem we have $L(x_e) \simeq R^r$ and (3.4) can be interpreted as a stable system of O.D.E.'s on R^r.

4. Feedback Stabilization of Nonlinear Minimum Phase Systems. In this section, we illustrate the use of the frequency domain heuristics for nonlinear systems derived in section 3 in the design and analysis of stabilizing feedback laws. In

particular, we sketch some initial results for the construction of both dynamic compensators (e.g. lead-lag laws) and nonlinear state-space feedback (e.g. PD) controllers. While the results presented here do comprise the first general approach to nonlinear stabilization which can, for example, stabilize open sets (in the Whitney topology) of unstable control systems, we view these as preliminary results and one of the primary research goals we are proposing is the systematic development and extension of these methods.

The design and stability analysis of the dynamic compensation schemes proposed here fall into three sequential strategies. First, the analysis of high gain feedback for globally minimum phase systems of strong relative degree one; second, the effects of adding stable zeroes to systems of higher relative degree; and, third, the effects of adding sufficiently stable poles to stable high gain systems. Since, of course, root-locus arguments cannot be applied directly to nonlinear systems to determine global results, we shall require an alternative approach. Explicitly, we make heavy use of singular perturbation methods pioneered as a tool for root-locus analysis in the linear case by Kokotovic et al. [19] and used as an analogue to root-locus arguments in the nonlinear case by Byrnes-Isidori ([16]-[18]) and also by Marino ([20]).

We begin with a simple example. Consider first the system, defined on R

$$\dot{x} = x^2 + u, \; y = x \tag{4.1}$$

Trivially, (4.1) is minimum phase and has relative degree 1, leading to the classical control strategy $u = - ky$ which in fact locally stabilizes (4.1). Moreover, for all x_0 there exists k, viz, $k > |x_0|$, for which $x_t \to 0$ is closed-loop. While the closed-loop system is never globally asymptotically stable for any fixed choice of k, this feedback strategy has the pleasant property of stabilizing arbitrarily large relatively compact sets of initial data. In this sense, classical controllers can be designed to globally stabilize all "physical" initial conditions of a (strong) relative degree 1, globally minimum phase nonlinear systems.

Theorem 4.1. Suppose the system (3.1) on R^n has strong relative degree 1, is globally minimum phase, has 0 as an equilibrium, and the vector field G is complete.

Consider the output feedback law $u = - ky$. For any bounded open set $U \subset R^n$, there exists k_U such that for all $k \geqq k_U$ and all $x_0 \in U$ the solution x_t of the closed-loop system tends to 0.

Because the design philosophy and the analysis of such laws is rather orthogonal to conventional nonlinear feedback control, we will present a fairly complete proof of Theorem 4.1.

Proof. We first note that, by Milnor's Theorem, $L(x_e)$ is diffeomorphic to R^{n-1} and is, in particular, path-connected so that normal forms exist. As in (3.10) choose coordinates (z,y) in which the closed-loop system takes the form

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

$$\varepsilon \dot{y} = \varepsilon L_f h(z,y) - yL_g h(z,y)$$

of a singularly perturbed system, where $\varepsilon = 1/k$. Setting $\varepsilon = 0$, we obtain the constraints

$$yL_g h(z,y) = 0$$

which imply, by hypothesis,

$$y = y(z) = 0$$

In particular, the reduced system (see [19]) is precisely the zero dynamics which is assumed to be globally asymptotically stable. The boundary layer equation is simply

$$\dot{\hat{y}} = - \hat{y} L_g h(z_0,\hat{y}) \qquad \hat{y}(0) = y_0$$

which is uniformly asymptotically stable on bounded sets, so that Tychonov's Theorem applies. Thus, for initial data (z_0, y_0) one obtains the asymptotic expression

$$z_t = \bar{z}_t + 0(1/k)$$

$$y_t = \bar{y} + 0(/k)$$

Therefore, there exists k_0, $k_0 = k(z_0, y_0)$, such that

$$\lim_{t\to\infty} z_t = 0$$

$$\lim_{t\to\infty} y_t = 0$$

holds for each $k > k_0$ on an open neighborhood of (z_0, y_0). If U is any bounded neighborhood of (z_0, y_0) a standard compactness argument yields the existence of

$k >> 0$ such that for all x_0 the solution x_t of the corresponding closed-loop system tencs to 0. Q.E.D.

We now consider a system (3.1) with relative regree r. Let $c(s) = c_0 + c_1 s + c_{r-2}s^{r-2} + s^{r-1}$ be a Hurwitz polynomial and consider the "derived" output

$$\bar{y} = c_0 y + c_1 \frac{dy}{dt} + \dots + c_{r-2} \frac{d^{r-2}y}{dt^{r-2}} + \frac{d^{r-1}y}{dt^{r-1}}$$

$$= c_0\, x_{2,1} + c_1\, x_{2,2} + \dots + c_{r-2}x_{2,r-1} + x_{2,r} \tag{4.2}$$

$\bar{y}(t)$ is thus the output of a cascade connection of (3.1) with (4.2). Heuristically, the augmented system should also be "minimum phase". While this is correct locally, appeal to the normal form (3.12) and to Tychonov's theorem (cf) yields, however, a far more subtle perturbation problem. Combining the methods sketched above with Lyapanov methods and LaSalle's Theorem, we are nonetheless able to rigorously prove global stability for a restricted class of minimum phase systems, e.g., for systems with a stable inverse, see [21]. As an easy example we note:

Theorem 4.2. Suppose the system (3.1) on R^n is minimum phase, has 0 as an equilibrium and has strong relative degree r. Then, for any bounded open set U there exists $\alpha_U < 0$ such that if

$$c(s) = 0 \to \mathrm{Re}(s) < \alpha_U$$

the cascade system (3.1) - (4.2) is minimum phase on U with strong relative degree 1.

The final phase of our construction is to analyze the effect of adding to (3.1) a "pole" which lies sufficiently far to the left of the imaginary axis. The key inductive result in the analysis of lead-lag compensators follows from a standard singular perturbation argument.

Theorem 4.3. Suppose $\nu_\infty = 1$ and that for some k the closed-loop system is locally and globally asymptotically stable to x_0 on R^n. If U is any bounded neighborhood of x_0 there exists a positive ε_U, sufficiently small, so that the closed-loop system with compensator

$$\hat{u}(s) = \frac{k}{1 + \varepsilon s}\, \hat{y}(s) \quad \varepsilon < \varepsilon_U$$

is locally and globally asymptotically stable on U.

Summarizing these results we obtain the following heuristic design principle for stabilizing nonlinear minimum phase systems of arbitrary relative degree:

Suppose (3.1) is a globally minimum phase system having strong relative degree r. For any bounded open set U R^n, there exists k_u, α_u, β_u such that if the proper transfer function

$$k(s) = k\ p(s)/q(s) \tag{4.3}$$

satisfies $k > k_U$, $p(s) = 0 \rightarrow \text{Re}(s) < \alpha_u$, $q(s) = 0 \rightarrow \text{Re}(s) < \beta_u$, then the closed-loop system satisfies:

For all initial data $x_0 \in U$ and z_0 an initial condition for a realization of (4.3)

$$x_t \rightarrow 0,\ z_t \rightarrow 0 \text{ as } t \rightarrow \infty.$$

5. Examples and Illustrations. In this section we illustrate the design techniques sketched in the previous sections.

Example 5.1. The system on $\mathbf{R}^2$ defined via

$$\begin{aligned} \dot{z} &= -z + z^2 y \\ \dot{y} &= \cos(zy) + e^{+z(1+y^2)}u \end{aligned} \tag{5.1}$$

has (strong) relative degree 1 with positive "high frequency" gain and stable zero dynamics

$$\dot{z} = -z. \tag*{(5.1)'}$$

Clasical control suggests using the output feedback law

$$u = -ky \tag{5.2}$$

to achieve closed-loop stability. In fact, since $z = 0$ is a hyperbolic equilibrium, by setting $\varepsilon = 1/k$ and applying a singular perturbation argument (based on Tychonov's Theorem) to the closed-loop system

$$\begin{aligned} \dot{z} &= -z + z^2 y \\ \varepsilon\dot{y} &= \varepsilon \cos(zy) - e^{+z^2(1+y^2)}y \end{aligned}$$

one can conclude that for a fixed bounded open subset U $\mathbf{R}^2$ there exists a k_U such that the feedback law 5.2 with $k > k_U$ stabilizes the closed-loop system (5.1) - (5.2) for all initial data $(z_0, y_0) \in U$.

Remark. Examples (e.g., (5.1) with the second equation replaced by y = u) show that, in general, the required gain k_U must grow with U, in sharp contrast to the linear case. Nonetheless, for fixed U, the law (5.2) with $k > k_U$ does have an infinite gain margin (in the sense of [37]), retaining some of the robustness features of classical linear control.

Example 5.2. The system evolving on R^3 defined via

$$\dot{z}_1 = (1-y)z_2 - e^y(z_1^3-z_1) + y^3 \sin z_2$$

$$\dot{z}_2 = - z_1 e^{-yz_2} \tag{5.3}$$

$$\dot{y} = \cos(z_1 z_2) + u$$

has (strong) relative degree 1 (with positive high frequency gain) and has as zero dynamics the van der Pol oscillator

$$\begin{aligned}\dot{z}_1 &= z_2 - z_1^3 + z_1 \\ \dot{z}_2 &= - z_1.\end{aligned} \tag*{(5.3)'}$$

Implementing the output feedback law (5.2), nonlinear root-locus theory (i.e. the closed-loop dynamics should limit to the zero dynamics in a suitable sense) would predict the existence of a stable limit cycle for k >> 0. In fact, since the limit cycle in (5.3)' is normally hyperbolic, by setting $\varepsilon = 1/k$ and appealing to Anosov's Theorem, we see that for a fixed bounded set U of initial data x_0 there exists a k_U so that in the closed-loop system (5.2)-(5.3) with $k > k_U$, x_t approaches a unique (stable) limit cycle.

In Examples (5.1)-(5.2), we could appeal to a singular perturbation argument because the ω-limit sets for the "zero dynamics" were normally hyperbolic and because the systems were in a "normal form" explicitly displaying the "fast" and "slow" state variables, giving a nonlinear generalization (see [25]-[26] and also [56], [57]) of the linear, relative degree one (SISO or MIMO) case treated by Kokotovic et al. in [58]. We now consider an example where, in fact, it will be necessary to choose the outputs in order to realize a state feedback law.

Example 5.3. Consider the system evolving on $\mathbf{R}^3$ according to

$$\dot{x}_1 = x_2^3$$

$$\dot{x}_2 = x_3^3 \tag{5.4}$$

$$\dot{x}_3 = u.$$

In order to render (5.4) a relative degree 1 system we choose a "dummy output" y of the form

$$y = h(x_1,x_2,x_3) = x_3 + \gamma(x_1,x_2)$$

and then choose $\gamma(x_1,x_2)$ so that the zero dynamics are stable. In "normal form" (5.4) takes the form

$$\begin{aligned} \dot{x}_1 &= x_2^3 \\ \dot{x}_2 &= (y - \gamma(x_1,x_2))^3 \\ \dot{y} &= u + \dot{\gamma}(x_1,x_2) \end{aligned} \qquad (5.4)'$$

choosing $\gamma(x_1,x_2) = x_1e^{x_1x_2}$, we have the zero dynamics

$$\begin{aligned} \dot{x}_1 &= x_2^3 \\ \dot{x}_2 &= -x_1^3e^{3x_1x_2} \end{aligned}$$

which can be seen to be locally and globally asymptotically stable by applying LaSalle's theorem to the "energy" function $V(x_1,x_2) = x_1^4 + x_2^4$. We implement the control law

$$u = - ky - \dot{\gamma}(x_1,x_2) \qquad (5.5)$$

For example, with k = 1 (5.5) specializes to the control law

$$u = - x_3 + - (x_1 + x_2^3 + x_1x_2^4 + x_3^3x_1^2)e^{x_1x_2} \qquad (5.6)$$

which is locally asymptotically stabilizing by the center manifold theorem. We stress, however, that (5.6) was derived in a systematic way using geometric nonlinear control theory. For a given bounded open set U, there exists a k_U such that

$$u = - kx_3 - (- kx_1 + x_2^3 + x_1x_2^4 + x_3^3x_1^2)e^{x_1x_2}$$

stabilizes all initial data $x_0 \in U$, for $k > k_U$, giving an infinite gain margin even in the critically stable case.

Example 5.4. We describe here an example from spacecraft attitude control which exhibits some of the problems we propose to study. The example considered in some depth in Crouch [22] concerns the specific case of atitude control of a rigid spacecraft with actuator failure, in this case thruster jets, so that there are only two remaining control torques acting about principal axes. The equations describing the system are then given by

$$\begin{aligned}
\dot{\omega}_1 &= a_1\omega_2\omega_3 + u_1 \qquad & a_1 &= (J_2 - J_3)/J_1 \\
\dot{\omega}_2 &= a_2\omega_1\omega_3 + u_2 \qquad & a_2 &= (J_3 - J_1)/J_2 \qquad\qquad (5.7) \\
\dot{\omega}_3 &= a_3\omega_1\omega_2 \qquad & a_3 &= (J_1 - J_2)/J_3 \\
\dot{R} &= S(\omega)R
\end{aligned}$$

$$S(\omega) = \begin{bmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_2 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{bmatrix}$$

attitude R of the spacecraft relative to inertial axes, and ω_i are the components of angular velocity. There are two problems of special interest, controlling the system to the equilibrium state $\omega = 0$, $R = R_0$ some desired attitude, and controlling the system to the periodic trajectory consisting of rotation at a constant rate $\omega_3 = \lambda$ about the third principal axis. It is shown in Crouch [39] that if $a_3 \neq 0$ the system above is controllable, and locally controllable about each of the equilibrium (trajectories) above. However of more interest would be the development of closed-loop state feedback controls which would locally and perhaps globally stabilize the system about these equilibria.

We first illustrate how to stabilize the angular velocity equations (see also [23] - [25]) using a multivariable extension of our techniques. We proceed by choosing y_1, y_2 so that the system has relative degree 1 through each channel and is minimum phase. Explicitly, we take

$$y_1 = \omega_1 + \gamma_1(\omega_3)$$

$$y_2 = \omega_2 + \gamma_2(\omega_3)$$

leading to the zero dynamics (i.e. $y_1 = y_2 = 0$)

$$\dot{\omega}_3 = a_3\gamma_1(\omega_3)\ \gamma_2(\omega_3)$$

Let us assume $j > j_2$, i.e. that $a_3 > 0$. Then, choosing

$$\gamma_1(\omega_3) = -\omega_3, \quad \gamma_2(\omega_3) = \omega_3^2$$

gives the stable zero dynamics

$$\dot{\omega}_3 = -a_3\omega_3^3$$

with "high frequency" dynamics

$$\dot{y}_1 = \dot{\omega}_1 + \dot{\gamma}_1 = a_1\omega_2\omega_3 + u_1 - a_3\omega_1\omega_2$$

$$\dot{y}_2 = \dot{\omega}_2 + \dot{\gamma}_2 = a_2\omega_1\omega_3 + u_2 + 2a_3\omega_1\omega_2\omega_3$$

In particular, by the center manifold theorem the feedback law

$$u_1 = -a_1\omega_2\omega_3 - y_1 - \dot{\gamma}_1$$

$$u_2 = -a_2\omega_1\omega_3 - y_2 - \dot{\gamma}_2$$

locally asymptotically stabilizes the system about the equilibrium, $\omega_1 = 0$. In the original coordinates, we have

$$\begin{aligned} u_1 &= -a_1\omega_2\omega_3 - \omega_1 + \omega_3 + a_3\omega_1\omega_2 \\ u_2 &= -a_2\omega_1\omega_3 - \omega_2 - \omega_3^2 - 2a_3\omega_1\omega_2\omega_3 \end{aligned} \tag{5.8}$$

At one time, it was the hope that feedback laws such as (5.8) would ultimately lead to the design of feedback laws stabilizing the full set of satellite equations (5.7). It is now known [26] that there does not exist a smooth (i.e. C^∞) state feedback law, $u = F(\omega,R)$, which makes the equilibrium $\omega = 0$, $R = R_0$, locally asymptotically stable. However, using our design philosophy we can derive state feedback control laws for which the closed-loop trajectories asymptotically approach a motion about the third principal axis. Explicitly, using Euler angles to parametrize the frame R (see e.g. [22]) the feedback law

$$\begin{aligned} u_1 &= - a_1 a_3\omega_2\omega_3 - K(\omega_1 + \phi + A_1a_3\omega_3 + B_1a_3^2\omega_3^2) - \cos(\eta)\omega_1 + a_3\sin(\eta)\omega_3 - A_1a_3\omega_2 \\ &\quad -2B_1a_3^2\omega_1 w_2\omega_3 \qquad (5.9) \\ u_2 &= - a_2a_3\omega_1\omega_3 - K(\omega_2 + \eta + A_2 + B_2a_3^2\omega_3^2) \\ &\quad - \sin(\eta) \tan(\phi)\, \omega_1 - \omega_2 - a_3 \cos(\eta) \tan(\phi)\, \omega_3 - 2B_2a_3^2\omega_3^2 \end{aligned}$$

where

$$A_1A_2 = 0 \text{ and } A_1(A_1 - B_2) - A_2(A_2 + B_1) < 0$$

satisfies: For a given bounded set U of initial conditions x_0, there is a k_U such that (5.9) for any $k > k_U$, drives $(\omega_1)_t \to 0$, $\eta_t \to 0$, $\phi_t \to 0$, as $t \to \infty$.

Example 5.5. Consider the rigid body model (2.1) for a robotic manipulator, we wish to analyze the effect of the PD control

$$\tau = - k_1(k_2(q - q_d) + \dot{q}) \qquad (5.10)$$

where k_1, $k_2 > 0$ or, more generally, are matrices with $\sigma(k_1) \subset \mathbb{C}^+$. We analyze (5.10) in two stages, first we set $y = q - q_d z = \dot{q}$ and consider the stable differentiator

$$\tilde{z} = k_2y + z$$

which has the effect of making the robot arm minimum phase, relative degree 1 in each channel. Explicitly,

$$\begin{aligned} \dot{y} &= \tilde{z} - k_2 y \\ \tilde{z} &= f(\tilde{z},y) + M(y)^{-1}\tau \end{aligned} \tag{5.11}$$

which has zero dynamics, i.e. constraining $\tilde{z} = 0$,

$$\dot{y} = - k_2 y.$$

Now, implementing (5.10) takes the form

$$\tau = - k_1 \tilde{z}$$

which can be analyzed as above using Tychonov's Theorem.

REFERENCES

1. E. Freund, "A Nonlinear Control Concept for Computer Controlled Manipulators," Proc. IFAC Symp. on Multivariable Technological Systems, Fredericton, 1977.

2. T.J. Tarn, A.K. Bejczy, A. Isidori and Y. Chen, "Nonlinear Feedback in Robot Arm Control," Proc. of 23rd IEEE Conf. on Dec. and Control, Las Vegas, 1984, 736-751.

3. S. Arimoto, S. Kawamura and F. Miyazaki, "Bettering operation of dynamic systems by learning: A new control theory for servo mechanism or mechatronics systems, Proc. 23rd CDC, pp. 1064-1069, 1984.

4. T. Mita and E. Kato, "Iterative Control and its Applications to Motion Control of Robot Arm - A Direct Approach to Servo Problems," Proc. of 24th Conf. Dec. & Control (1985), Ft. Lauderdale, 1393-1398.

5. M. Le Borgne, J.M. Ibarra, and B. Espian, "Adaptive Control of High Velocity Manipulators," Proc. of the 11th Int. Symp. on Industrial Robots, Tokyo, Oct. 1981, 227-236.

6. T.J. Tarn and A.K. Bejczy, Nonlinear Feedback for Puma 560 Robot Arm, Robotics Laboratory Report, Washington University (1985).

7. Haruhiko Asada, "M.I.T. Direct-Drive Arm Project," Robots 8 Conference Proceedings: 16-10 - 16-21.

8. A. Isidori, A.J. Krener, C. Gori-Giorgi and s. Monaco, "Nonlinear Decoupline via Feedback, a Differential Geometric Approach," IEEE Trans. Aut. Control AC-21, (1981), 331-345.

9. R. Hirschorn, "(A,B)-invariant Distributions and Disturbance Decoupling of Nonlinear Systems," SIAM J. Control and Opt. 19, (1981), 1-19.

10. R. Hirschorn, "Invertibility of Nonlinear Control Systems," SIAM J. Control and Opt. 17, (1979), 289-297.

11. A. Isidori and A.J. Krener, Nonlinear zero distributions, 19th IEEE Conf. Decision and Control, (1980).

12. A. Isidori and C.H. Moog, On the nonlinear equivalent of the notion of transmission zeros, Modeling and Adaptive Control (C.I. Byrnes and A. Kurszauski, eds.), Lecture Notes in Control and Information Sciences, Springer Verlag, to appear.

13. M. Takegaki and S. Arimoto, "A New Feedback Method for Dynamic Control of Manipulators," J. Dyn. Syst., 102, (1981), 119-125.

14. D.E. Koditschek, "Adaptive Stategies for the Control of Natural Motion," Proc. of 24th CDC, Ft. Lauderdale, (1985), 1405-1409.

15. D. W. Parish, Nonlinear Control and Output Decoupling of Robot Arm Dynamics, M. E. Thesis, ASU, 1986.

16. C.I. Byrnes and A. Isidori, "A Frequency Domain Philosophy for Nonlinear Systems, with Application to Stabilization and to Adaptive Control," Proc. of 23rd IEEE Conf. on Dec. and Control, Las Vegas, 1984, 1569-1573.

17. C.I. Byrnes and A. Isidori, "Asymptotic Expansions, Root Loci and the Global Stability of Nonlinear Feedback Systems," Algebraic and Geometric Methods in Nonlinear Control Theory (M. Fliess and M. Hazewinkel, Eds.), D. Reidel Publishing Co., 1985.

18. C.I. Byrnes and A. Isidori, "Global Feedback Stabilization of Nonlinear Minimum Phase Systems," Proc. of 24th IEEE Conf. on dec. and Control, Ft. Lauderdale (1985), 1031-1037.

19. K.-K. D. Young, P.V. Kokotovic and V.I. Utkin, "A Singular Perturbation Analysis of High-Gain Feedback Systems," IEEE Trans. Aut. Control AC-22 , (1977), 931-938.

20. R. Marino, "Nonlinear Compensation by High Gain Feedback," Int. J. Control, Vol 42, 1369-1385, (1985).

21. C. I. Byrnes and A. Isidori, in preparation.

22. P.E. Crouch, Spacecraft Attitude Control and Stabilization: Applications of Geometric Control Theory to Rigid Body Models, IEEE Trans. Vol AC-29, 321-331, (1984).

23. D. Aeyels, Stabilization by Smooth Feedback of the Angular Velocity of a Rigid Body, Systems and Control Letters, Vol 6, 59-63, (1985).

24. R.W. Brockett, Asymptotic Stability and Feedback Stabilization, Differential Geometric Control Theory, Progress in Mathematics (R.W. Brockett, R.S. Millmann, H.J. Sussmann, eds.) Birkhäuser Boston, 181-191, (1983).

25. P.E. Crouch, M. Irving, On Sufficient Conditions for Local Asymptotic Stability of Nonlinear Systems whose Linearization is Uncontrollable, Control Theory Centre Report No. 114, University of Warwick (1983).

26. C.I. Byrnes and A. Isidori, "Attitude Stabilization of Rigid Spacecraft." Submitted to IEEE Transactions on Automatic Control.

Global Observability and Detectability: an overview

by

C. I. Byrnes[**]
Department of Mathematics
Arizona State University
Tempe, Arizona, 85287

and

C. F. Martin[*]
Department of Mathematics
Texas Tech University
Lubbock, Texas 79409

[*] Supported in part by grants from NSA and NASA and in part by[**] AFOSR and NSF.

1 Introduction

The central problem of control theory is to design a control law that will cause a system to meet some set of design specifications. For example the problem may be as simple mathematically as to find a function $u(t)$ so that the system

$$\dot{x} = Ax + Bu, \quad x(0) = x_0$$

takes on the specified value x_T at time T or it could be as complex as defining the control law that will allow a robot to function on an assembly line. In either case the primary function is to assure that the *control law* will meet some set of specifications. In the first example the problem usually becomes in practice one of designing the control law and then becomes the problem of keeping the system as close to the designed trajectory as possible, i.e. of defining a feedback control law that stabilizes around the trajectory. At this point in both examples a new complication enters the picture–in order to control the system the designer must be able to calculate where in the statespace the system is located. A problem of *observation* has complicated the design. In almost all control problems there is ultimately an underlying problem of determining the position of the system in the statespace or a problem of determining the particular trajectory that a system is following, [5,6].

It has only been in the last decade that problems of observability have gained importance in themselves. The paper of Herman and Krener, [11] was fundamental in that it placed the theory of local observability of nonlinear systems on a solid framework of differential geometry and showed that in fact the problem of observability was not simply the dual of the problem of controllability. However from a practical point of view the problem of local observability is not as critical as the problem of being able to determine in a universal way where the system is located in the statespace. With the work of Elliott and Tarn and their students [1,13,14] the study of *global observability* was initiated. The problem of being able to distinguish between any two points in the statespace is very complicated mathematically and as we will see in this paper has very little connection to the a problem of local observability.

The intent of this paper is to give a partial survey of the theory of observability of dynamical systems. We will begin in section 2 with a recap of the theory of observability of linear systems and some of the attendant problems—tracking, observers, inverse systems, etc. In section 3 we will consider the problem of local observability of non-linear systems and will recount the developments of the paper of Hermann and Krener, [11] and related work on the geometric aspects of the local controllability. In section 4 we consider the problem of global observability of nonlinear systems which have differential constraints. In particular the work of Elliott, Tarn, and Aeyels, [1,13,14] will be considered along with the somewhat different approach of Sussman, [18]. In section 5, we consider the recent work of Drager, Martin, Byrnes and others on the global observability of translational flows on the torus or more generally flows on abelian groups. This work shows that there are fundamental differences between the local theory and any systematic development of a global theory of observability. In section 6 we consider the work of McMahon on the existence of a vector field on a manifold which is observable by every nonconstant continuous function. Relevant to this section is paper by Byrnes, Dayawansa and Martin, [24], which contains results about which manifolds can admit such vector fields. In section 7 we consider the problem of observability of continuous time system with discrete sampling and show that this leads to a version of the easy Whitney embedding theorem. Here the work of Aeyles, [22], Martin and Smith, [21], and a recent paper of Dayawansa and Martin, [20] are relevant along with a paper of Bendsøe, [23].

2 Linear Theory

Perhaps the simplest problem in control theory is the following:
Let

$$\dot{x} = Ax \tag{1}$$
$$y = cx. \tag{2}$$

What conditions must be imposed on the matrices A and c in order that the output function $y(t)$ uniquely determines the solution of the system $\dot{x} = Ax$? Here we assume that the A is an $n \times n$ matrix and that c is an $1 \times n$ matrix. We first note that the solution of the system is given by the action of the one parameter group $\exp(At)$ on the underlying space $\mathbf{R}^n$. That is, $x(t) = \exp(At)x_0$ for some initial data point x_0 and hence that the output function is given by the analytic function $y(t) = c\exp(At)x_0$. Since the output function is analytic the function is determined uniquely by the sequence $\{y^{(n)}(0)\}_{n=0}^{\infty}$. Calculating the derivatives of the output and evaluating at $t = 0$ we see that the $n'th$ term is $cA^{n-1}x_0$ and hence we have that the condition for observability is that the infinite system of equations

$$\begin{aligned} cx_0 &= 0 \\ cAx_0 &= 0 \\ &\vdots \\ cA^k x_0 &= 0 \\ &\vdots \end{aligned}$$

has a unique solution, namely $x_0 = 0$. Thus the condition is that the matrix

$$\begin{pmatrix} c \\ cA \\ \vdots \\ cA^k \\ \vdots \end{pmatrix}$$

has rank n. However by the Cayley Hamilton theorem the rank of the matrix is determined by its first n rows and hence we have derived the classical result that the system is observable iff the rank of the matrix

$$\begin{pmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{pmatrix}$$

is equal to n.

We first note that we have made fundamental use of the fact that the semigroup $\exp(At)$ is analytic and hence that the function $y(t)$ has a Taylor series expansion $\Sigma\, c'A^n x_0 t^n/n!$. As we will see in the next section this construction is basic to the determining conditions for observability in the general nonlinear case as is posed by Hermann and Krenner. Each coefficient in the Taylor series is a function of the initial data and hence the problem of uniquely determining the solution is problem of solving an infinite set of equations for the unknown initial data. In the linear case, because of the recursion induced by the Cayley–Hamilton Theorem the problem reduces to a linear problem in a finite number of equations and hence either has infinitely many solutions or a unique solution. Unfortunately there is no hope of this generalizing to the nonlinear case or to more general systems defined on manifolds.

Although the proof of observability using the Taylor series is simple there is another proof that appears to be more relevant to the problem of determining conditions for global observability. Consider the linear system

$$\begin{aligned} \dot{x} &= Ax \\ y &= c'x. \end{aligned}$$

The above system is observable iff there exists a matrix P such that

$$PA = AP$$

and

$$cP = c$$

then $P = I$. This result is well known, at least in the control case but the proof is interesting. Suppose P exists with the hypothesized communativity properties and suppose P is not the identity. Let x be a vector such that $Px \neq x$. then we have that $c'\exp(At)x = c'P\exp(At)x = c'\exp(At)Px$ and hence that the system is unobservable. On the other hand we now need to show that if the system is unobservable then there exists such a P. We first assume that A is cyclic. Let x be a nonzero vector such that $c'\exp(At)x = 0$. Now we note that by differentiating this expression we have that $c'A^kx = 0$ for all k. We define a matrix Q by the following,

$$Q = [x, Ax, A^2x, \cdots, A^{n-1}x].$$

From Q we will construct a matrix that satisfies the commutativity properties. First consider

$$AQ = A[x, Ax, A^2x, \cdots, A^{n-1}x]$$

$$
\begin{aligned}
&= [Ax, A^2x, \cdots, A^{n-1}x, A^nx] \\
&= [x, Ax, A^2x, \cdots, A^{n-1}x] \begin{bmatrix} 0 & \cdots & & \alpha_1 \\ 1 & & & \vdots \\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & \cdots & 1 & \alpha_n \end{bmatrix} \\
&= QRAR^{-1}
\end{aligned}
$$

Hence we have that

$$A(QR) = (QR)A$$

and QR is nonzero. We also note that since $c'A^kx = 0$ that

$$c'QR = 0.$$

Now let

$$P = QR + \alpha I.$$

It is clear that there exists an α such that P is invertible and P is the desired matrix. If A is not cyclic then the preceding argument can be modified by decomposing the statespace into the sum of a cyclic invariant subspace and a complimentary subspace.

The fact that there exists such a P iff the system is unobservable is equivalent to saying that there is a linear symmetry for the system. This phenomena will occur in several different contexts and seem to underlie much of the theory of global observability, especially in the case that there is group involved at the level of the statespace.

3 Nonlinear Theory

Consider a system with controls

$$\Sigma : \begin{array}{rcl} \dot{x} &=& f(x,u) \\ y &=& g(x) \end{array} \tag{3}$$

where $x \in \mathbf{R}^n$, f is a C^∞ vector field, g is a C^∞ real valued function and $u \in \Omega$ a subset of $\mathbf{R}$. We could of course allow u and y to be vector valued but for the purposes of this paper the scalar case will suffice. We could also assume that the system was evolving on C^∞ manifold but the theory we are presenting is basically local and hence we will restrict ourselves to the Euclidean case. We denote the solution of the differential equation $\dot{x} = f(x,u)$ with initial value x_0 by $\omega_{x_0}(u,t)$. Following Hermann and Krener, [11], we say that two points x_0 and x_1 are *indistinguishable* iff for every input function $u(t)$

$$g(\omega_{x_0}(u,t)) = g(\omega_{x_1}(u,t)).$$

Indistinguishability is an equivalence relation on $\mathbf{R}^n$ and we denote the equivalence class of x by $I(x)$. We define the system Σ to be *observable at* x iff $I(x) = \{x\}$ and we say that the system is *observable* if it is observable at x for all x. An equally valid theory could be developed for the case at hand by removing the dependence on the control, that is we could assume that either the control is not present or it is fixed.

The definition above is inherently global. In order to take advantage of differentiable structure we have hypothesized it is necessary to restrict the definition. Let U be an open subset of $\mathbf{R}^n$ and let $x_0, x_1 \in U$. We say that $x - 0$ is *U-indistinguishable* from x_1 iff for every control u such that $\omega_{x_0}(u,t)$ and $\omega_{x_1}(u,t)$ both lie entirely within U fail to distinguish between x_0 and x_1. U-indistinguishability is not an equivalence relation because it may fail to be transitive. We will, however, denote the set of points U-indistinguishable from x by $I_U(x)$. We now define the system Σ to be *locally observable at* x_0 if and only if $I_U(x_0) = \{x_0\}$ and simply *locally observable* if it is locally observable at x_0 for all x_0.

If we are only interested in distinguishing a point x_0 from its immediate neighbors we can weaken the definitions in the following way. In analogy with the definition of *observable* we will say that the system Σ is *weakly observable at* x_0 iff there is a neighborhood U of x_0 such that $I(x_0) \cap U = \{x_0\}$ and we say the system Σ is *weakly observable* if it is weakly observable at x_0 for all values of x_0. Again this concept require arbitrarily large times and the trajectories may wander far from the neighborhood U. In analogy with the definition of locally observability we define the system to be *locally weakly observable at* x_0 if there exists a neighborhood U of x_0 such that for every open neighborhood V contained in U we have $I_V(x_0) = \{x_0\}$ and simply *locally weakly observable* if this true for every x_0. We have the following relations holding between the four definitions of observability.

$$\begin{tikzcd}
\text{locally observable} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{observable} \arrow[d, Rightarrow] \\
\text{locally weakly observable} \arrow[r, Rightarrow] & \text{weakly observable}
\end{tikzcd}$$

It is easy to see that for linear autonomous systems as we considered in the last section these four concepts are equivalent. We will develop a simple test for local weak observability that reflects the controllability rank conditions for linear systems.

Let $C^\infty(\mathbb{R}^n)$ denote the linear space of all C^∞ real valued infinitely differentiable functions on $\mathbf{R}^n$. Let $\mathcal{X}$ denote the Lie algebra of all C^∞ vector fields on $\mathbf{R}^n$. $C^\infty(\mathbb{R}^n)$ is a $\mathcal{X}$-module with the operation being given by

$$h * \phi(x) = \frac{\partial \phi}{\partial x}(x)h(x).$$

This is , of course just Lie differentiation. Let $\mathcal{F}$ denote the sub Lie algebra of $\mathcal{X}$ generated by all vector fields of the form $f(\ ,u)$ where u is some constant. We finally let $\mathcal{G}$ denote the $\mathcal{F}$-module generated by a function $g \in C^\infty(\mathbb{R}^n)$. Recall that this module is the central object in the description of local controllability of non linear systems.

We let $\mathcal{X}^*$ denote the space of all one forms on $\mathbf{R}^n$, that is just the space of linear combinations of gradients of elements of $C^\infty(\mathbb{R}^n)$. Vector fields act on one forms according to the definition

$$L_h(\omega)(x) = \left(\frac{\partial \omega^*}{\partial x}(x)h(x)\right)^* + \omega(x)\frac{\partial h}{\partial x}(x)$$

where ω is a one form, h is a vector filed and * denotes transpose. A standard result is that if $\omega = d\phi$ then L_h and d commute. Thus $d\mathcal{G}$ is also an $\mathcal{F}$-module. We denote by $d\mathcal{G}(x_0)$ the space of one forms evaluated at the point x_0. The system Σ is said to satisfy the *observability rank condition at* x_0 if the dimension of $d\mathcal{G}(x_0)$ is n. We can now state the canonical theorem from the paper of Herman and Krener, [11].

Theorem 3.1 *If Σ satisfies the observability rank condition at x_0 then Σ is locally weakly observable at x_0.*

The proof is simple and is based on an application of the inverse function theorem. However it should be noted that the proof o the observability rank condition for linear systems is likewise a simple (linear) application of the inverse function theorem. The approach of Hermann and Krener can be pushed somewhat further but ultimately it must be concluded that the methods are essentially local and that only by very artificial hypothesis can the methods give global results.

4 Global Observability and Differentiability Constraints

Again we consider the system Σ but we will now assume that the control is not present. So we are asking if we can distinguish between trajectories of an autonomous dynamical system. Assume for the moment that the functions f and g are analytic and hence that we can construct the Taylor series of the output function. Consider the simple example of the differential equation

$$\dot{x} = \sin x.$$

The solution of the differential equation with initial data $x(0) = a$, $\dot{x}(0) = b$ has Taylor series

$$x(t) = a + bt + \sin a\, t^2/2 + b\cos a\, t^3/6 + (\sin a \cos b - b^2 \sin a)t^4/24 + \cdots.$$

If the system is observed with a linear function of the form

$$y(t) = \alpha x(t) + \beta\dot{x}(t)$$

then the output function has Taylor series

$$\begin{aligned} y(t) &= (\alpha a + \beta b) + (\alpha b + \beta \sin a)t + (\alpha \sin a + \beta b \cos a)t^2/2 \\ &\quad + (\alpha b \cos a + \beta \sin a \cos a - \beta b^2 \sin a)t^3/6 + \cdots \end{aligned}$$

to determine if the system is observable one need only determine if the initial data, a and b, can be recovered from the coefficients of the Taylor series. That is to say can we solve the infinite set of equations

$$\begin{aligned}
(\alpha a + \beta b) &= \tau_1 \\
\alpha b + \beta \sin a &= \tau_2 \\
\alpha \sin a + \beta b \cos a &= \tau_3 \\
\alpha b \cos a + \beta \sin a \cos a - \beta b^2 \sin a &= \tau_4 \\
&\vdots
\end{aligned}$$

for the unknown initial data a and b? This is a formidable task even for this simple system. This approach has been used by Elliott and Tarn and their associates, [14,13].

The basic idea of this attack is very powerful and in general depends on being able to expand the output function as a series of functions in such a way that the coefficients are uniquely determined by the initial data of the differential equation. The series need not be a Taylor series and we will see in later sections that there are times when the expansion can be accomplished in terms of a Dirichlet series or a Fourier series to prove observability. The main shortcoming of the attack is that the resulting set of equations is in general very difficult to solve. One could visualize a much more sophisticated attack on the problem of observability based on the idea of approximation of the output function in much more general function spaces.

A more subtle attack on the problem of observability was begun by Dirk Aeyls in [1]. In this seminal paper Aeyls takes advantage of the fact that the class of Morse-Smale systems have very well behaved trajectories. Let M be a compact C^∞ manifold and let X be a C^∞ vector field on M. We recall that a vector field is *Morse-Smale* if

1. The number of fixed points and periodic orbits is finite and each is hyperbolic.

2. All stable and unstable manifolds intersect transversally.

3. The nonwandering set consists of fixed points and periodic orbits only.

Morse-Smale vector fields have the property that every orbit converges asymptotically to an equilibrium point or to a periodic orbit. Thus there arises the possibility of waiting until an orbit is in the neighborhood of an critical set to attempt to distinguish it from other orbits. The main theorem of the Aeyls, [1], is the following

Theorem 4.1 *Let there be given a Morse-Smale system on a compact manifold with a nonzero number of critical elements and let there be given a smooth output function h into Euclidean space $\mathbb{R}^r$. Then the system is globally observable if*

1. *the rank condition for is satisfied at the critical elements,*
2. *h separate critical points,*
3. *the images of periodic orbits under h are different and every output trajectory corresponding to a closed orbit has minimal period, equal to the period of the closed orbit.*

The rank condition is just the rank condition for local observability in a neighborhood of the critical elements as developed in the paper of Hermann and Krener, [11]. The proof of the theorem is quite technical but the ideas are quite intuitive as is expressed by the following example which again is contained in the paper of Aeyls, [1].

Loosely speaking, the Aeyl's proof of global observability fro Morse-Smale systems is carried out by selecting subsets of the manifold such that at some specific time they are "pushed forward" by the flow into some of the neighborhoods of the critical sets,N_{x_i}, where they are distinguished at some well-picked time instant, either by the rank conditions for local observability, RC, or by the fact taht the observation function distinguishes critical points, MS. This will be illustrated through a discussion of a particular example. Consider the unit sphere $S^2 \subset R^3$ centered at the origin. A flow is defined on S^2 with two critical points, a source at the "north-pole" $x_n = (0,0,1)$ and a sink in the "south-pole" $x_s = (0,0,-1)$. The other orbits of the flow are the "meridian lines." Let h be an output function which assumes different values at x_n and x_s. Let the rank condition be satisfied at both poles. Let N_{x_n} and N_{x_s} be neighborhoods of the critical points. Let V_{x_s} be an open set covering M less P_{x_n}, where $P_{x_n} \subset N_{x_n}$ is a neighborhood of x_n. Any pair of points in V_{x_s} is distinguishable at some time T_α by RC at x_s. There remains to be shown how to distinguish V_{x_s} from its complement in M, or, to be sure, from N_{x_n} which contains the complement. This is carried out in two steps. First, corresponding to T_α, there exists a neighborhood V_{x_n} of x_n, contained in N_{x_n}, such that $\Phi(T_\alpha, V_{x_n}) \subset N_{x_n}$. Therefore, V_{x_s} is distinguishable from V_{x_n} by MS at time T_α. Finally, V_{x_s} is distinguishable from N_{x_n} less V_{x_n} by RC in x_s at some finite time $T_\beta > T_\alpha$. Thus any two orbits are distinguished.

In a general proof problem of saddle points must be faced. This adds conceptual and technical difficulty. Indeed, at this point it is clear – following the ideas explained in the example – how to construct a proof of global observability for the case of a Morse-Smale vectorfield containing a finite number of sources and sinks and no saddles. When saddles are present, one might at first consider neighborhoods $N_{x_{sa}}$ around the saddle points. Then one might be tempted to say that, since all points – except for the sources – eventually wind up in the neighborhoods N_x of the sinks and the saddles, the example again contains all the ideas on how to give a proof in the general case. Such a reasoning would indeed show how to distinguish all pairs of points on the manifold if one is willing to accept an *infinitely long observation interval.* Indeed, points belonging to the stable manifold of one saddle $N_{x_{sa}}$ but sitting close to the stable manifold of another saddle take a *long time* before they are trapped in $N_{x_{sa}}$–the closer they are to the stable manifold of the other saddle, the longer it takes, by continuity of the flow. These points have a somewhat similar behavior to points in the neighborhood of the north-pole-source of the example. Therefore, in the distinguishability process of a formal proof, these points should somehow be treated together with the stable manifolds to which they are close to – and not together with the stable manifolds to which they belong to – *in order to reduce the observation interval to finite time.* The general proof relies heavily on the cellular structure induced on the manifold by the stable manifolds of the Morse-Smale vectorfield.

Another approach to observability in the large was undertaken by Sussman in [18]. There the idea is to define an equivalence relation on the points of the manifold in terms of indistinguishability, that is, two points are equivalent if and only if the orbits emanating form them are not distinguished by the output function. It is fairly easy to see that the relation is an equivalence relation and the natural approach is to consider the quotient of the manifold and the equivalence relation. Sussman gives various conditions under which the resulting object is a manifold and determines conditions under which the system descends to the quotient as an observable system. The resulting system is of course globally observable. Sussman's technique is not really a method for determining if the system is observable but a method for constructing globally observable systems. The main object of his construction was to produce realization of nonlinear systems that are globally observable and controllable.

5 Translational Flows on the Torus

The work of Hermann, Krener, Aeyls, Elliott, Tarn and others on the problems of observability of nonlinear systems and on the problems of global observability did not really attempt to determine necessary conditions for global observability. The conditions that were imposed were of the nature of differentiable conditions of linear control theory. There are just two sets of results that had promise for studying really general systems and had the potential of giving fundamental insight into the problem of observability. The first was the work of Aeyls, using Morse-Smale systems. The conditions that he imposed were not that different from the general conditions of Hermann and Krener but were conditions that attacked the problem of global observability directly rather than obtaining global observability from accidental conditions. The second set of results were the results of Kuo, Elliott and Tarn. Their methods were quite direct and consisted of examining the series expansion of the output function. This approach leads to sufficient conditions for the observability of the systems. A natural extension of their work is to consider the expansion of the output map in terms of series other than a Taylor series.

A first attack on the problem of global observability that did not attempt to impose differential constraints was by Drager and Martin in [9]. There the following problem was considered. Let $\mathbf{T}^n$ denote the $n-$dimensional torus and consider the vector fields on the $\mathbf{T}^n$ that generate the irrational winding lines. These are the simplest vector fields on the torus and are natural to consider. The question that was posed was to determine necessary and sufficient conditions for the observability of these flows. Thus the problem is just to determine the conditions on the observation function that will render the system observable. In [9] this problem wasn't solved but never-the-less an interesting result was observed. If the observation function was assumed to be continuous and was assumed to have a unique maximum then, using a result of Kronecker on the approximation of real numbers with linear rational combinations of irrational numbers it was shown that the system consisting of the irrational winding lines and any continuous function with a unique maximum value was observable. the result is not particularly difficult but it was the first case in the literature that observability was obtained without the assumption of smoothness. It was stated in [9] that the underlying phenomena was ergodicity. However later developments seem to belie this statement.

In a sequel to this paper Byrnes and Crouch, [3], showed that this result followed without the assumption of a unique maximum and that the relevant conditions was that the observation function had a special point, value that was obtained exactly once–a minimum or a maximum. However the only technical requirement was that the observation was continuous. More interestingly they showed that the vector fields could be replaced with vector fields that had the property that they were *minimal distal.* This simply means that the if the initial data for two orbits is separated then the time parameterized orbits remained separated by at least a distance ϵ. The condition of minimality ensures that the orbits are dense. The idea of the proof is to follow one of the orbits until it is sufficiently close to the special point and so that the value of the observation function is distinct from the value on the other orbit. In this paper it was also recognized that the general case should consist of a compact abelian group instead of the torus, $\mathbf{T}^n$.

In the setting of a compact abelian group three more papers quickly followed. Drager and Martin reproved their original result using Fourier analysis on the torus and showed that a sufficient condition was that the observation function should be continuous and that no Fourier coefficient should vanish. This paper was distinguished only by the neatness of the proof and was not a real extension of the theory. the late Douglas McMahon, to whom this paper is dedicated, mad a major extension with the following result, the system consisting of a dense translational flow on a compact abelian group and a continuous observation function, is observable if and only if there is no subgroup that leaves the observation function invariant. Independently Balog, Bennett and Martin showed that the observation function need not be continuous but that the characteristic function of certain 'nice sets' sufficed, i.e. those sets that had the property that they consisted of the interior of their closure provided that there were again no symmetries. These last two results were very satisfying since they mimic the result for linear systems. McMahon proved his results using harmonic analysis techniques and the result of Balog, et. al. was proved using very different techniques of point set topology. There has been a recent announcement by Drager, Foote and McMahon of a result that incorporates both of the above results into a single theorem with the proof being based on techniques from harmonic analysis. Byrnes and McMahon have announced a major new formulation of the theorem in terms of the dual group that generalizes the results of drager and Martin seems to imply the results of Drager Foote and McMahon–namely that the necessary and sufficient conditions for observability is that the characters not represented in the fourier series of the observation function should not include any group of characters.

6 Universal Observability

The winding lines on the torus of the last section are easily proved to be observable by a large set of continuous functions but it is easily seen that there are continuous functions, even analytic functions, which fail to observe even the simplest winding lines. Consider for example the observation function

$$f(\theta) = \cos 2\theta$$

and the uniform rotation on the circle $\mathbf{S}^1$. An easy calculation shows that the flows starting at θ_0 and $\theta_0 + \pi$ are indistinguishable. Interestingly enough the system is locally observable but not observable.

At first thought it would seem that one would always be able to construct an observation function that would render the flows generated by any vector field unobservable. But this is not the case. In a very clever example Douglas McMahon, [16], constructed a vector field on a compact homogeneous space that is observable by *every* continuous nonconstant function. The manifold is the homogeneous space

$$\mathcal{H} = Sl(2, \mathbb{R})/H$$

where H is a nonarithmetical, co-compact subgroup. The flow is generated by the matrix element

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The proof is dependent on technical results from the field of topological dynamics but ultimately rests on the fact that it can be shown that the flow is strongly mixing in the sense that the flow generated in the product space $\mathcal{H} \times \mathcal{H}$ has the property that any flow that initiates at a non diagonal point is dense in the product. It's clear that this property implies that every nonconstant continuous function is observing since we simply select two points at which the function has distinct values and eventually any two distinct orbits will pass through any arbitrarily small neighborhood of the two points.

At the present time this is the only known example of a universally observable vector field. There are a few properties for which it can be demonstrated that any manifold that has a universally observable vector field must possess. First it is clear that the closure of any two orbits must be the entire manifold. For if not there would exist a continuous function that is zero on the union of the orbit and one at some point not in the closure. Thus the two orbits would not be distinguished. Clearly there can be at most one singular orbit and there can exist no periodic orbits. thus there is at most one equilibrium point and every other orbit is dense. Byrnes, Dayawansa and Martin, [24], have shown that there can exist no equilibrium points and hence every orbit is dense. By constructing the one point compactification of the manifold and suitably modifying the flows in the neighborhood of the point at infinity it can by be shown as a corolary of the fact that every orbit is dense that the manifold is compact. From this it follows that every manifold that admits a universally observably vector field has Euler Characterisctic zero. It is not known if every universally observable vector field has the stronger property that orbits are dense in the product space.

McMahon's example is dependent on the existence of non-arithmetical co-compact subgroups of $Sl(2, \mathbb{R})$. For $n > 2$ such subgroups fail to exist due to rigidity theorems. Thus his construction does not generalize. For two dimensional manifolds the only possible example is the 2-torus. At this point it is undecided whether or not the torus admits a universally observable vector field. This question is discussed but not decided in [24].

7 Another Point Of View

Let (X, f) be an smooth observable system on a compact n-dimensional manifold M. The integral curves of X, $\phi_t(x_0)$ composed with f, $f(\phi_t(x_0))$, defines a map from the manifold into the space of real valued functions of a real variable. Since the manifold is compact the functions are bounded and we may as well assume we are mapping into the space of continuous functions with the supremum topology. Elementary properties of the system imply that the mapping

$$x_0 \longmapsto f(\phi_t(x_0))$$

is smooth. Observability implies that the map is one to one and the various forms of local observability imply that the mapping is locally one to one. We would like for the mapping to be nonsingular but in general additional conditions are necessary to ensure this.

A very realistic problem has been posed in Aeyels, [22] and in Smith and Martin, [21], which has interesting implications in the above setting. Suppose that instead of the function $f(\phi_t(x_0))$ we are given the value of this function at n preselected times, $t_1, \, ,t_2, \, \cdots \, ,t_n$. Does this preserve observability? In general the answer is no, even in the case of linear systems, [21]. In the case of nonlinear systems a positive answer would have provided a one-to-one mapping of M into $\mathbb{R}^n$. This wouldn't have necessarily have been impossible but would have certainly have been pathological. Aeyels showed that generically it suffices to evaluate the output function at $2n+1$ points and a mapping of the manifold into $\mathbb{R}^{2n+1}$ is obtained. It is possible to extend Aeyel's result and construct a nonsingular map from M into $\mathbb{R}^{2n+1}$—the easy imbedding theorem. The paper of Bendsøe, [23], is relevant in that he constructs a globally observable vector field on an arbirary compact manifold. An interesting open problem is to determine if there

[9] C. Byrnes and A. Isidori. "Global feedback stabilization of nonlinear systems," **Proc. of the 24th IEEE conf. on Dec. and Control**, Ft. Lauderdale, 1985.

[10] J. Clark, C. Ong, T. Tarn and G. Huang. "Quantum nondemolition filters," **Math. Systems Theory**, Vol 18, pp33-55, 1985.

[11] W. Dayawansa and C. Martin, "Observing linear dynamics with polynomial output functions," to appear in **System and Control Letters**.

[12] L. Drager and C. Martin. "Global observability of a class of nonlinear discrete time systems," **System and Control Letters**, Vol 6, pp. 65-68, 1985.

[13] L. Drager and C. Martin. "Observability of flows on the torus: an application of number theory," to appear **Math. Systems Theory**.

[14] L. Drager and C. Martin. "Global observability of ergodic translations on compact groups," to appear.

[15] D. Gilliam, Z. Li and C. Martin, "The observability of the heat equation with sampling in time and space," submitted.

[16] R. Hermann and A. Krener. "Nonlinear controllability and observability', **IEEE Tran. Aut. Cont.**, Vol AC-22, pp. 728-740, 1977.

[17] G. Huang, T, Tarn and J. Clark. "On the controllability of quantum-mechanical Systems," **Journal of Mathematical Physics**, Vol 24, pp.2608-2618, 1983.

[18] S. Kuo, D. Elliott and T. Tarn. "Observability of nonlinear systems," **Information and Control**, Vol 22, pp.89-99, 1973.

[19] S. Kuo, D. Elliott and T. Tarn. "Exponential observers for nonlinear dynamic systems," **Information and Control**, Vol 29, pp. 204-216, 1976.

[20] C. Martin and J. Smith, "Discrete observability of linear systems," to appear.

[21] C. Ong, G. Huang, T. Tarn and J. Clark. "Invertibility of quantum-mechanical control systems" **Math. Systems Theory**, Vol 17, pp. 335-350, 1985.

[22] D. McMahon. "An example of a universally observable dynamical system," to appear in **System and control Letters.**

[23] T. Tarn and Y. Rasis. "Observers for nonlinear stochastic systems," **IEEE Tran. Aut. Cont.**, Vol AC-21, pp. 444-448, 1976.

[24] H. Sussman, "A generalization of the closed subgroup theorem to quotients of arbitrary manifolds," **J. of Differential Geometry**, vol. 10, pp. 151-166, 1975.

ADAPTIVE CONTROLLERS FOR UNCERTAIN DYNAMICAL SYSTEMS

Martin Corless
School of Aeronautics and Astronautics
Purdue University
West Lafayette, Indiana 47907, USA

George Leitmann
Department of Mechanical Engineering
University of California
Berkeley, California 94720, USA

ABSTRACT

The fundamental feedback control problem of obtaining desired system behavior in the presence of uncertainties is considered for a class of uncertain systems described by differential equations.

Taking a deterministic point of view, a class of adaptive controllers which yield stable behavior is proposed.

The use of these controllers is illustrated by examples and applications.

I. INTRODUCTION

In order to control the behavior of a system in the "real" world, be it physical, biological or socio-economic, the system analyst seeks to capture the system's salient features in a mathematical model. This abstraction of the "real" system always contains uncertain elements; these may be parameters, constant or varying, which are unknown or imperfectly known, or they may be unknown or imperfectly known inputs into the system. Despite such imperfect knowledge about the chosen mathematical model, one often seeks to devise controllers which will "steer" the system in some desired fashion, for example so that the system response will approach or track a desired reference response; by suitable definition of the system (state) variables, such problems can always be cast into the form of stability problems.

Two main avenues are open to the analyst seeking to control an uncertain dynamical system. He may choose a stochastic approach in which information about the uncertain elements as well as about the system response is statistical in nature; e.g., see Refs. (1-2). Loosely speaking, when modelling via random variables, one is content with desirable behavior on the average. The other approach to the control of uncertain systems, and the one for which we shall opt in the present discussion, is deterministic. Available, or assumed, information about uncertain elements is deterministic in nature. Here one seeks controllers which assure the desired response of the dynamical system.

In this paper, the mathematical model is embodied in ordinary differential equations, the state equations of the system. We divide the systems under consideration into three subclasses depending on the type of potentially destabilizing uncertainties present in the system description and on the way the control enters into the description. For each of the systems considered there exists a state feedback controller which assures that the zero state is globally uniformly asymptotically stable. However, these controllers depend on constants in the system description which are not known; e.g., such constants are the values of unknown constant disturbances or unknown bounds on time-varying parameters or inputs. We propose controllers which may be regarded as adaptive versions of the feedback controllers mentioned above; in place of the unknown constants, one employs quantities which change or adapt as the state of the system evolves. Under some circumstances, these adaptive quantities may be considered to be estimates of the unknown constants. The method of devising these adaptive controllers is based on the constructive use of Lyapunov theory as suggested, in a somewhat different context, in Refs. (3-8).

II. SYSTEMS UNDER CONSIDERATION

All of the systems under consideration belong to one main class S4. However, we introduce first three subclasses S1, S2, S3, each of which is included in the main class.

System Class S1

The systems in this class are described by

$$\dot{x}(t) = f(t,x(t)) + B^{(1)}(t,x(t))[\psi(u^{(1)}(t) + d_a) + d_b], \tag{2.1}$$

where $t \in \mathbb{R}$, $x(t) \in \mathbb{R}^n$ is the state and $u^{(1)}(t) \in \mathbb{R}^{m_1}$ is the control; d_a and d_b are *unknown* (arbitrary) constants and the functions $f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $B\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n\times m_1}$, and $\psi\colon \mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$ are *uncertain*, i.e., they are not assumed known but are only assumed to satisfy certain conditions (A1, A2, A3(1), A5(1)).

Concerning the function f we introduce the following assumption.

Assumption A1. *1) f is Caratheodory*[1] *and*

$$f(t,0) = 0 \quad \forall\, t \in \mathbb{R}\,. \tag{2.2}$$ [2]

2)[3] *There exist a C^1 function $V\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_+$ and functions $\gamma_1, \gamma_2, \gamma_3\colon \mathbb{R}_+ \to \mathbb{R}_+$, where γ_1, γ_2 belong to class KR and γ_3 belongs to class K, such that for all $(t,x) \in \mathbb{R} \times \mathbb{R}^n$*

[1] See Ref. (9), Appendix, sec. A, or just note that if f is continuous, it is Caratheodory.

[2] We use "0" to denote a zero vector.

[3] See Appendix.

$$\gamma_1(||x||) \leq V(t,x) \leq \gamma_2(||x||), \tag{2.3}$$

$$\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)f(t,x) \leq -\gamma_3(||x||). \tag{2.4}$$

Assumption A1 asserts that there exists a Lyapunov function V which guarantees that the zero state is a g.u.a.s. (globally uniformly asymptotically stable) equilibrium state of the system described by

$$\dot{x}(t) = f(t,x(t)); \tag{2.5}$$

see Refs. (5, 10-12).

Concerning the function ψ the following is assumed.

Assumption A2. *1) ψ is onto; i.e., given any $\eta \in \mathbb{R}^{m_1}$, there exists $w \in \mathbb{R}^{m_1}$ such that $\psi(w) = \eta$.*

This assumption and A1 imply that, given any constants d_a, $d_b \in \mathbb{R}^{m_1}$, there exists a constant control $u^{(1)}(t) \equiv v$ given by

$$v = w - d_a, \ \psi(w) = -d_b, \tag{2.6}$$

such that the zero state is a g.u.a.s. equilibrium point of (2.1).

2) ψ^T is the derivative of some C^1 function Ψ: $\mathbb{R}^{m_1} \to \mathbb{R}$.

3) For each $w \in \mathbb{R}^{m_1}$ the function W: $\mathbb{R}^{m_1} \to \mathbb{R}$ given by

$$W(\hat{w}) = \Psi(\hat{w}) - \Psi(w) - \frac{\partial \Psi}{\partial w}(w)(\hat{w} - w) \tag{2.7}$$

satisfies

$$\hat{w} \neq w \Rightarrow W(\hat{w}) > 0, \tag{2.8}$$

$$\lim_{||\hat{w}|| \to \infty} W(\hat{w}) = \infty. \tag{2.9}$$

We also make the following additional assumptions.

Assumption A3. *1) The function $B^{(1)}$ is strongly Caratheodory.*[4]

Assumption A4. *One of the following two conditions is satisfied.*

C1. *There exists a continuous function γ_4: $\mathbb{R}_+ \to \mathbb{R}_+$ which satisfies*

$$\lim_{r \to \infty} \gamma_4(r) = \infty, \tag{2.10}$$

such that for all $(\hat{w},w) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_1}$

$$[\psi(\hat{w}) - \psi(w)]^T(\hat{w} - w) \geq \gamma_4(||\hat{w} - w||)||\hat{w} - w||. \tag{2.11}$$

C2. *For each $d \geq 0$ there exists $b_1(d) \geq 0$ such that for all $(t,x) \in \mathbb{R} \times \mathbb{R}^n$*

$$||x|| \leq d \Rightarrow ||\alpha^{(1)}(t,x)|| \leq b_1(d), \tag{2.12}$$

[4] See Ref. (9), Appendix, sec. A, or just note that if $B^{(1)}$ is continuous, it is strongly Caratheodory.

where

$$\alpha^{(1)}(t,x) = B^{(1)^T}(t,x)\,\frac{\partial V^T}{\partial x}\,(t,x). \tag{2.13}$$

Assumption A5. *1) At each $t \in \mathbb{R}$, $\alpha^{(1)}(t,x(t))$ is* known.

Remarks 2.1. 1) If f is linear time-invariant, i.e.,

$$f(t,x) = Ax \qquad \forall (t,x) \in \mathbb{R} \times \mathbb{R}^n, \tag{2.14}$$

where $A \in \mathbb{R}^{n\times n}$ and A is asymptotically stable (i.e., all of its eigenvalues have negative real parts), then A1(1) is satisfied and A1(2) is satisfied by taking any positive-definite symmetric $Q \in \mathbb{R}^{n\times n}$ and letting

$$V(t,x) = \frac{1}{2}\,x^T P x \qquad \forall (t,x) \in \mathbb{R} \times \mathbb{R}^n \tag{2.15}$$

where $P \in \mathbb{R}^{n\times n}$ is the unique positive-definite symmetric solution of

$$PA + A^T P + Q = 0; \tag{2.16}$$

see Refs. (5, 10-12). If, in addition, $B^{(1)}$ is constant, i.e.,

$$B^{(1)}(t,x) = B \qquad \forall (t,x) \in \mathbb{R} \times \mathbb{R}^n \tag{2.17}$$

where $B \in \mathbb{R}^{n\times m_1}$, then

$$\alpha^{(1)}(t,x) = B^T P x \tag{2.18}$$

and C2, and hence A4, is satisfied.

2) As a particular example of a function which satisfies A2 and C1 consider any function ψ: $\mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$ given by

$$\psi(w) = Fw, \tag{2.19}$$

where $F \in \mathbb{R}^{m_1\times m_1}$ is symmetric positive-definite. The existence of F^{-1} implies A2(1) is satisfied. Assumptions A2(2) and A2(3) are shown to hold by letting

$$\Psi(w) = \frac{1}{2}\,w^T F w,$$

and C1 is assured with

$$\gamma_4(r) = \lambda_{\min}(F)r, \tag{2.20}$$

where $\lambda_{\min}(F)$ denotes the smallest eigenvalue of F, and $\lambda_{\min}(F) > 0$.

3) As a more general example of a function which satisfies A2 see Ref. (9).

4) Assumption A5(1) is made in order to ensure that there is sufficient information available to implement the proposed controllers for this system class. Note that this assumption is completely independent of ψ and does not require complete knowledge of f and $B^{(1)}$. For example, some of the controlled systems presented in Refs. (13-15) contain an uncertain f which satisfies A1 for a known V. Ther e $B^{(1)}$ is known, so that the function $\alpha^{(1)}$ is known. For another example see sec. V.A.

5) For some previous literature on controllers for systems subject to unknown constant disturbances see Refs. (16-23).

System Class S2

The systems in this class are described by

$$\dot{x}(t) = f(t,x(t)) + B^{(2)}(t,x(t))[Fu^{(2)}(t) + Dh(t,x(t))] \tag{2.21}$$

where t,x(t), and f are as defined for S1 and $u^{(2)}(t) \in \mathbb{R}^{m_2}$ is the control; the matrix $D \in \mathbb{R}^{m_2 \times p}$ is *unknown* and the matrix $F \in \mathbb{R}^{m_2 \times m_2}$ and functions $B^{(2)}: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n \times m_2}$ and $h: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^p$ are *uncertain*, i.e., they are not assumed known but are only assumed to satisfy certain conditions (A3(2), A5(2), A6).

In addition to assuming that f satisfies A1, we also make the following assumptions for this class.

Assumption A6. *The matrix F is symmetric positive-definite.*

Note that this assumption and A1 imply that for each F and D there exists a state feedback control given by

$$u^{(2)}(t) = Kh(t,x(t)), \; K = -F^{-1}D, \tag{2.22}$$

such that the zero state is a g.u.a.s. equilibrium point of (2.21). However, the matrices F and D are not assumed to be known.

Assumption A3. *2) The functions $B^{(2)}$ and h are strongly Caratheodory.*

Assumption A5. *2) At each $t \in \mathbb{R}$, $\alpha^{(2)}(t,x(t))$ and $h(t,x(t))$ are* known, *where for all $(t,x) \in \mathbb{R} \times \mathbb{R}^n$*

$$\alpha^{(2)}(t,x) = B^{(2)^T}(t,x) \frac{\partial V^T}{\partial x}(t,x). \tag{2.23}$$

The following condition, which is *not* an assumption, will affect the choice of one of the parameters in the proposed controllers for this system class.

C3. *For each $d \geq 0$ there exists $b_2(d) \geq 0$ such that for all $(t,x) \in \mathbb{R} \times \mathbb{R}^n$*

$$||x|| \leq d \Rightarrow ||h(t,x)|| \; ||\alpha^{(2)}(t,x)|| \leq b_2(d). \tag{2.24}$$

Remarks 2.2. 1) Quite frequently the error equation which arises in the problem of requiring a linear time-invariant system with unknown parameters to track a reference model falls into this class of systems; see Refs. (6-8, 24-32).

2) For a particular example of a system in this class see sec. V.B.

System Class S3

In this class we consider systems which contain potentially destabilizing uncertainties of a more general nature than those considered in S1 and S2. The systems are described by

$$\dot{x}(t) = f(t,x(t)) + B^{(3)}(t,x(t))g(t,x(t),u^{(3)}(t)) \tag{2.25}$$

where t,x(t), and f are as defined for S1 and $u^{(3)}(t) \in \mathbb{R}^{m_3}$ is the control; the functions $B^{(3)}: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n \times m_3}$ and $g: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{m_3} \to \mathbb{R}^{m_3}$ are *uncertain*, i.e., they are not assumed known but are only assumed to satisfy certain conditions (A3(3), A5(3), A7). In addition to assuming that f satisfies A1, the following assumptions are made for this class.

Assumption A7. *1) There exist an* uncertain *function* ρ*:* $\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_+$ *and an* unknown *constant* $\beta_o > 0$ *such that for all* $(t,x,u) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{m_3}$

$$u^T g(t,x,u) \geq \beta_o ||u|| \; [||u|| - \rho(t,x)]. \tag{2.26}$$

2) There exist an unknown *constant* $\beta \in (0,\infty)^k$ *and a* known *function* Π*:* $\mathbb{R} \times \mathbb{R}^n \times (0,\infty)^k \to \mathbb{R}_+$ *such that for all* $(t,x) \in \mathbb{R} \times \mathbb{R}^n$

$$\rho(t,x) = \Pi(t,x,\beta). \tag{2.27}$$

That is, we do not assume that the bound $\rho(t,x)$ is known; we only assume that it depends in a known manner on an unknown constant vector β.

3) For each $(t,x) \in \mathbb{R} \times \mathbb{R}^n$*, the function* $\Pi(t,x,\cdot):(0,\infty)^k \to \mathbb{R}_+$ *is* C^1*, concave*[5]*, and non-decreasing with respect to each coordinate of its argument,* β.

Assumption A3. *3) The function* $B^{(3)}$ *is Caratheodory and* g*,* Π*, and* $\frac{\partial \Pi}{\partial \beta}$ *are strongly Caratheodory.*

Assumption A5. *3) At each* $t \in \mathbb{R}$*,* $\alpha^{(3)}(t,x(t))$ *and* $x(t)$ *are* known *where for all* $(t,x) \in \mathbb{R} \times \mathbb{R}^n$

$$\alpha^{(3)}(t,x) = B^{(3)^T}(t,x) \frac{\partial V^T}{\partial x} (t,x). \tag{2.28}$$

Remarks 2.3. 1) In the earlier literature (see Refs. (15, 33-40)) systems of this class have been considered where g is of the form

$$g(t,x,u) = [I + E(t,x)]u + e(t,x), \tag{2.29}$$

and $E(t,x)$ and $e(t,x)$ satisfy

$$\left.\begin{array}{ll} ||E(t,x)|| \leq c, & c < 1, \\ ||e(t,x)|| \leq \rho_o(t,x), & \end{array}\right. \quad \forall (t,x) \in \mathbb{R} \times \mathbb{R}^n \tag{2.30}$$

for a known constant c and function ρ_o. Hence A7(1) is satisfied by taking

$$\beta_o = 1-c, \tag{2.31}$$

$$\rho(t,x) = \rho_o(t,x)/(1-c). \tag{2.32}$$

In other words, in these references β_o and ρ are assumed completely known whereas here only A7(2) and A7(3) are assumed.

2) As an example of a function which satisfies the assumptions on Π, consider any function Π: $\mathbb{R} \times \mathbb{R}^n \times (0,\infty)^k \to \mathbb{R}_+$ given by

$$\Pi(t,x,\beta) = \kappa_o(t,x) + \kappa^T(t,x)\beta$$

where κ_o: $\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_+$ and κ: $\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_+^k$ are known strongly Caratheodory functions.

[5] That is, $-\Pi(t,x,\cdot)$ is convex.

3) For a particular example of a system in this class, see sec. V.C. For applications of this class, see Refs. (44-46).

System Class S4

This is the main class of systems. It includes each of the preceding subclasses, S1, S2, and S3. Systems in this class are described by

$$\begin{aligned}
&\dot{x}(t) = f(t,x(t)) + G(t,x(t), u(t)),\\
&G(t,x,u) = B^{(1)}(t,x)[\psi(u^{(1)} + d_a) + d_b] + B^{(2)}(t,x)[Fu^{(2)} + Dh(t,x)]\\
&\qquad + B^{(3)}(t,x)g(t,x,u^{(3)}), \qquad (2.33)\\
&(u^{(1)^T} u^{(2)^T} u^{(3)^T}) = u^T,
\end{aligned}$$

where all quantities are as previously defined.

All of the preceding assumptions A1 - A7 are made for this class.

For an example of a system in this class see Ref. (47).

III. PROPOSED CONTROLLERS

In this section we present a class of controllers which guarantee the desired stability properties for each of the system classes considered in the previous section. In essence, each controller proposed is an adaptive version of a zero state stabilizing controller which depends on unknown constants.

Controller Class C1 (for S1)

The controllers in this class are given by

$$u^{(1)}(t) = \hat{v}(t) - \bar{\ell}_1\ell_1\tilde{\alpha}(t), \qquad (3.1)$$

$$\dot{\hat{v}}(t) = -\ell_1\tilde{\alpha}(t), \qquad (3.2)$$

where $\hat{v}(t_o) \in \mathbb{R}^{m_1}$ is arbitrary,

$$\ell_1 > 0,\ \bar{\ell}_1 \geq 0, \qquad (3.3)$$

$\bar{\ell}_1 > 0$ if condition C2 is not satisfied, and

$$\tilde{\alpha}(t) = \alpha^{(1)}(t,x(t)). \qquad (3.4)$$

Note that $u^{(1)}(t)$ is also given by

$$u^{(1)}(t) = \hat{v}(t) + \bar{\ell}_1\dot{\hat{v}}(t) \qquad (3.5)$$

$$u^{(1)}(t) = -\bar{\ell}_1\ell_1\tilde{\alpha}(t) - \ell_1\int_{t_o}^{t}\tilde{\alpha}(\tau)d\tau + \hat{v}(t_o). \qquad (3.6)$$

Thus, these controllers can be considered as versions of the classical PI (proportional plus integral) controllers.

For an application of these controllers, see sec. V.A.

Controller Class C2 (for S2)

The controllers in this class are given by

$$u^{(2)}(t) = \hat{K}(t)h(t,x(t)) - \bar{\ell}_2\tilde{\alpha}(t)h^T(t,x(t))\Gamma h(t,x(t)), \tag{3.7}$$

$$\dot{\hat{K}}(t) = -\tilde{\alpha}(t)h^T(t,x(t))\Gamma, \tag{3.8}$$

where $\hat{K}(t) \in \mathbb{R}^{m_2 \times p}$, $\hat{K}(t_o)$ being arbitrary, Γ is any positive-definite symmetric $p \times p$ matrix,

$$\bar{\ell}_2 \geq 0, \tag{3.9}$$

$\bar{\ell}_2 > 0$ if condition C3 is not satisfied, and

$$\tilde{\alpha}(t) = \alpha^{(2)}(t,x(t)). \tag{3.10}$$

Note that $u^{(2)}(t)$ is also given by

$$u^{(2)}(t) = [\hat{K}(t) + \bar{\ell}_2\dot{\hat{K}}(t)]h(t,x(t)) \tag{3.11}$$

$$u^{(2)}(t) = [-\bar{\ell}_2\Omega(t) - \int_{t_o}^{t} \Omega(\tau)d\tau + \hat{K}(t_o)]h(t,x(t)) \tag{3.12}$$

$$\Omega(t) = \tilde{\alpha}(t)h(t,x(t))^T\Gamma. \tag{3.13}$$

It may readily be seen that an S2 system with

$$h(t,x) = 1 \qquad \forall(t,x) \in \mathbb{R} \times \mathbb{R}^n \tag{3.14}$$

is also an S1 system and its C2 controllers are the same as its C1 controllers. C1 controllers are special cases of C2 controllers.

For an application of C2 controllers, see sec. V.B.

Controller Class C3 (for S3)

The controllers in this class are given by

$$u^{(3)}(t) = p(t,x(t), \hat{\beta}(t), \varepsilon(t)), \tag{3.15}$$

$$p(t,x,\hat{\beta},\varepsilon) = -\Pi(t,x,\hat{\beta})s(t,x,\hat{\beta},\varepsilon), \tag{3.16}$$

$$\dot{\hat{\beta}}(t) = L^{(3)} \frac{\partial\Pi^T}{\partial\beta} (t,x(t),\hat{\beta}(t))||\alpha^{(3)}(t,x(t))||, \tag{3.17}$$

$$\dot{\varepsilon}(t) = -\ell_4\varepsilon(t), \tag{3.18}$$

$$\hat{\beta}(t_o) \in (0,\infty)^k, \; \varepsilon(t_o) \in (0,\infty), \tag{3.19}$$

$$\ell_4 > 0, \tag{3.20}$$

where $L^{(3)} \in \mathbb{R}^{k \times k}$ is diagonal with positive elements and s: $\mathbb{R} \times \mathbb{R}^n \times (0,\infty)^{k+1} \to \mathbb{R}^{m_3}$ is *any* strongly Caratheodory function which satisfies

$$s(t,x,\hat{\beta},\varepsilon)||\alpha^{(3)}(t,x)|| = ||s(t,x,\hat{\beta},\varepsilon)||\alpha^{(3)}(t,x), \tag{3.21}$$

i.e., the two vectors have the same direction, and

$$||\mu(t,x,\hat{\beta})|| > \varepsilon \Rightarrow s(t,x,\hat{\beta},\varepsilon) = \frac{\alpha^{(3)}(t,x)}{||\alpha^{(3)}(t,x)||}, \tag{3.22}$$

$$\mu(t,x,\hat{\beta}) = \Pi(t,x,\hat{\beta})\alpha^{(3)}(t,x) \tag{3.23}$$

for all $(t,x,\hat{\beta},\varepsilon) \in \mathbb{R} \times \mathbb{R}^n \times (0,\infty)^{k+1}$. A particular example of such a function s is given by

$$s(t,x,\hat{\beta},\varepsilon) = \text{sat}[\mu(t,x,\hat{\beta})/\varepsilon] \tag{3.24}$$

where

$$\text{sat}(\eta) = \begin{cases} \eta & , \quad ||\eta|| \leq 1, \\ \dfrac{\eta}{||\eta||} & , \quad ||\eta|| > 1. \end{cases} \tag{3.25}$$

These controllers can be considered as adaptive versions of those presented in Refs. (34-38) and modified in Ref. (39) for systems of this class where g is of the form considered in Remark 2.3(1).

For applications of C3 controllers, see sec. V.C and Refs. (44-46).

Controller Class C4 (for S4)

Roughly speaking, the controllers in this class are combinations of controllers from the preceding three classes. More precisely, they are given by

$$u^T(t) = (u^{(1)^T}(t) \quad u^{(2)^T}(t) \quad u^{(3)^T}(t)), \tag{3.26}$$

where $u^{(1)}(t)$, $u^{(2)}(t)$, and $u^{(3)}(t)$ are given by controllers in classes C1, C2, and C3, respectively.

For an application of a C4 controller, see Ref. (47).

IV. PROPERTIES OF SYSTEMS WITH PROPOSED CONTROLLERS

Before stating a theorem, let us consider any system belonging to class S4 subject to any corresponding controller in class C4. By defining the parameter "estimate" vector

$$\hat{q} = (\hat{v}^T \; \hat{k}_1 \; \hat{k}_2 \ldots \hat{k}_{m_2} \; \hat{\beta}^T \; \varepsilon)^T, \tag{4.1}$$

where $\hat{k}_i$, $i = 1,2,\ldots,m_2$, are the rows of $\hat{K}$, and by appropriately defining $\bar{f}^{(1)}$: $\mathbb{R} \times \mathbb{R}^n \times \mathcal{Q} \to \mathbb{R}^n$ and $\bar{f}^{(2)}$: $\mathbb{R} \times \mathbb{R}^n \times \mathcal{Q} \to \mathbb{R}^r$, (see Ref. (9), Appendix) where

$$\mathcal{Q} = \mathbb{R}^{m_1} \times \mathbb{R}^{m_2 p} \times (0,\infty)^{k+1}, \tag{4.2}$$

$$r = m_1 + m_2 p + k + 1, \tag{4.3}$$

such a controlled system can be described by

$$\begin{aligned} \dot{x}(t) &= \bar{f}^{(1)}(t,x(t),\hat{q}(t)), \\ \dot{\hat{q}}(t) &= \bar{f}^{(2)}(t,x(t),\hat{q}(t)). \end{aligned} \tag{4.4}$$

This is a system whose complete state $(x,\hat{q})$ belongs to $\mathbb{R}^n \times \mathcal{Q}$.

Defining the parameter vector

$$q = (v^T \; k_1 \; k_2 \ldots k_{m_2} \; \beta^T \; 0)^T, \tag{4.5}$$

where k_i, $i = 1,2,\ldots,m_2$, are the rows of K, we are now ready to state a theorem.

Theorem 4.1. *Consider any system belonging to class S4 and subject to any corresponding controller in class C4. The resulting controlled system can be described by (4.4) and has the following properties.*

P1) Existence of Solutions. *For each $(t_o, x_o, \hat{q}_o) \in \mathbb{R} \times \mathbb{R}^n \times Q$ there exists a solution $(x(\cdot),\hat{q}(\cdot))$: $[t_o,t_1) \to \mathbb{R}^n \times Q$ of (4.4) with $(x(t_o),\hat{q}(t_o)) = (x_o,\hat{q}_o)$.*

P2) Uniform Stability of (0,q). *For each $\eta > 0$ there exists $\delta > 0$ such that if $(x(\cdot), \hat{q}(\cdot))$ is any solution of (4.4) with $||x(t_o)||$, $||\hat{q}(t_o) - q|| < \delta$ then $||x(t)||$, $||\hat{q}(t) - q|| < \eta$ for all $t \in [t_o,t_1)$.*

P3) Uniform Boundedness of Solutions. *For each r_1, $r_2 > 0$ there exist $d_1(r_1,r_2)$, $d_2(r_1,r_2) \geq 0$ such that if $(x(\cdot), \hat{q}(\cdot))$ is any solution of (4.4) with $||x(t_o)|| \leq r_1$ and $||\hat{q}(t_o) - q|| \leq r_2$ then $||x(t)|| \leq d_1(r_1,r_2)$ and $||\hat{q}(t) - q|| \leq d_2(r_1,r_2)$ for all $t \in [t_o,t_1)$.*

P4) Extension of Solutions. *Every solution of (4.4) can be extended into a solution defined on $[t_o,\infty)$.*

P5) Convergence of x(·) to Zero. *If $(x(\cdot), \hat{q}(\cdot))$: $[t_o,\infty) \to \mathbb{R}^n \times Q$ is a solution of (4.4) then*

$$\lim_{t \to \infty} x(t) = 0. \tag{4.6}$$

Proof. The details of a proof may be found in Ref. (9), Appendix, sec. D.

Remark 4.1. The above theorem also applies to a system of class S1, S2, or S3 subject to a controller belonging to C1, C2, or C3, respectively. For example, an S1 system subject to a C1 controller may be considered an S4 system subject to a C4 controller with $B^{(2)}(t,x) = B^{(3)}(t,x) \equiv 0$.

V. APPLICATIONS AND EXAMPLES

A. LUR'E TYPE SYSTEMS

Consider a system described by

$$\dot{z}(t) = Az(t) + B\psi(u(t) + d_a) + d, \tag{5.1}$$

$$y(t) = Cz(t), \tag{5.2}$$

where $t \in \mathbb{R}$, $z(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^m$ is the output; $d_a \in \mathbb{R}^m$ and $d \in \mathbb{R}^n$ are *unknown* constant disturbances; the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{m\times n}$ are *uncertain*, i.e., they are not assumed known but are only assumed to satisfy the following assumptions. 1) A is asymptotically stable. 2) (A,B) is controllable. 3) (C,A) is observable. 4) The transfer function G, given by

$$G(s) = C(sI - A)^{-1}B, \tag{5.3}$$

is strictly positive real; see Refs. (48,49). 5) ψ: $\mathbb{R}^m \to \mathbb{R}^m$ satisfies A2.

For a given constant reference output $y^* \in \mathbb{R}^m$, it is desired to obtain a controller which assures that, for any initial condition of the controlled system, z(·) is bounded and

$$\lim_{t \to \infty} y(t) = y^*. \tag{5.4}$$

It is assumed additionally that, for each t, y(t) is known.

If we let

$$x(t) = z(t) - z^o, \tag{5.5}$$

$$z^o = A^{-1}B(CA^{-1}B)^{-1}(y^* + CA^{-1}d) - A^{-1}d, \tag{5.6}$$

then, utilizing (5.5), (5.6), (5.1), and (5.2), the system can be described by

$$\dot{x}(t) = Ax(t) + B[\psi(u(t) + d_a) + d_b], \tag{5.7}$$

$$d_b = (CA^{-1}B)^{-1}(y^* + CA^{-1}d), \tag{5.8}$$

and the output tracking error is given by

$$e(t) = y(t) - y^* = Cx(t). \tag{5.9}$$

System description (5.7) is in the form of (2.1) with

$$f(t,x) = Ax, \quad B^{(1)}(t,x) = B, \quad u^{(1)}(t) = u(t). \tag{5.10}$$

As a consequence of assumptions (1) - (4), there exist positive-definite symmetric matrices P, Q $\epsilon\ \mathbb{R}^{n\times n}$ such that

$$PA + A^TP + Q = 0, \qquad C = B^TP; \tag{5.11}$$

see Refs. (48, 49). Hence, assumptions A1, A2, A3(1), A4, and A5(1) hold (see Remark 2.1(1)) and system description (5.7) belongs to S1.

Taking u(t) to be given by a C1 controller for (5.7), one has (utilizing (3.1) - (3.4), (2.18), (5.11), and (5.9))

$$u(t) = \hat{v}(t) - \bar{\ell}_1\ell_1 e(t), \tag{5.12}$$

$$\dot{\hat{v}}(t) = -\ell_1 e(t),$$

where $\ell_1 > 0$ and $\bar{\ell}_1 \geq 0$.

As a consequence of Theorem 4.1, the use of control given by (5.12) results in $(x(\cdot), \hat{v}(\cdot))$ being bounded and $\lim_{t\to\infty} x(t) = 0$. Hence z(·) is bounded and $\lim_{t\to\infty} y(t) = y^*$.

For numerical simulation, we have taken a two-dimensional system described by

$$\dot{z}_1(t) = z_2(t) + d_1,$$

$$\dot{z}_2(t) = -a_1z_1(t) - a_2z_2(t) + \psi(u(t) + d_a) + d_2,$$

$$y(t) = b_1z_1(t) + z_2(t),$$

$$a_1 = 6, \quad a_2 = 5, \quad b_1 = 1, \quad \psi(w) = w^3, \quad d_1 = 3, \quad d_2 = d_a = 1.$$

For $y^* = 1$, three different controls were considered.

1) The constant control, $u(t) \equiv v$, which assures the desired performance but which requires knowledge of the system and disturbances. For this example v = 1.

2) Control given by (5.12) with $\ell_1 = 1$, $\bar{\ell}_1 = 0$ and $\hat{v}(0) = 0$.

3) Control given by (5.12) with $\ell_1 = 1$, $\bar{\ell}_1 = 1$ and $\hat{v}(0) = 0$.

The results of simulations with $z_1(0) = z_2(0) = 0$ are presented graphically in Figures 5.1(1) - 5.1(3).

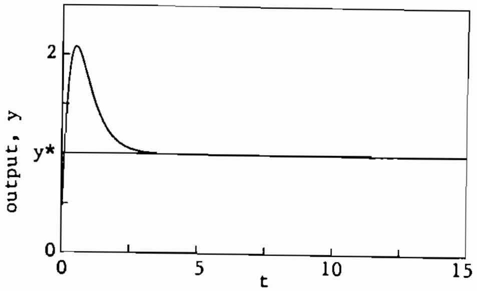


Fig. 5.1(1). Output response for constant control.

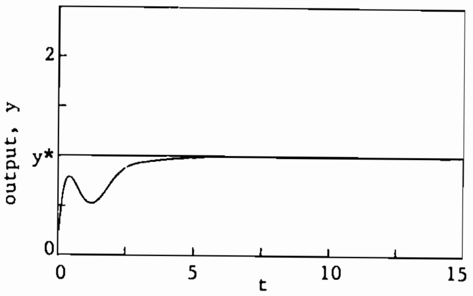


Fig. 5.1(2). Output response for $\ell_1=1$, $\bar{\ell}_1=0$.

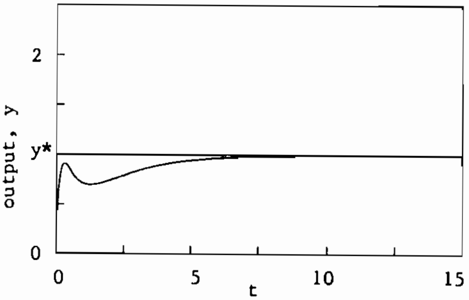


Fig. 5.1(3). Output response for $\ell_1=1$, $\bar{\ell}_1=1$.

B. A BIOLOGICAL EXAMPLE

Consider a species of animals whose population dynamics can be described by

$$\dot{y}(t) = (r/k)y(t)[k - y(t)] + u(t),$$
$$y(t) \geq 0 \qquad (5.13)$$

where y(t) is the biomass of the species and u(t) is a control on the biomass growth or decay rate at time t, r is the intrinsic growth rate and k is the environmental carrying capacity; see Ref. (50). Suppose that r and k are *unknown* positive constants but the biomass is accessible.

The control problem considered here may be stated as follows. Given a "desirable" positive biomass y*, obtain a control policy the utilization of which assures that, for any positive initial biomass, the resulting biomass evolution y(·) is bounded and positively valued and y(t) → y* as t → ∞.

If, for positive biomasses, we introduce a new state variable x, defined by

$$x(t) = \ell n(y(t)/y^*), \qquad (5.14)$$

then

$$y(t) = y^* e^{x(t)} \qquad (5.15)$$

and, utilizing (5.13) - (5.15), the system can be described by

$$\dot{x}(t) = (r/k)y^*[1 - e^{x(t)}] + (e^{-x(t)}/y^*)[u(t) + Dy^*e^{x(t)}], \tag{5.16}$$

$$D = r(1 - y^*/k).$$

The control problem is now equivalent to that of obtaining a controller which assures that all solutions of (5.16) are bounded and converge to zero.

Equation (5.16) has the form of (2.21) with

$$f(t,x) = (r/k)y^*[1 - e^x], \quad B^{(2)}(t,x) = e^{-x}/y^*, \quad F = 1,$$

$$h(t,x) = y^*e^x, \quad u^{(2)}(t) = u(t). \tag{5.17}$$

Assumptions A1(1) and A3(2) are clearly satisfied. If one considers the function V given by

$$V(t,x) = y^*(e^x - x - 1) \tag{5.18}$$

then A1(2) can be shown to hold; see Ref. (9). From (2.23), (5.17), and (5.18) one has

$$\alpha^{(2)}(t,x) = 1 - e^{-x}, \tag{5.19}$$

and A5(2) and C3 are satisfied.

Thus, system description (5.16) belongs to S2, and a class of its C2 controllers are given by (utilizing (3.7) - (3.10), (5.15), (5.17) and (5.19))

$$\begin{aligned} u(t) &= \hat{K}(t)y(t) - \bar{\ell}_2\Gamma y(t)\eta(t), \\ \dot{\hat{K}}(t) &= -\Gamma\eta(t), \\ \eta(t) &= y(t) - y^*, \\ \Gamma &> 0, \ \bar{\ell}_2 \geq 0. \end{aligned} \tag{5.20}$$

As a consequence of Theorem 4.1, the use of a control given by (5.20) assures the desired system behavior.

For numerical simulations of this system we have taken $k = 1.5$, $r = 0.25$ and $y^* = 0.75$. We have considered the performance of the following three controllers for $y(0) = 2.039$.

1) The controller depending on knowledge of system parameters which assures the desired performance, i.e.,

$$u(t) = Ky(t), \quad K = -D = -0.125.$$

2) The controller given by (5.20) with $\Gamma = 0.1$, $\hat{K}(0) = 0$ and $\bar{\ell}_2 = 0$.

3) The controller given by (5.20) with $\Gamma = 0.1$, $\hat{K}(0) = 0$ and $\bar{\ell}_2 = 1$.

The results of the simulations are presented graphically in Figures 5.2(1) - 5.2(3).

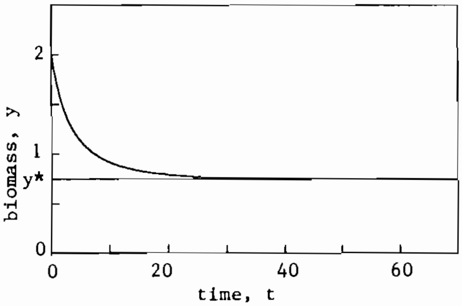


Fig. 5.2(1). Biomass evolution for $u(t) = Ky(t)$.

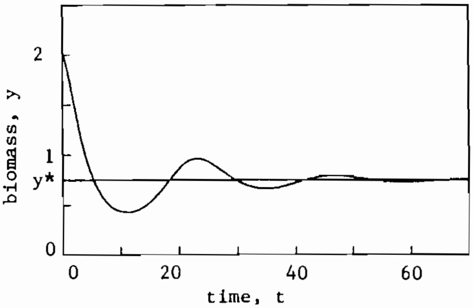


Fig. 5.2(2). Biomass evolution for $\Gamma=0.1$, $\bar{\ell}_2=0$.

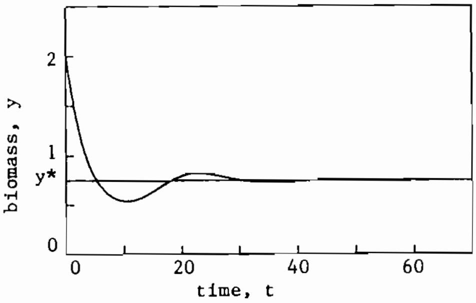


Fig. 5.2(3). Biomass evolution for $\Gamma=0.1$, $\bar{\ell}_2=1$.

C. SIMPLE PENDULUM

Consider a simple pendulum of mass m and length ℓ subjected to a control moment M and an *unknown* bounded disturbance $v(\cdot)$, in the form of a horizontal acceleration of its point of support; see Figure 5.3. Letting $u = M/m\ell^2$ and letting x_1 denote the angle between the arm of the pendulum and a vertical reference line, the equations of motion are

$$\begin{aligned}\dot{x}_1(t) &= x_2(t)\\ \dot{x}_2(t) &= -a \sin x_1(t) + u(t) - (v(t)/\ell) \cos x_1(t).\end{aligned} \tag{5.21}$$

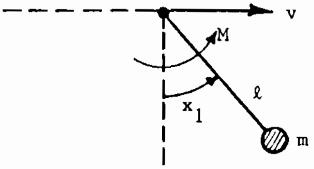


Fig. 5.3. A disturbed simple pendulum.

Assume that a is a known positive constant and the state $(x_1(t)x_2(t))$ is accessible. No information is assumed about the bound on $v(\cdot)$. We shall consider the problem of obtaining a controller which assures that all possible state trajectories of the system are bounded and converge to the zero state.

Since the zero state of the uncontrolled disturbance-free system is not g.u.a.s., the following control is proposed.

$$\begin{aligned} &u(t) = -bx_1(t) - cx_2(t) + u^{(3)}(t), \\ &c > 0, \quad b > -ad, \\ &d = \inf\{\sin x_1/x_1 : x_1 \in \mathbb{R},\ x_1 \neq 0\}, \end{aligned} \tag{5.22}$$

where $u^{(3)}(t)$ will be specified later. The system can now be described by

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -a \sin x_1(t) - bx_1(t) - cx_2(t) - (v(t)/\ell)\cos x_1(t) + u^{(3)}(t) \end{aligned} \tag{5.23}$$

which is in the form of (2.25) with

$$\begin{aligned} &f(t,x) = \begin{bmatrix} x_2 \\ -a \sin x_1 - bx_1 - cx_2 \end{bmatrix}, \\ &B^{(3)}(t,x) = (0,1)^T, \quad g(t,x,u) = u - (v(t)/\ell) \cos x_1. \end{aligned} \tag{5.24}$$

Assumption A1(1) is satisfied. Considering the function V defined by

$$V(t,x) = (b + c^2/2)x_1^2 + cx_1x_2 + x_2^2 + 2a(1 - \cos x_1), \tag{5.25}$$

A1(2) can be shown to hold; see Ref. (9). Letting $\beta = \sup\{|v(t)|/\ell : t \in \mathbb{R}\}$, $|(v(t)/\ell) \cos x_1| \leq \beta|\cos x_1|$; hence A7 and A3(3) are satisfied (see Remarks 2.3(1) and (2)) by taking

$$\beta_o = 1, \quad \rho(t,x) = \Pi(t,x,\beta) = \beta|\cos x_1|. \tag{5.26}$$

Thus, system description (5.23) belongs to S3.

Letting $u^{(3)}(t)$ be given by a C3 controller, one has

$$u^{(3)}(t) = -\tilde{s}(t)\hat{\beta}(t)|\cos x_1(t)|, \tag{5.27}$$

$$\dot{\hat{\beta}}(t) = \ell_3|\tilde{\alpha}(t)\cos x_1(t)|, \quad \hat{\beta}(t_o) > 0, \tag{5.28}$$

where

$$\begin{aligned} &\ell_3 > 0, \\ &\tilde{\alpha}(t) = cx_1(t) + 2x_2(t), \\ &\tilde{s}(t) = s(t,x(t),\hat{\beta}(t),\varepsilon(t)), \end{aligned} \tag{5.29}$$

s being any strongly Caratheodory function which assures the satisfaction of

$$\tilde{s}(t)|\tilde{\alpha}(t)| = |\tilde{s}(t)|\tilde{\alpha}(t),$$
$$|\tilde{\alpha}(t)\hat{\beta}(t)\cos x_1(t)| > \varepsilon(t) \Rightarrow \tilde{s}(t) = \frac{\tilde{\alpha}(t)}{|\tilde{\alpha}(t)|}, \qquad (5.30)$$

and

$$\dot{\varepsilon}(t) = -\ell_4\varepsilon(t), \quad \varepsilon(t_o) > 0, \quad \ell_4 > 0. \qquad (5.31)$$

If one lets

$$\tilde{s}(t) = \mathrm{sat}[\tilde{\alpha}(t)\hat{\beta}(t)|\cos x_1(t)|/\varepsilon(t)]$$

then

$$u^{(3)}(t) = -\mathrm{sat}(\hat{\eta}(t))\hat{\beta}(t)\cos x_1(t),$$
$$\hat{\eta}(t) = \tilde{\alpha}(t)\hat{\beta}(t)[\cos x_1(t)]/\varepsilon(t). \qquad (5.32)$$

For numerical simulations of this system we have taken $a = 1$, $\ell = 1$ and $v(t) = \cos t$. We have considered the system behavior under control given by (5.22) with $b = 1$, $c = 1$ and $u^{(3)}$ given by four different controllers.

1) $u^{(3)}(t) = 0$.

2) The controller which cancels the disturbance term in the system description, i.e.,

$$u^{(3)}(t) = (v(t)/\ell) \cos x_1(t).$$

This, of course, requires complete knowledge of the disturbance.

3) A non-adaptive controller which requires knowledge of the bound β,

$$u^{(3)}(t) = -\mathrm{sat}(\eta(t))\beta \cos x_1(t),$$
$$\eta(t) = \tilde{\alpha}(t)\beta[\cos x_1(t)]/\varepsilon,$$
$$\varepsilon = 0.01, \quad \beta = 1;$$

see Ref. (39).

4) The adaptive controller given by (5.32), (5.31), and (5.28) with $\ell_3 = 1$, $\ell_4 = 0.1$, $\hat{\beta}(0) = 0.01$ and $\varepsilon(0) = 0.01$.

The results of simulations for $x_1(0) = 1$ and $x_2(0) = 0$ are presented graphically in Figures 5.4(1) - 5.4(4).

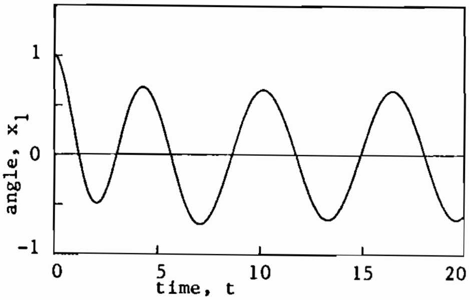


Fig. 5.4(1). Pendulum position for linear controller only.

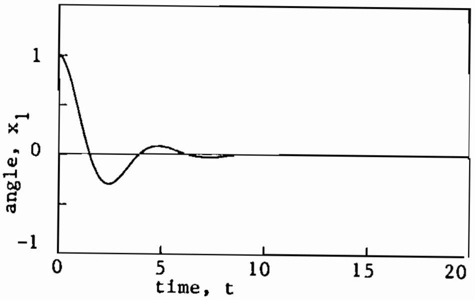


Fig. 5.4(2). Pendulum position for disturbance-cancelling controller.

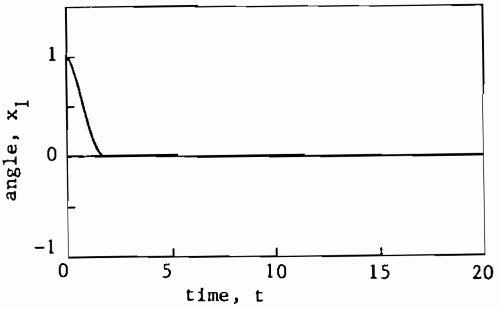


Fig. 5.4(3). Pendulum position for non-adaptive saturation controller.

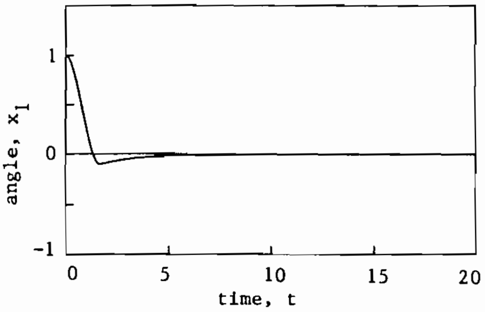


Fig. 5.4(4). Pendulum position for adaptive controller.

VI. APPENDIX

Definition 6.1. *A function* $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ *belongs to* class K (KR) *iff it is continuous, nondecreasing, and satisfies*

$$\gamma(0) = 0, \quad r > 0 \Rightarrow \gamma(r) > 0,$$

$$(\lim_{r\to\infty} \gamma(r) = \infty).$$

ACKNOWLEDGMENTS

The contents of this paper are based on research supported by the National Science Foundation under grant ECS-7813931 and carried out, in part, while George Leitmann was the recipient of an Alexander von Humboldt Foundation US Senior Scientist Award.

REFERENCES

1. Kushner, H.J., *IEEE Int. Conv. Rec. 14,* 143 (1966).
2. Astrom, K.J., "Introduction to Stochastic Control Theory." Academic Press, New York (1970).
3. Lur'e, A.I., "Some Nonlinear Problems in the Theory of Automatic Control" (in Russian). Gostekhizdat, Moscow (1951). German translation, Akademie Verlag (1957). English translation, Her Majesty's Stationary Office (1957).
4. LaSalle, J.P., *SIAM J. Contr. 1,* 3 (1962).
5. Kalman, R.E., and Bertram, J.E., *J. Basic Engineering 82,* 371 (1960).
6. Rang, E.R., "Adaptive Controllers Derived by Stability Considerations," Memorandum MR7905. Minneapolis-Honeywell Regulator Co. (1962).
7. Butchart, R.L., and Shakcloth, B., Proc. 2nd IFAC Symp. on Theory of Self-Adaptive Control Systems, 145. Plenum Press (1966).
8. Parks, P.C., *IEEE Trans. Automat. Contr. AC-11,* 362 (1966).
9. Corless, M., and Leitmann, G., in "Dynamical Systems and Microphysics," Blaquière, A., and Leitmann, G. (eds.). Academic Press, New York (1984).
10. LaSalle, J.P., and Lefchetz, S., "Stability by Liapunov's Direct Method with Applications." Academic Press, New York (1961).
11. Hahn, W., "Stability of Motion." Springer-Verlag, Berlin (1967).
12. Cesari, L., "Asymptotic Behavior and Stability Problems in Ordinary Differential Equations." Springer-Verlag, New York (1971).
13. Molander, P., "Stabilisation of Uncertain Systems," LUTFD2/(TFRT-1020)/1-111. Lund Institute of Technology (1979).

14. Thorp, J.S., and Barmish, B.R., *J. Optimiz. Theory Appl. 35,* 559 (1981).
15. Barmish, B.R., Corless, M., and Leitmann, G., *SIAM J. Contr. and Optimiz. 21,* 246 (1983).
16. Johnson, C.D., *IEEE Trans. Automat. Contr. AC-13,* 416 (1968).
17. Sobral, M., Jr., and Stefanek, R.G., *IEEE Trans. Automat. Contr. AC-15,* 498 (1970).
18. Bélanger, P.R., *IEEE Trans. Automat. Contr. AC-15,* 695 (1970).
19. Davison, E.J., and Smith, H.W., *Automatica 7,* 489 (1971).
20. Davison, E.J., and Goldenberg, A., *Automatica 11,* 461 (1975).
21. Desoer, C.A., and Wang, Y.T., *IEEE Trans. Automat. Contr. AC-23,* 70 (1978).
22. Grujić, Lj. T., and Porter, B., *Int. J. Syst. Sci. 11,* 177 (1980).
23. Porter, B., and Grujić, Lj. T., *Int. J. Syst. Sci. 11,* 837 (1980).
24. Winsor, C.A., and Roy, R.J., *IEEE Trans. Automat. Contr. AC-13,* 204 (1968).
25. Lindorff, D.P., and Carroll, R.L., *Int. J. Contr. 18,* 897 (1973).
26. Landau, I.D., *Automatica 10,* 353 (1974).
27. Monopoli, R.V., *IEEE Trans. Automat. Contr. AC-19,* 474 (1974).
28. Narendra, K.S., and Valavani, L.S., *IEEE Trans. Automat. Contr. AC-23,* 570 (1978).
29. Narendra, K.S., and Valavani, L.S., *IEEE Trans. Automat. Contr. AC-25,* 243 (1980).
30. Narendra, K.S., Lin, Y.H., and Valavani, L.S., *IEEE Trans. Automat. Contr. AC-25,* 440 (1980).
31. Morse, A.S., *IEEE Trans. Automat. Contr. AC-25,* 433 (1980).
32. Parks, P.C., *IEEE Proc. Pt. D 128,* 195 (1981).
33. Monopoli, R.V., "Engineering Aspects of Control System Design Via the 'Second Method' of Lyapunov," CR-654. NASA (1966).
34. Gutman, S., and Leitmann, G., Proc. 2nd IFIP Conf. Optimiz. Techniques. Springer-Verlag, Berlin (1976).
35. Gutman, S., and Leitmann, G., Proc. IEEE Conf. Decision Contr. (1976).
36. Gutman, S., *IEEE Trans. Automat. Contr. AC-24,* 437 (1979).
37. Leitmann, G., *J. Dynam. Syst. Meas. Contr. 101,* 212 (1979).
38. Leitmann, G., *Acta Astronautica 7,* 1457 (1980).
39. Corless, M., and Leitmann, G., *IEEE Trans. Automat. Contr. AC-26,* 1139 (1981).
40. Leitmann, G., *J. Dynam. Syst. Meas. Contr. 103,* 95 (1981).
41. Ryan, E.P., *Int. J. Contr. 38* (1983).
42. Petersen, I.R., and Barmish, B.R., Proc. Amer. Contr. Conf. (1984).
43. Ryan, E.P., and Corless, M., *IMA J. Math. Contr. Inf. 1,* 223 (1984).
44. Lee, C.S., and Leitmann, G., 2nd Workshop on Renewable Resource Management (1985).
45. Singh, S.N., *IEEE Trans. Automat. Contr. AC-30,* 1099 (1985).
46. Singh, S.N., Proc. IEEE Conf. Decision Contr. (1986).
47. Corless, M., and Leitmann, G., *Annales de la Fondation Louis de Broglie 9,* 65 (1984).
48. Narendra, K.S., and Taylor, J.H., "Frequency Domain Criteria for Absolute Stability." Academic Press, New York (1973).
49. Anderson, B.D.O., and Vongpanitlerd, S., "Network Synthesis: A State Space Approach." Prentice-Hall (1973).
50. Lee, C.S., and Leitmann, G., *Int. J. Syst. Sci. 14,* 979 (1983).

DETERMINISTIC CONTROL OF UNCERTAIN SYSTEMS

M. Corless
School of Aeronautics and Astronautics
Purdue University
West Lafayette, Indiana 47907 USA

G. Leitmann
Department of Mechanial Engineering
University of California
Berkeley, California 94720 USA

FOREWORD

In order to control the behavior of a system in the "real" world, the system analyst seeks to capture the system's salient features in a mathematical model. This abstraction of the "real" system usually contains uncertain elements, for example, uncertainties due to parameters, constant or varying, which are unknown or imperfectly known, or uncertainties due to unknown or imperfectly known inputs into the system. Despite such imperfect knowledge about the chosen mathematical model, one often seeks to devise controllers which will "steer" the system in some desired fashion, for example, so that the system response will approach or track a desired reference response; by suitable definition of the system (state) variables such a problem can usually be cast into that of stabilizing a prescribed state.

Two main avenues are open to the analyst seeking to control an uncertain dynamical system. He may choose a stochastic approach in which information about the uncertain elements as well as about the system response is statistical in nature; for example, see Refs. (1,2). Loosely speaking, when modelling via random variables, one is content with desirable behavior on the average. The other approach to the control of uncertain systems, and the one for which we shall opt in the present discussion, is deterministic. Available, or assumed, information about uncertain elements is deterministic in nature. here one seeks controllers which assure the desired response of the dynamical system.

I. INTRODUCTION

We consisder the problem of obtaining memoryless stabilizing feedback controllers for uncertain dynamical systems described by ordinary differential equations. Various classes of controllers are presented. The design of all of these controllers is based on Lyapunov theory.

Before proceeding with the problem, we introduce some basic notions and results for ordinary differential equations.

II. BASIC NOTIONS

Let $T = (\underline{t},\infty)$ where $\underline{t} \in [-\infty,\infty)$; let X be a non-empty open subset of $\mathbb{R}^n$; and let $f: T \times X \to \mathbb{R}^n$. Consider the first order *ordinary differential equation (o.d.e.)*

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

where $\dot{x}(t)$ denotes the derivative of the function $x(\cdot)$ at t. By a *solution* of (2.1) we shall mean an absolutely continuous function $x(\cdot):[t_0,t_1) \to X$, where $t_0 \in T$ and $t_1 \in (t_0,\infty]$, which satisfies (2.1) almost everywhere[1] on $[t_0,t_1)$.

When considering a system described by an equation of the form (2.1), we shall refer to X as the *state space*, a member of X as a *state*, equation (2.1) as the *state equation*, and a solution of (2.1) as a *state evolution*, *state motion*, or *state history*.

A. EXISTENCE AND CONTINUATION OF SOLUTIONS

Since, in this paper, we consider systems described by o.d.e.'s, the two properties introduced in this section are of fundamental importance.

Definition 2.1. Equation (2.1) has (global) existence of solutions *iff, given any pair* $(t_0,x^o) \in T \times X$, *there exists a solution* $x(\cdot):[t_0,t_1) \to X$ *of (2.1) with* $x(t_0) = x^o$.

The following theorem (see Ref. (3) or (4) for a proof) yields sufficient conditions for existence of solutions.

Theorem 2.1. If f is a Carathêodory[2] *function, equation (2.1) has global existence of solutions.*

Definition 2.2 Equation (2.1) has indefinite continuation of solutions *iff, given any solution* $x(\cdot):[t_0,t_1) \to X$ *of (2.1), there exists a solution* $x^c(\cdot):[t_0,\infty) \to X$ *of (2.1) with* $x^c(t) = x(t)$ *for all* $t \in [t_0,t_1)$.

The following theorem, which may be deduced from the results presented in Ref. (4), chapter 1, provides useful sufficient conditions for indefinite continuation of solutions.

Theorem 2.2 Suppose f is Caratheodory and for each solution $x(\cdot):[t_0,t_1) \to X$ *of (2.1) with* $t_1 < \infty$, *there exists a compact subset C of X such that* $x(t) \in C$ *for all* $t \in [t_0,t_1)$. *Then, equation (2.1) has indefinite continuation of solutions.*

[1] That is, everywhere except possibly on a set of Lebesgue measure zero.

[2] See Appendix, sec. A, or just note that if f is continuous, it is Caratheodory.

B. BOUNDEDNESS AND STABILITY

In this section, we formalize the notion of a system described by (2.1) exhibiting "desirable" behavior with respect to a state $x^\star \in \bar{X}$, where $\bar{X}$ is the closure of $\bar{X}$.

Definition 2.3. The solutions of (2.1) are globally uniformly bounded (g.u.b.) *iff, given any compact subset* C *of* X*, there exists* $d(C) \in \mathbb{R}_+$ *such that, if* $x(\cdot):[t_0,t_1) \to X$ *is any solution of (2.1) with* $x(t_0) \in C$*, then* $\|x(t)\| \leq d(C)$ *for all* $t \in [t_0,t_1)$.

Definition 2.4 $x^\star$ *is* uniformly stable (u.s.) *for (2.1) or (2.1) is* uniformly stable *about* $x^\star$ *iff, given any neighborhood*[3] B *of* $x^\star$*, there exists a neighborhood* B_0 *of* $x^\star$ *such that, if* $x(\cdot):[t_0,t_1) \to X$ *is any solutions of (2.1) with* $x(t_0) \in B_0$*, then* $x(t) \in B$ *for all* $t \in [t_0,t_1)$.

Definition 2.5. $x^\star$ *is a* global uniform attractor (g.u.a.) *for (2.1) iff, given any neighborhood* B *of* $x^\star$ *and any compact subset* C *of* X*, there exists* $T(C,B) \in \mathbb{R}_+$ *such that, if* $x(\cdot):[t_0,\infty) \to X$ *is any solution of (2.1) with* $x(t_0) \in C$*, then* $x(t) \in B$ *for all* $t \geq t_0 + T(C,B)$.

Definition 2.6. $x^\star$ *is* globally uniformly asymptotically stable (g.u.a.s) *for (2.1) or (2.1) is* globally uniformly asymptotically stable *about* $x^\star$ *iff:*

1) The solutions of (2.1) are g.u.b.

2) $x^\star$ *is u.s. for (2.1).*

3) $x^\star$ *is a g.u.a. for (2.1).*

Remark 2.1. Frequently, in the definition of uniform stability of $x^\star$ in the literature, $x^\star$ is assumed to be an *equilibrium state* for (2.1), i.e., $x^\star \in X$ and $f(t,x^\star) = 0$[4] for all $t \in T$, or, equivalently, the function $x(\cdot):T \to X$, $x(t) = x^\star$, is a solution of (2.1). However, one may readily show that, if a state $x^\star \in X$ is uniformly stable and if $x(\cdot):[t_0,t_1) \to X$ is any solution of (2.1) with $x(t_0) = x^\star$, it necessarily follows that $x(t) = x^\star$ for all $t \in [t_0,t_1)$.

C. LYAPUNOV FUNCTIONS AND A SUFFICIENT CONDITION FOR G.U.A.S.

In this section, we restrict the discussion to differential equations of the form

[3] By a neighborhood of $x^\star$, we mean a set containing on open set which contains $x^\star$.

[4] We use "0" to denote a zero vector.

(2.1) with $X = \mathbb{R}^n$, i.e.,

$$\dot{x}(t) = f(t,x(t)) \tag{2.2}$$

where $f: T \times X \to \mathbb{R}^n$, $X = \mathbb{R}^n$, and $T = (\underline{t},\infty)$ with $\underline{t} \in [-\infty,\infty)$. In particular, we present a theorem (Theorem 2.3) which yields a sufficient condition assuring that (2.2) is g.u.a.s. about the zero state. The condition utilizes the notion of a Lyapunov function which we shall define presently.

Definition 2.7. A function $V: T \times \mathbb{R}^n \to \mathbb{R}_+$ is a candidate Lyapunov function *iff it is continuously differentiable and there exist functions $\gamma_1, \gamma_2: \mathbb{R}_+ \to \mathbb{R}_+$ of class* KR[5] *such that for all $(t,x) \in T \times \mathbb{R}^n$*

$$\gamma_1(\|x\|) \leq V(t,x) \leq \gamma_2(\|x\|) \ . \tag{2.3}$$

Remark 2.2 Suppose $V: T \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$V(t,x) = W(x)$$

for all $(t,x) \in T \times \mathbb{R}^n$, where $W: \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function satisfying

$$W(0) = 0,$$
$$x \neq 0 \Rightarrow W(x) > 0,$$
$$\lim_{\|x\| \to \infty} W(x) = \infty \ ,$$

for all $x \in \mathbb{R}^n$. Then, V is a candidate Lyapunov function. To see this, define $\gamma_1, \gamma_2: \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\gamma_1(r) = \inf_{\|x\| \geq r} W(x) \ ,$$

$$\gamma_2(r) = \sup_{\|x\| \leq r} W(x),$$

for all $r \in \mathbb{R}_+$.

Definition 2.8. A function $V: T \times \mathbb{R}^n \to \mathbb{R}_+$ is a Lyapunov function *for (2.2) iff it is a candidate Lyapunov function and there exists a function $\gamma_3: \mathbb{R}_+ \to \mathbb{R}_+$ of class* K[6] *such that for all $(t,x) \in T \times \mathbb{R}^n$*

[5] See Appendix, sec. B.

[6] See Appendix, sec. B.

$$\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)f(t,x) < -\gamma_3(\|x\|). \tag{2.4}$$

We may now introduce a sufficient condition for (2.2) to be g.u.a.s. about zero.

Theorem 2.3. If there exists a Lyapunov function for (2.2) then (2.2) is g.u.a.s. about zero.

Proofs of various versions of Theorem 2.3 can be found in the literature; see, for example, Refs. (6-10). Also, Theorem 2.3 is a corollary of Theorem 6.1 of which there is a proof in Ref. (69).

The following corollary is readily deduced from Theorem 2.1, 2.2, and 2.3.

Corollary 2.1. If $f{:}T \times \mathbb{R}^n \to \mathbb{R}^n$ is Carathéodory and there exists a Lyapunov function for (2.2), then, (2.2) has existence and indefinite continuation of solutions and is g.u.a.s. about zero.

D. SYSTEMS WITH CONTROL

In this section, we present a notion which is basic in this paper. It is the notion of a system with control, i.e., a system whose state evolution depends not only on an initial state but also on an externally applied control input. For some non-empty set $U \subset \mathbb{R}^m$, the set of control values, and some function $F{:}T \times X \times U \to \mathbb{R}^n$ (T and X are as before) such a system is described by

$$\dot{x}(t) = F(t,x(t),u(t)) \tag{2.5}$$

where $u(t) \in U$ is the control value at t.

Thus, if, for any function $c{:}T \to U$, one lets

$$u(t) = c(t) \tag{2.6}$$

in (2.5), the resulting *open-loop* controlled system is described by

$$\dot{x}(t) = F(t,x(t),c(t)), \tag{2.7}$$

i.e., it is described by (2.1) with

$$f(t,x) = F(t,x,c(t)); \tag{2.8}$$

hence the state evolution depends on the choice of c. Such a function c is often called on open-loop control function.

In this paper we shall consider control to be given by a memoryless *feedback* controller, i.e., we shall consider

$$u(t) = p(t,x,(t)) \tag{2.9}$$

for some feedback control function $p:\mathcal{T} \times \mathcal{X} \to \mathcal{U}$. Substituting (2.9) into (2.5), a feedback controlled system can be described by

$$\dot{x}(t) = F(t,x(t),\ p(t,x(t))), \tag{2.10}$$

i.e., it can be described by (2.1) with

$$f(t,x) = F(t,x,p(t,x)). \tag{2.11}$$

Since we shall be considering the g.u.a.s. property as a criterion for desirable system behavior we introduce the following definitions.

Definition 2.9. A feedback control function $p:\mathcal{T} \times X \to \mathcal{U}$ stabilizes *(2.5) about* x^* *iff (2.10) has existence and indefinite continuation of solutions and is g.u.a.s. about* x^*.

Definition 2.10. (2.5) is stabilizable *about* x^* *iff there exists a* $p:\mathcal{T} \times X \to \mathcal{U}$ *which stabilizes (2.5) about* x^*.

III. UNCERTAIN SYSTEMS

In the previous chapter, we introduced the notion of a system with control, i.e., a system described by

$$\dot{x}(t) = F(t,x(t),\ u(t)) \tag{3.1}$$

for some function $F:\mathcal{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}^n$. Clearly, such a control system is completely specified by specifying F.

When modelling a "real" system, on usually does not have, or cannot obtain, an "exact" model. The model usually contains uncertain elements, for example, uncertainties due to parameters, constant or varying, which are unknown or imperfectly known, or uncertainties due to unknown or imperfectly known inputs into the system.

When saying that a system described by (3.1) is uncertain, we are really saying that F is uncertain, i.e., we do not know exactly what the function F is. Therefore, in our model of an uncertain system, we model the uncertainty with a statement of the form

$$F \in \mathcal{F} \tag{3.2}$$

where $\mathcal{F}$ is some *known*, non-empty, class of functions which map $\mathcal{T} \times \mathcal{X} \times \mathcal{U}$ into $\mathbb{R}^n$.

$\mathcal{F}$ reflects our knowledge of the system.

As a simple first example of an uncertain system, consider a scalar system subject to an uncertain Lebesgue measurable input disturbance $v:\mathbb{R} \to \mathbb{R}$

$$\dot{x}(t) = -x(t) + v(t) + u(t). \tag{3.3}$$

Table 3.1 lists three different possible assumptions on the knowledge of v.

In case 1, the disturbance is simply an unknown constant and the system model is given by (3.1)-(3.2) where a member of $\mathcal{F}$ is any function $F:\mathbb{R}^3 \to \mathbb{R}$ which satisfies

$$F(t,x,u) = -x+d+u \quad \forall (t,x,u) \in \mathbb{R}^3 \tag{3.4}$$

for some $d \in \mathbb{R}$.

In case 2, the disturbance is an unknown Lebesgue measurable function with known upper and lower bounds, $\overline{\rho}$ and $\underline{\rho}$, respectively. In this case, a member of $\mathcal{F}$ in system description (3.1)-(3.2) is any function $F:\mathbb{R}^3 \to \mathbb{R}$ which satisfies

$$F(t,x,u) = -x+v(t)+u \quad \forall (t,x,u) \in \mathbb{R}^3 \tag{3.5}$$

for some measurable function $v:\mathbb{R} \to [\underline{\rho},\overline{\rho}]$.

In case 3, which includes cases 1 and 2, the disturbance is modelled by a bounded measurable function with no assumption on the knowledge of its bounds. In this case, a member of $\mathcal{F}$ in (3.1)-(3.2) is any function $F:\mathbb{R}^3 \to \mathbb{R}$ which satisfies (3.5) for some bounded measurable v. This case is treated in Refs. (69) and (70).

1. $v(t) = d \;\forall\; t \in \mathbb{R}$; $d \in \mathbb{R}$ unknown

2. $v:\mathbb{R} \to [\underline{\rho},\overline{\rho}]$; $\underline{\rho}$, $\overline{\rho}$ known

3. $v:\mathbb{R} \to \mathbb{R}$, bounded

TABLE 3.1

As a second example, consider a scalar system

$$\dot{x}(t) = v(t)x(t)+u(t) \tag{3.6}$$

where the Lebesgue measurable function $v:\mathbb{R} \to \mathbb{R}$ models an uncertain parameter. Again, Table 3.1 lists some possible assumptions on the knowledge of v.

As a generalization of the previous two examples, consider a system described by

$$\dot{x}(t) = g(t,x(t),\ u(t),\ v(t)) \tag{3.7}$$

where $g:\mathcal{T} \times \mathcal{X} \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}^n$ is known Caratheodory function, $v:\mathcal{T} \to \mathcal{V}$ is an uncertain Lebesgue measurable function, and $\mathcal{V} \subset \mathbb{R}^p$. This can be considered the general model for a system with uncertain parameters or inputs, the uncertainties being modelled by v. Table 3.2 lists three different possible assumptions on the knowledge of v.

1. $v(t) = d \ \forall \ t \in \mathbb{R}$; $d \in \mathcal{V}$, d unknown
2. $v:\mathcal{T} \to \mathcal{V}^o$, $\mathcal{V}^o$ known
3. $v:\mathcal{T} \to \mathcal{V}$, v bounded

TABLE 3.2

An uncertain system described by case 2 of the previous example is an example of the type of uncertain system we shall be considering in this article. Basically, the uncertain systems considered here are specified by specifying for each $(t,x,u) \in \mathcal{T} \times \mathcal{X} \times \mathcal{U}$, the set of possible values which F(t,x,u) may assume. For cases 1 and 3, see Refs. (69) and (70).

IV. INITIAL PROBLEM STATEMENT - STABILIZATION

A. *PROBLEM STATEMENT*

Basically, the type of problem we shall consider initially in this paper is as follows.

Given an uncertain system described by

$$\dot{x}(t) = F(t,x(t), u(t)), \tag{4.1}$$

$$F \in \mathcal{F}, \tag{4.2}$$

where $\mathcal{F}$ is a known, non-empty class of function which map $\mathcal{T} \times \mathcal{X} \times \mathcal{U}$ into $\mathbb{R}^n$, and a "desirable" state $x^* \in \mathcal{X}$, obtain a feedback control function $p:\mathcal{T} \times \mathcal{X} \to \mathcal{U}$ which stabilizes (4.1) about x^*.

Since the only information available on F is a class of functions $\mathcal{F}$ to which F belongs, we attempt therefore to solve the above problem by looking for a feedback control function which stabilizes (4.1) about x^* for all $F \in \mathcal{F}$.

We now introduce:

Definition 4.1. A feedback control function $p:\mathcal{T} \times \mathcal{X} \to \mathcal{U}$ stabilizes (4.1)-(4.2) *about* x^* *iff* p *stabilizes (4.1) about* x^* *for each* $F \in \mathcal{F}$.

The problem we shall consider is that of obtaining a feedback control function which stabilizes (4.1)-(4.2) for a given $\mathcal{F}$.

Definition 4.2 (4.1)-(4.2) is stabilizable *about* x^* *iff there exists* $p:\mathcal{T} \times \mathcal{X} \to \mathcal{U}$ *which stabilizes (4.1)-(4.2) about* x^*.

Remark 4.1. Note that stabilizability of (4.1) for each $F \in \mathcal{F}$ does not imply stabilizability of (4.1)-(4.2). It might be the case that, for each $F \in \mathcal{F}$, there exists p (dependent on F) which stabilizes (4.1), but there does not exist p which stabilizes (4.1) for all $F \in \mathcal{F}$.

For example, consider the pair of scalar systems

$$\begin{aligned} \dot{x}(t) &= u(t), \\ \dot{x}(t) &= -u(t). \end{aligned} \tag{4.3}$$

Although each system is stabilizable (e.g., let $p(t,x) = -x$ and $p(t,x) = x$, respectively), it is unlikely that there exists a feedback control function which stabilizes both of them. However, there may exist a non-memoryless or dynamic controller which stabilizes both; see Refs. (69) and (70).

B. A USEFUL THEOREM FOR THE SYNTHESIS OF STABILIZING CONTROLLERS

In this section, we present a theorem (Theorem 4.1) which is useful in the synthesis of zero-state stabilizing feedback control functions for uncertain systems whose state space is $\mathbb{R}^n$. For a given uncertain system, the theorem yields criteria which, if satisfied by a feedback control function, ensure that the feedback control function is a stabilizing controller.

Theorem 4.1. Consider an uncertain system described by (4.1)-(4.2) with $X = \mathbb{R}^n$ *and suppose that* $p: T \times X \to U$ *is such that*

$$\dot{x}(t) = F(t,x(t),\ p(t,x(t))) \tag{4.4}$$

has existence and indefinite continuation of solutions for all $F \in \mathcal{F}$*. If, for each* $F \in \mathcal{F}$*, there exists a Lyapunov function for (4.4), then* p *stabilizes (4.1)-(4.2) about zero.*

Proof. The proof follows readily from Theorem 2.3.

In the next section, we consider a particular class of uncertain systems. For each member of that class, we present a class of candidate stabilizing feedback control functions whose design is based on meeting the Lyapunov criterion in Theorem 4.1.

V. *L-G* CONTROLLERS

In this chapter, we consider first a class of uncertain systems which have been treated previously in the literature; see Refs. (33,36,42,60). For each member of this class, we present a class of (previously obtained) candidate stabilizing controllers. We then enlarge the class of systems for which the presented controllers are candidate stabilizing controllers. Finally, we present a theorem which yields some properties of systems subject to the controllers presented.

A. ORIGINAL CLASS OF UNCERTAIN SYSTEMS

A member of the class of uncertain systems under consideration here is described by (3.1)-(3.2), i.e.,

$$\dot{x}(t) = F(t,x(t),\ u(t)) \tag{5.1}$$

$$F \in \mathcal{F} \tag{5.2}$$

where $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, and $\mathcal{F}$ satisfies Assumptions A1 and A2.

Assumption A1. ($F(t,x,\cdot)$ is affine) *For each* $F \in \mathcal{F}$*, there exist functions* $f: T \times X \to \mathbb{R}^n$ *and* $B: T \times X \to \mathbb{R}^{n \times m}$ *such that for all* $(t,x,u) \in T \times X \times U$

$$F(t,xu) = f(t,x) + B(t,x)u. \tag{5.3}$$

If A1 is satisfied, then each $F \in F$ has a unique representation in the form of (5.3), i.e., for each $F \in F$, there exists a unique pair (f,B) for which (5.3) is satisfied for all $(t,x,u) \in T \times X \times U$; this pair is given by

$$f(t,x) = F(t,x,0), \tag{5.4}$$

$$B(t,x)u = F(t,x,u) - F(t,x,0) \quad \forall\, u \in U \tag{5.5}$$

for all $(t,x) \in T \times X$.

We let S_F denote the set of pairs (f,B) which, for some $F \in F$, satisfy (5.4)-(5.5) for all $(t,x) \in T \times X$.

Assumption A2. There exist Carathéodory functions $f^o: T \times X \to \mathbb{R}^n$ and $B^o: T \times X \to \mathbb{R}^{n\times m}$, a candidate Lyapunov function $V: T \times X \to \mathbb{R}_+$, a strongly Carathéodory[7] *function $\rho^o: T \times X \to \mathbb{R}_+$, and a constant $c \in \mathbb{R}_+$ such that:*

1) V is a Lyapunov function for

$$\dot{x}(t) = f^o(t,x(t)). \tag{5.6}$$

2) For each $(f,B) \in S_F$, there exist Carathéodory functions $e: T \times X \to \mathbb{R}^m$ and $E: T \times X \to \mathbb{R}^{m\times m}$ each that

$$f = f^o + B^o e, \tag{5.7}$$

$$B = B^o + B^o E, \tag{5.8}$$

and

$$\|e(t,x)\| \leq \rho^o(t,x), \tag{5.9}$$

$$\|E(t,x)\| \leq c < 1 \tag{5.10}$$

for all $(t,x) \in T \times X$.

Thus, utilizing (5.1) (5.3), (5.7) and (5.8), any system under consideration here is described by

$$\dot{x}(t) = f^o(t,x(t)) + B^o(t,x(t))[e(t,x(t)) + E(t,x(t))u(t) + u(t)], \tag{5.11}$$

[7] See Appendix, sec. A, or just note that if a function is continuous, it is strongly Caratheodory.

where f^0 has a Lyapunov function V, and e and E satisfy (5.9) and (5.10) for some strongly Carathéodory function ρ^0 and constant $c > 0$.

Remarks 5.1. 1) For a certain or completely known system, i.e., for a system for which e and E are known, it should be clear that, under condition (5.10), (5.11) is stabilizable. If one lets

$$
\begin{aligned}
u(t) &= p(t,x(t)),\\
p(t,x) &= -\,[I+E(t,x)]^{-1}e(t,x),
\end{aligned}
\tag{5.12}
$$

then (5.11) reduces to (5.6) which, as a consequence of f^0 being Carathéodory, part 1 of Assumption A2, and Corollary 2.1, has existence and indefinite continuation of solutions and is g.u.a.s. about zero.

2) In the literature, conditions (5.7) and (5.8) are sometimes referred to as *matching conditions;* see Refs. (36,42,46,59).

When Assumption A1 is satisfied, the existence of f^0, B^0 satisfying (5.7) and (5.8) are equivalent to either of the following two conditions:

Condition C1. There exist Carathéodory functions $f^0: T \times X \to \mathbb{R}^n$ and $B^0: T \times X \to \mathbb{R}^{n\times m}$ such that for all $F \in \mathcal{F}$ and $(t,x,u) \in T \times X \times U$,

$$
F(t,x,u) - f^0(t,x) \in \mathcal{R}\,(B^0(t,x)) \qquad \text{5.13)}
$$

where $\mathcal{R}(B^0(t,x))$ denotes the range space of $B^0(t,x)$.

Condition C2. There exist Carathéodory functions $f^0: T \times X \to \mathbb{R}^n$ and $B^0: T \times X \to \mathbb{R}^{n\times m}$ such that for all $F \in \mathcal{F}$ and $(t,x,u) \in T \times X \times U$,

$$
[I-B^0(t,x)B^{0\dagger}(t,x)]\;[F(t,x,u) - f^0(t,x)] = 0 \tag{5.14}
$$

where $B^{0\dagger}(t,x)$ denotes the pseudoinverse[8] of $B^0(t,x)$; see Refs. (13,14).

3. Suppose one has an uncertain system described by (5.1)-(5.2) which satisfies A1 and A2.2 but for which A2.1 is relaxed to:

Assumption A3. There exists a strongly Carathêodory function $p^0: T \times X \to U$ such that V is a Lyapunov function for

$$
\dot{x}(t) = \bar{f}^0(t,x(t)), \tag{5.15}
$$

where

[8] If $B \in \mathbb{R}^{n\times m}$ and rank $(B) = m$, then $B^\dagger = [B^TB]^{-1}B^T$.

$$\bar{f}^0 = f^0 + B^0 p^0 \ . \tag{5.16}$$

Assumption A3 assures that

$$\dot{x}(t) = f^0(t,x(t)) + B^0(t,x(t))u(t) \tag{5.17}$$

is stabilizable about zero.

Letting

$$\bar{u}(t) = u(t) - p^0(t,x(t)), \tag{5.18}$$

one has

$$u(t) = p^0(t,x(t)) + \bar{u}(t) \tag{5.19}$$

and, utilizing Assumptions A1 and A2.2, one has[9] for each $F \in \mathcal{F}$

$$\begin{aligned} \dot{x} &= F \\ &= f^0 + B^0[e+Eu+u] \\ &= f^0 + B^0[e+Ep^0 + p^0 + E\bar{u}+\bar{u}] \\ &= f^0 + B^0p^0 + B^0[e+Ep^0 + E\bar{u}+\bar{u}] \\ &= \bar{f}^0 + B^0[\bar{e}+E\bar{u}+\bar{u}] \end{aligned}$$

where

$$\bar{e} = e+Ep^0. \tag{5.20}$$

Thus,

$$\dot{x}(t) = \bar{F}(t,x(t), \bar{u}(t)) \tag{5.21}$$

where

$$\bar{F}(t,x,u) = \bar{f}^0(t,x) + B^0(t,x)[\bar{e}(t,x) + E(t,x)\bar{u}+\bar{u}] \ , \tag{5.22}$$

$$\|\bar{e}(t,x)\| \leq \bar{\rho}^0(t,x), \tag{5.23}$$

$$\bar{\rho}^0(t,x) = \rho^0(t,x) + c\|p^0(t,x)\| \tag{5.24}$$

for all $(t,x,u) \in \mathcal{T} \times \mathcal{X} \times \mathcal{U}$. Hence, one may obtain a new system description which

[9] Sometimes, for the sake of brevity, we shall omit arguments.

satisfied A1 and A2; $\bar{f}^0$, B^0, V, $\bar{\rho}^0$, and c assure A1 and A2.

B. L-G CONTROLLERS

Consider any uncertain system described by (5.1)-(5.2) (with $X = \mathbb{R}^n$) which satisfies Assumptions A1 and A2. An *L-G* controller for such a system is any function $p:T \times X \to U$ which satisfies

$$p(t,x) = -\frac{\alpha(t,x)}{\|\alpha(t,x)\|}\rho^c(t,x) \quad \text{if} \quad \alpha(t,x) \neq 0, \tag{5.25}$$

$$\rho^c(t,x) > \rho(t,x), \tag{5.26}$$

where

$$\alpha(t,x) = B^{0T}(t,x)\frac{\partial V^T}{\partial x}(t,x), \tag{5.27}$$

$$\rho(t,x) = \rho^0(t,x)/(1-c) \tag{5.28}$$

for all $(t,x) \in T \times X$, and (f^0,B^0,V,ρ^0,c) assure satisfaction of A2.

For previous literature on the above controllers, see Refs. (6,17-37).

C. EXTENSION OF ORIGINAL SYSTEM CLASS

In this section, we present a class of uncertain systems which is a generalization of the class presented in sec. V.A. An uncertain system in this class is described by (5.1)-(5.2) where $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, and F satisfies the following assumption.

Assumption A4. There exist a Carathêodory function $B^0:T \times X \to \mathbb{R}^{n\times m}$, a candidate Lyapunov function $V:T \times X \to \mathbb{R}_+$, and a strongly Carathêodory function $\rho:T \times X \to \mathbb{R}_+$ such that each $F \in \mathrm{F}$ can be expressed as

$$F(t,x,u) = f^s(t,x) + B^0(t,x)g(t,x,u) \tag{5.29}$$

for all $(t,x,u) \in T \times X \times U$, for some functions $f^s:T \times X \to \mathbb{R}^n$ and $g:T \times X \times U \to \mathbb{R}^m$ which satisfy:

1) f^s is Carathêodory and V is a Lyapunov function for

$$\dot{x}(t) = f^s(t,x(t)). \tag{5.30}$$

2) g is strongly Carathêodory and

$$\|u\| > \rho(t,x) \Rightarrow u^T g(t,x,u) > o \tag{5.31}$$

for all $(t,x,u) \in T \times X \times U$.

An *L-G* controller for a system in this class is any function $p:T \times X \to U$ which satisfies (5.25), (5.26), and (5.27), where (B^o,V,ρ) is any triple which assures A4.

Remarks 5.2. 1) To demonstrate that a member of the class of systems considered in sec. V.A is a member of the class treated here, let

$$f^s = f^o \tag{5.32}$$

$$g(t,x,u) = e(t,x) + E(t,x)u+u \tag{5.33}$$

for all $(t,x,u) \in T \times X \times U$; hence (5.29) is satisfied. Now note that

$$\begin{aligned} u^T g(t,x,u) &= u^T e(t,x) + u^T E(t,x) + u^T u \\ &> - \|e(t,x)\| \|u\| - \|E(t,x)\| \|u\|^2 + \|u\|^2 \\ &> - \rho^o(t,x)\|u\| - c\|u\|^2 + \|u\|^2 \\ &= (1-c)\|u\|^2 - \rho^o(t,x)\|u\| \ ; \end{aligned}$$

hence

$$u^T g(t,x,u) > (1-c)\|u\|^2 - \rho^o(t,x)\|u\| \tag{5.34}$$

for all $(t,x,u) \in T \times X \times U$, and (5.31) is satisfied with

$$\rho = \rho^o/(1-c). \tag{5.35}$$

2) The function f^s need not be the same for each $F \in F$; however, each f^s must have the same Lyapunov function V.

D. PROPERTIES OF SYSTEMS WITH L-G CONTROLLERS

We have the following theorem.

Theorem 5.1. Consider any uncertain system described by (5.1)-(5.2) (with $X = \mathbb{R}^n$*) which satisfies Assumption A4. If p is any corresponding L-G controller (as given by (5.25)-(5.27)) for which*

$$\dot{x}(t) = F(t,x(t), p(t,x(t))) \tag{5.36}$$

has existence and indefinite continuation of solutions for all $F \in \mathcal{F}$, then p stabilizes (5.1)-(5.2) about zero.

Proof. See Ref. (69).

Looking at Theorem 5.1, it can be seen that we have not completely solved the original problem for an uncertain system presented in sec. V.C. To do so, we need to exhibit an *L-G* controller p which assures existence and indefinite continuation of solutions to (5.36) (as defined in Definitions 2.1 and 2.2) for all $F \in \mathcal{F}$. This, however, is not possible in general. Except in special cases[10], it is not possible to obtain a function p, satisfyng (5.25) and (5.26) for all $(t,x) \in T \times X$, which is continuous in x. Thus, one cannot assure that $f:T \times X \rightarrow \mathbb{R}^n$, given by

$$f(t,x) = F(t,x,p(t,x)) \;\forall\; (t,x) \in T \times X$$

is continuous in x. Hence (5.36) does not satisfy the usual requirements for existence of solutions; see Theorem 2.1

In view of the above, we need to relax the requirements of the original problem statement. Here are two possible relaxations:

1) Relax the requirements which must be met by a function in order to be considered a solution of (5.36). This is the approach taken in Refs. (29-37) where the notion of generalized solutions is introduced and the *L-G* controllers solve the relaxed problem for the systems considered there.

2) Relax the requirement of g.u.a.s. of (5.36) about zero. This is what is done in Refs. (39-49,60) and what we do in the next chapter by introducing the notion of practical stabilization. In this approach, one may solve the relaxed problem with controllers which are continuous in the state; hence they are more desirable from the viewpoint of practical implementation as well.

VI. RELAXED PROBLEM STATEMENT - PRACTICAL STABILIZATION

Before introducing a relaxed problem statement, we need some new notions. Consider a system described by (2.1), i.e.,

$$\dot{x}(t) = f(t,x(t)) \tag{6.1}$$

[10] For example, $\rho(t,x) = \gamma(t,x)\|\alpha(t,x)\|$ for all $(t,x,u) \in T \times X \times U$, where $\gamma:T \times X \rightarrow \mathbb{R}_+$ is Carathéodory.

where $f: T \times X \to \mathbb{R}^n$ and suppose that $x^* \in \overline{X}$.

For any subset B of $\mathbb{R}^n$, we have the following definition.

Definition 6.1. The solutions of (6.1) are globally uniformly ultimately bounded (g.u.b.) *within B iff, given any compact subset C of X, there exists $T(C) \in \mathbb{R}_+$ such that, if $x(\cdot): [t_o, \infty) \to X$ is any solution of (6.1) with $x(t_o) \in C$, $x(t) \in B$ for all $t \geq t_o + T(C)$.*

If $B \subset \mathbb{R}^n$ is a neighborhood of x^*, we have the following definition.

Definition 6.2. System (6.1) B-tracks x^* *or* tracks x^* to within B *iff:*

1) The solutions of (6.1) are g.u.b.

2) There exists a neighborhood B_o of x^* such that, if $x(\cdot):[t_o, t_1) \to X$ is any solution of (6.1) with $x(t_o) \in B_o$, then $x(t) \in B$ for all $t \in [t_o, t_1)$.

3) The solutions of (6.1) are g.u.u.b. within B.

Remark 6.1. If (6.1) B-tracks x^* for any neighborhood B of x^*, it is g.u.a.s. about x^*.

The following theorem yields sufficient conditions for B-tracking of the zero state when $X = \mathbb{R}^n$.

Theorem 6.1. Consider any system described by (6.1) with $X = \mathbb{R}^n$ and suppose there exist a candidate Lyapunov function $V: T \times X \to \mathbb{R}_+$, a class K function $\gamma_3: \mathbb{R}_+ \to \mathbb{R}_+$, and a constant $c_3 \in \mathbb{R}_+$, which satisfy

$$\lim_{r \to \infty} \gamma_3(r) > c_3 \quad , \tag{6.2}$$

such that for all $(t,x) \in T \times X$,

$$\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x) f(t,x) \leq -\gamma_3(\|x\|) + c_3 \ . \tag{6.3}$$

Then, (6.1) tracks the zero state to within any neighborhood[11] *B of $B(\overline{d})$, where*

$$B(\overline{d}) = \{x \in \mathbb{R}^n : \|x\| \leq \overline{d}\}, \tag{6.4}$$

$$\overline{d} = \overline{\gamma}_1^{-1}(\gamma_2(\overline{r})), \qquad (6.5)^{12}$$

[11] If $V \subset \mathbb{R}^n$, then a neighborhood of V is any subset of $\mathbb{R}^n$ which contains an open set containing V.

[12] For the definition of $\overline{\gamma}^{-1}$ and $\overline{\gamma}^{-1}$ and some of their properties see Appendix, sec. B.

$$\bar{r} = \bar{\gamma}_3^{-1}(c_3), \tag{6.6}$$[12]

and γ_1, γ_2*:* $\mathbb{R}_+ \to \mathbb{R}_+$ *are any class* KR *functions for which*

$$\gamma_1(\|x\|) \leq V(t,x) \leq \gamma_2(\|x\|) \tag{6.7}$$

for all $(t,x) \in T \times X$.

Proof: See Ref. (69).

We now introduce:

Definition 6.3. A collection P *of feedback control functions* $p{:}T \times X \to U$ practically stabilizes *(3.1)-(3.2) about* x^* *iff, given any neighborhood* B *of* x^*, *there exists* $p \in P$ *such that for all* $F \in F$,

$$\dot{x}(t) = F(t,x(t),p(t,x(t)))$$

has existence and indefinite continuation of solutions and B-tracks x^*.

The relaxed problem we shall consider is that of obtaining a collection P of feedback control functions which practically stabilizes (3.1)-(3.2) about x^* for a given F.

Definition 6.4. (3.1)-(3.2) is practically stabilizable *about* x^* *iff there exists there exists a collection* P *of feedback control functions which practically stabilizes (3.1)-(3.2) about* x^*.

We now present a theorem which is useful in the synthesis of zero-state practically stabilizing sets of feedback controllers for uncertain systems whose state space is $\mathbf{R}^n$. For a given uncertain system, the theorem yields criteria which, if satisfied by a collection of feedback control functions, assure that the collection is a practically stabilizing collection.

Theoren 6.2. Consider an uncertain system described by (3.1)-(3.2) with $X = \mathbb{R}^n$ *and suppose that* P *is a collection of feedback control functions* $p{:}T \times X \to U$. *If there exists a candidate Lyapunov function* $V{:}T \times X \to \mathbb{R}_+$ *and a class* K *function* $\gamma_3{:}\ \mathbb{R}_+ \to \mathbb{R}_+$ *such that given any* $c_3 > 0$ *there exists* $p \in P$ *which assures that for all* $F \in F$,

$$\dot{x}(t) = F(t,x(t),\ p(t,x(t)))$$

has existence and indefinite continuation of solutions and

[13] For the definition of $\bar{\gamma}_1^{-1}$ and $\bar{\gamma}_3^{-1}$ and some of their properties see Appendix, sec. B.

$$\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)F(t,x,p(t,x)) \leq -\gamma_3(\|x\|) + c_3 \tag{6.8}$$

for all $(t,x) \in T \times X$, then P practically stabilizes (3.1)-(3.2) about zero.

Proof: This theorem follows from Theorem 6.1 and the fact that given any neighborhood $\mathcal{B}$ of the origin in $\mathbb{R}^n$, there exists $c_3 > 0$ which satisfies (6.2) and assures that $\mathcal{B}$ is a neighborhood of $B(\overline{d})$ as given by (6.4)-(6.6).

In the next chapter, we present some practically stabilizing controller sets whose design is based on meeting the criteria in the above theorem.

VII. MODIFIED *L-G* CONTROLLERS

Consider again a member of the class of uncertain systems presented in sec. V.A., i.e., consider an uncertain system described by (5.1)-(5.2) where $F: T \times X \times U \to \mathbb{R}^n$ $X, = \mathbb{R}^n$, $U = \mathbb{R}^m$, and F satisfies Assumptions A1 and A2.

In this chapter we present some zero-state practically stabilizing controller sets for such a system. Each controller presented is a continuous-in-state approximation to some *L-G* controller presented in sec. V.B.

Taking any quintuple (f^0, b^0, V, ρ^0, c) which assures satisfaction of Assumption A2, a proposed set of modified *L-G* controllers for practical stabilizability is the set P of strongly Carathéodory functions $p_\varepsilon : T \times X \to U$, $\varepsilon > 0$, which satisfy

$$\|p_\varepsilon(t,x)\| \leq \rho^c(t,x), \tag{7.1}$$

$$p_\varepsilon(t,x) = -\frac{\mu(t,x)}{\|\mu(t,x)\|}\rho^c(t,x) \quad \text{if} \quad \|\mu(t,x)\| > \varepsilon, \tag{7.2}$$

$$\rho^c(t,x) > \rho(t,x), \tag{7.3}$$

where

$$\mu(t,x) = \rho^c(t,x)B^{0T}(t,x)\frac{\partial V^T}{\partial x}(t,x), \tag{7.4}$$

$$\rho(t,x) = \rho^0(t,x)/(1-c) \tag{7.5}$$

for all $(t,x) \in T \times X$.

As a particular example of a function satisfying the above requirements on p_ε, consider p_ε given by

$$p_\varepsilon(t,x) = \begin{cases} -\dfrac{\mu(t,x)}{\|\mu(t,x)\|}\,\rho(t,x) & \text{if} \quad \|\mu(t,x)\| > \varepsilon \\[2ex] -\dfrac{\mu(t,x)}{\varepsilon}\,\rho(t,x) & \text{if} \quad \|\mu(t,x)\| < \varepsilon, \end{cases} \tag{7.6}$$

$$\mu(t,x) = \rho(t,x)B^{0T}(t,x)\,\frac{\partial V^T}{\partial x}(t,x) \tag{7.7}$$

for all $(t,x) \in T \times X$.

We now have the following theorem.

Theorem 7.1. Consider any uncertain system described by (5.1)-(5.2) (with $X = \mathbb{R}^n$ *and* $U = \mathbb{R}^m$*) which satisfies Assumptions A1 and A2; let* P *be any corresponding set of modified L-G controllers as defined above; and suppose that* $\gamma_1, \gamma_2 : \mathbb{R}_+ \to \mathbb{R}_+$ *are KR functions and* $\gamma_3 : \mathbb{R}_+ \to \mathbb{R}_+$ *is a K function which assure that V is a Lyapunov function for* $\dot{x}(t) = f^\rho(t,x(t))$*. Then, for each* $p_\varepsilon \in P$ *for which*

$$\lim_{r\to\infty} \gamma_3(r) > 2\varepsilon, \tag{7.8}$$

and for each $F \in F$,

$$\dot{x}(t) = F(t,x(t),\, p_\varepsilon(t,x(t))) \tag{7.9}$$

has existence and indefinite continuation of solutions and tracks the zero state to within any neighborhood B *of* $B(\bar{d}_\varepsilon)$ *where*

$$B(\bar{d}_\varepsilon) = \{x \in \mathbb{R}^n : \|x\| < \bar{d}_\varepsilon\}, \tag{7.10}$$

$$\bar{d}_\varepsilon = \bar{\gamma}_1^{-1}(\gamma_2(\bar{r}_\varepsilon)), \tag{7.11}$$

$$\bar{r}_\varepsilon = \bar{\gamma}_3^{-1}(2\varepsilon). \tag{7.12}$$

Proof. See Ref. (69).

From Theorem 7.1, we may deduce the following corollary.

Corollary 7.1. Consider any uncertain system described by (5.1)-(5.2) (with $X = \mathbb{R}^n$ *and* $U = \mathbb{R}^m$*) which satisfies Assumptions A1 and A2 and let* P *be any corresponding set of modified L-G controllers as defined above. The,* P *practically stabilizes (5.1)-(5.2) about zero.*

Proof. This corollary follows from Theorem 7.1 and the fact that given any neighborhood B of the origin in $\mathbb{R}^n$, there exists $\varepsilon > 0$ which satisfies (7.8) and assures that B is a neighborhood of $B(\bar{d}_\varepsilon)$ as given by (7.10)-(7.12).

VIII. APPENDIX

A. *CARATHÉODORY FUNCTIONS*

In sec. A, T is any non-empty Lebesgue measurable subset of $\mathbb{R}$ and X is any non-empty subset of $\mathbb{R}^n$.

Definition 8.1. 1) A function $f:T \times X \to \mathbb{R}^p$ is Carathéodory *iff: for each $t \in T$, $(f(t,\cdot)$ is continuous; for each $x \in X$, $f(\cdot,x)$ is Lebesgue measurable; and, for each compact subset C of $T \times X$, there exists a Lebesgue integrable function $M_C(\cdot)$ such that, for all $(t,x) \in C$,*

$$\| f(t,x) \| \leq M_C(t).$$

2) A function $f:T \times X \to \mathbb{R}^p$ is strongly Carathéodory *iff it satisfies 1) with $M_C(\cdot)$ replaced by a constant M_C.*

B. K, KR *FUNCTIONS*

Definition 8.2 1) A function $\gamma:\mathbb{R}_+ \to \mathbb{R}_+$ belongs to class K iff it is continuous and satisfies

$$r_1 < r_2 \Rightarrow \gamma(r_1) < \gamma(r_2) \quad \forall \; r_1, r_2 \in \mathbb{R}_+ ,$$

$$\gamma(0) = 0, \; r > 0 \Rightarrow \gamma(r) > 0.$$

2) A function $\gamma:\mathbb{R}_+ \to \mathbb{R}_+$ belongs to class KR *iff it belongs to K and*

$$\lim_{r\to\infty} \gamma(r) = \infty.$$

Lemma 8.1. If γ belongs to K, then there exist functions $\underline{\gamma}^{-1}, \bar{\gamma}^{-1}:[0,\ell) \to \mathbb{R}_+$, where $\ell = \lim\limits_{r\to\infty} \gamma(r)$, such that

$$\underline{\gamma}^{-1}(s) = \inf\{ f \in \mathbb{R}_+ : \gamma(r) = s\} \; \forall \; s \in [0,\ell), \tag{6.1}$$

$$\bar{\gamma}^{-1}(s) = \sup\{ r \in \mathbb{R}_+ : \gamma(r) = s\} \; \forall \; s \in [0,\ell), \tag{6.2}$$

and these functions are strictly increasing and satisfy

$$\gamma(\underline{\gamma}^{-1}(s)) = s = \gamma(\bar{\gamma}^{-1}(s)) \;\forall\; s \in [0,\ell), \tag{6.3}$$

$$\underline{\gamma}^{-1}(\gamma(r)) \leq r \leq \bar{\gamma}^{-1}(\gamma(r)) \;\forall\; r \in \mathbb{R}_+. \tag{6.4}$$

Proof. See Ref. (69)

Acknowledgement

This article is based on research supported by the National Science Foundation and the U.S. Air Force Office of Scientific Research; it was carried out while G. Leitmann was the recipient of a U.S. Senior Scientist Award of the Alexander von Humboldt Foundation.

REFERENCES

1. H.J. Kushner: On the Status of Optimal Control and Stability for Stochastic Systems. IEEE Int. Conv. Rec. 14, 143, 1966.

2. J.K. Åström: Introduction to Stochastic Control Theory. Academic Press, New York, 1970.

3. E.A. Coddington and N. Levinson: Theory of Ordinary Differential Equations. McGraw-Hill, New York, 1955.

4. J.K. Hale: Ordinary Differential Equations. Krieger, 1980.

5. M.W. Hirsch and S. Smale: Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, New York, 1974.

6. R.E. Kalman and J.E. Bertram: Control System Analysis and Design via the "Second Method" of Lyapunov, I: Continuous-Time Systems. J. Basic Engineering 82, 371, 1960.

7. J.P. LaSalle and S. Lefchetz: Stability by Liapunov's Direct Method with Aplications. Academic Press, New York, 1961.

8. W. Hahn: Stability of Motion. Springer-Verlag, Berlin, 1967.

9. L. Cesari: Asymptotic Behavior and Stability Problems in Ordinary Differential Equations. Springer-Verlag, New York, 1971.

10. J.P. LaSalle: The Stability of Dynamical Systems. SIAM, 1976.

11. M.G. Safonov: Stability and Robustness of Multivariable Feedback Systems. MIT Press, Cambridge, 1980.

12. J.C. Doyle and G. Stein: Multivariable Feedback Design: Concepts for a Classical/Modern Synthesis. IEEE Trans. on Automatic Control, AC-26, 4, 1981.

13. D.G. Luenberger: Optimization by Vector Space Methods. Wiley, New York, 1966.

14. D.M. Wiberg: State Space and Linear Systems. McGraw-Hill, New York, 1971.

15. A.I. Lur'e: Some Nonlinear Problems in the Theory of Automatic Control (in Russian). Gostekhizdat, Moscow, 1951. German translation, Akademie Verlag, 1957. English translation, Her Majesty's Stationery Office, 1957.

16. A.M. Letov: Stability of Nonlinear Regulating Systems (in Russian). Izdatel'stvo Technichesko-Teoreticheskoi Literatury, Moscow, 1955.

17. R.W. Bass: Discussion of "Die Stabilität von Regelsystemen mit nachgebender Rückführung" by A.M. Letov. Proc. Heidelberg Conf. Automatic Control, 209, 1957.

18. J.P. LaSalle: Stability and Control. SIAM J. Control, 1, 3, 1962.

19. G.W. Johnson: Synthesis of Control Systems with Stability Constraints Via the Direct Method of Lyapunov. IEEE Trans. Automatic Control AC-9, 270, 1964.

20. L.P. Grayson: Two theorems on the Second Method. IEEE Trans. Automatic Control, AC-9, 587, 1964.

21. R.V. Monopoli: Discussion on 'Two Theorems on the Second Method'. IEEE Trans. Automatic Control, AC-10, 369, 1965.

22. R.V. Monopoli: Synthesis Techniques Employing the Direct Method. IEEE Trans. Automatic. Control, AC-10, 369, 1965.

23. R.V. Monopoli: Corrections to: Synthesis Techniques Employing the Direct Method. IEEE Trans. Automatic Control., AC-11, 631, 1966.

24. L.P. Grayson: The Status of Synthesis Using Lyapunov's Method. Automatica 3, 91, 1965.

25. R.V. Monopoli: Engineering Aspects of Control System Design Via the 'Direct Method' of Lyapunov. CR-654, NASA, 1966.

26. G. Leitmann: A Simple Differential Game. J. Optimiz. Theory Appl. 2, 220, 1968.

27. S. Gutman: Differential Games and Asymptotic Behavior of Linear Dynamical Systems in the Presence of Bounded Uncertainty. (Ph.D. Dissertation) UC Berkeley, 1975.

28. S. Gutman and G. Leitmann: On a Class of Linear Differential Games. J. Optimiz. Theory Appl. 17, 511, 1975.

29. S. Gutman and G. Leitmann: Stabilizing Control for Linear Systems with Bounded Parameter and Input Uncertainty. Proc. 2nd IFIP Conf. Optimiz. Techniques, Springer-Verlag, Berlin, 1975.

30. G. Leitmann: On Stabilizing a Linear System with Bounded State Uncertainty, in Topics in Contemporary Mechanics. Springer-Verlag, Vienna, 1974.

31. S. Gutman: Uncertain Dynamical Systems - A Differential Game Approach. NASA TMX-73, 135, 1976.

32. G. Leitmann: Stabilization of Dynamical Systems under Bounded Input Disturbance and Parameter Uncertainty, in Differential Games and Control Theory II. (E.O. Roxin, P.-T. Liu, and R.L. Sternberg, ed.), 47, Marcel Dekker, New York, 1976.

33. S. Gutman and G. Leitmann: Stabilizing Feedback Control for Dynamical Systems with Bounded Uncertainty. Proc. IEEE Conf. Decision Control, 1976.

34. G. Leitmann: Guaranteed Asymptotic Stability for a Class of Uncertain Linear Dynamical Systems. J. Optimiz. Theory Appl. 27, 99, 1979.

35. G. Leitmann: Guaranteed Asymptotic Stability for Some Linear Systems with Bounded Uncertainties. J. Dynam. Syst. Meas. Contr., 101, 212, 1979.

36. S. Gutman: Uncertain Dynamical Systems--Lyapunov Min-Max Approach. IEEE Trans. Automatic Control, AC-24, 437, 1979.

37. S. Gutman and Z. Palmor: Properties of Min-Max Controllers in Uncertain Dynamical Systems. SIAM J. Contr. Optimiz., 20, 850, 1982.

38. T. Yoshizawa: Lyapunov's Function and Boundedness of Solutions. Bol. Soc. Mat. Mex. Ser. 2 5, 146, 1960.

39. G. Leitmann: Guaranteed Ultimate Boundedness for a Class of Uncertain Linear Dynamical Systems, in Differential Games and Control Theory III. (P.-T. Liu and E. Roxin, ed.), 29, Marcel Dekker, New York, 1978.

40. G. Leitmann: Guaranteed Ultimate Boundedness for a Class of Uncertain Linear Dynamical Systems. IEEE Trans. Automatic Control, AC-23, 1109, 1978.

41. G. Leitmann: On the Efficacy of Nonlinear Control in Uncertain Linear Systems. J. Dynam. Syst. Meas. Contr., 102, 95, 1981.

42. M. Corless and G. Leitmann: Continuous State Feedback Guaranteeing Uniform Ultimate Boundedness for Uncertain Dynamic Systems. IEEE Trans. Automat. Contr., AC-26, 1139, 1981.

43. M. Corless and G. Leitmann: Erratum to 'Continuous State Feedback Guaranteeing Uniform Ultimate Boundedness for Uncertain Dynamic Systems'. IEEE Trans. Automat. Contr., AC-28, 249, 1983.

44. G. Leitmann: Deterministic Control of Uncertain Systems. Acta Astronautica, 7, 1457, 1980.

45. W. Breinl and G. Leitmann: Zustandsrückführung fur dynamische Systeme mit Parameterunsicherheiten. Regelungstechnik, 31, 95, 1983.

46. E.P. Ryan and M. Corless: Ultimate Boundedness and Asymptotic Stability of a Class of Uncertain Dynamical Systems via Continuous and Discontinuous Feedback Control. IMA J. Math. Contr. Inf., 1, 223, 1984.

47. E.P. Ryan, G. Leitmann and M. Corless: Practical Stabilizability of Uncertain Dynamical Systems: Application to Robotic Tracking. J. Optimiz. Theory. Appl., 47, 235, 1985.

48. M. Corless, G. Leitmann and E.P. Ryan: Tracking in the Presence of Bounded Uncertainties. Proc. 4th IMA Int. Conf. Control Theory, Cambridge University, England, 1984.

49. A. Steinberg and M. Corless: Stabilizing Uncertain Systems wit Partial State Measurement. Proc. 23rd IEEE Conf. on Decision and Control, Phoenix, Arizona, 1984.

50. G. Leitmann and H.Y. Wan, Jr.: A Stabilization Policy for an Economy with Some Unknown Characteristics. J. Franklin Institute, 306, 23, 1978.

51. G. Leitmann and H.Y. Wan, Jr.: Macro-Economic Stabilization Policy for an Uncertain Dynamic Economy, in New Trends in Dynamic System Theory and Economics. Academic Press, New York, 1979.

52. G. Leitmann and H.Y. Wan, Jr.: Performance Improvement of Uncertain Macroeconomic Systems, in Dynamic Optimization and Mathematical Economics. (P.-T. Liu, ed.), Plenum Press, New York, 1979.

53. C.S. Lee and G. Leitmann: On Optimal Long-Term Management of Some Ecological Systems Subject to Uncertain Disturbances. Int. J. Syst. Sci., 14, 979, 1983.

54. V.I. Utkin: Variable Structure Systems with Sliding Mode: A Survey. IEEE Trans. Automatic Control, AC-22, 1977.

55. K.-K.D. Young: Design of Variable Structure Model-Following Control Systems. IEEE Trans. Automatic. Control., AC-23, 1079, 1978.

56. J.J. Slotine and S.S. Sastry: Tracking Control of Non-Linear Systems Using Sliding Surfaces, with Application to Robot Manipulators. Int. J. Control, 38, 465, 1983.

57. E.P. Ryan: A Variable Structure Approach to Feedback Regulation of Uncertain Dynamical Systems. Int. J. Control, 38, 1121, 1983.

58. J.S. Thorp and B.R. Barmish: On Guaranteed Stability of Uncertain Systems via Linear Control. J. Optimiz. Theory Appl., 35, 559, 1981.

59. B.R. Barmish and G. Leitmann: On Ultimate Boundedness Control of Uncertain Systems in the Absence of Matching Conditions. IEEE Trans. Automatic Control, AC-27, 1253, 1982.

60. B.R. Barmish, M. Corless and G. Leitmann: A New Class of Stabilizing Controllers for Uncertain Dynamical Systems. SIAM J. Contr. Optimiz. 21, 246, 1983.

61. C.V. Hollot and B.R. Barmish: Optimal Quadratic Stabilizability of Uncertain Linear Systems. Proc. 18th Allerton Conf. Communications Contr. Computing, 1980.

62. B.R. Barmish, I.R. Petersen and A. Feuer: Linear Ultimate Boundedness Control of Uncertain Dynamical Systems. Automatica, 19, 523, 1983.

63. I.R. Petersen and b.R. Barmish: Control Effort Considerations in the Stabilization of Uncertain Dynamical Systems. Proc. Amer. Contr. Conf., 1984.

64. S.S.L. Chang and T.K.C. Peng: Adaptive Guaranteed Cost Control of Systems with Uncertain Parameters. IEEE Trans. Automatic Control AC-17, 474, 1972.

65. R.V. Patel, M. Toda, and B. Sridhar: Robustness of Linear Quadratic State Feedback Designs in the Presence of System Uncertainty. IEEE Trans. Automatic Control, AC-22, 1977.

66. A. Vinkler and L.J. Wood: Multistep Guaranteed Cost Control of Linear Systems with Uncertain Parameters. J. Guidance Control, 2, 449, 1979.

67. P. Molander: Stabilisation of Uncertain Systems. LUTFD2/(TFRT-1020)/1-111. Lund Institute of Technology, 1979.

68. M. Eslami and D.L. Russell: On Stability with Large Parameter Variations: Stemming from the Direct Method of Lyapunov. IEEE Trans. Automatic Control, AC-25, 1980.

69. M. Corless: Control of Uncertain Systems. Ph.D. Dissertation, University of California, Berkeley, 1984.

70. M. Corless and G. Leitmann: Adaptive Control for Uncertain Dynamical Systems, in Dynamical Systems and Microphysics: Control Theory and Mechanics. (A. Blaquiere and G. Leitmann, eds.). Academic Press, New York, 1984.

71. M. Corless and G. Leitmann: Adaptive Long-Term Management of Some Ecological Systems Subject to Uncertain Disturbances, in Optimal Control Theory and Economic Analysis 2. (G. Feichtinger, ed.), Elsevier Science Publishers, Amsterdam, Holland, 1985.

72. M. Corless, D.P. Goodall, G. Leitmann and E.P. Ryan: Model-Following Controls for a Class of Uncertain Dynamical Systems. Proc. IFAC Conf. on Identification and System Parameter Estimation, York University, England, 1985.

73. E.P. Ryan, G. Leitmann and M. Corless: Practical Stabilizability of Uncertain Dynamical Systems: Application to Robotic Tracking. J. Optimiz. Theory Applic., 47, 235, 1985.

74. I.J. Ha and E.G. Gilbert: Robust Tracking in Nonlinear Systems and Its Application to Robotics. Proc. 24th Conf. on Decision and Contr., Ft. Lauderdale, Florida, 1985.

75. G. Leitmann, E.P. Ryan and A. Steinberg: Feedback Control of Uncertain Systems: Robustness with Respect to Neglected Actuator and Sensor Dynamics, Int. J. Control 43, 1243, 1986.

76. M. Corless and J. Manela: Control of Uncertain Discrete-Time Systems, Proc. American Control Conf., Seattle, Washington, 1986.

77. G. Leitmann, C.S. Lee and Y.H. Chen: Decentralized Control for a Large Scale Uncertain River System. Proc. IFAC Workshop on Modelling, Decisions and Games for Social Phenomena, Beijing, China, 1986.

NONLINEAR CONTROL THEORY AND DIFFERENTIAL ALGEBRA

Michel FLIESS

- o - o - o - o - o - o -

Laboratoire des Signaux et Systèmes
CNRS -ESE, Plateau du Moulon
91190 Gif-sur-Yvette, France

CONTENTS

INTRODUCTION

During the past twenty years, differential geometry and functional expansions have offered powerful means for getting many remarkable results on controllability, realization and static state feedback synthesis of nonlinear systems (see, e.g., the books [2,12,15,19,32] and the survey papers [13,36]). It was therefore believed that the aforementioned mathematical tools were able to give, at least locally, the right nonlinear analogues of most parts of the algebraic and geometric theories of constant linear systems. However, some basic problems remained unsolved, certainly the most important being the input-output inversion.

This note outlines a new framework for nonlinear systems which can be regarded as a sequel to our recent solution [8,10] of the inversion problem by differential algebra. Quite surprisingly, it turns out that many control theoretic concepts, which were taken for granted in the literature, should be reexamined in the light of our approach. The input-output behaviour of a large class of engineering systems like nonlinear circuits, swing dynamics or interconnected power systems is not given by the usual state-space equations, but by a finite number of possible implicit differential equations, some of which are algebraic, i.e., differential equations of order zero. This is interpreted as the differential analogue of the notion of algebraic field extension [25,28]. A new definition of state is given which employs non-differential transcendence basis and local differential algebra [31]. This gives a clear-cut answer to questions raised in circuit theory [4,17] when the impossibility of a global state-space description of many realistic examples is noticed.

Several other topics, such as inversion, series connection, exact model matching, and controller and observability canonical forms, are also treated. Feedback synthesis problems can also be studied by our methods [9]. Moreover, the parallelism with discrete-time systems can be restored when using difference algebra [11]. Finally, let us cite an earlier paper [14] on the connection between bilinear systems and the Picard-Vessiot theory, and a recent research announcement [29] employing differential Galois theory.

Acknowledgements. The author would like to thank Dr. T. Roska for some helpful comments.

I. A SHORT OVERVIEW OF DIFFERENTIAL ALGEBRA[1]

I.1. **Differential algebra** was originated more than fifty years ago by the American mathematician J.F. Ritt [30] at a time when **commutative algebra** began to arrive at its present shape [1]. Ritt's aim was to create a tool which would play the same role in respect to differential equations as commutative algebra to algebraic equations either in number theory or in algebraic geometry.

I.2. An **(ordinary) differential ring** is a commutative ring R, with $1 \neq 0$, equipped with one derivation $R \rightarrow R$, $a \rightarrow \frac{da}{dt} = \dot{a}$, such that

$$\forall \; a, b \in R, \quad \frac{d}{dt}(a+b) = \dot{a} + \dot{b},$$
$$\frac{d}{dt}(ab) = \dot{a}b + a\dot{b}.$$

Usual words from commutative algebra, like **ideal**, **field**,..., to which "differential" is added, have obvious meanings.

I.3. Let K and L be two differential fields such that $K \subset L$. As in usual algebra, two situations are possible:

- Each element of L satisfies an algebraic differential equation with coefficients in K, i.e., each element of L is **differentially algebraic** over K. Then L is said to be a **differentially algebraic extension** of K.

- There exists at least one element of L which does not satisfy any differential algebraic equation with coefficients in K, i.e., which is **differentially transcendental** over K. Then L is said to be a **differentially transcendental extension** of K. The maximum number of such elements, which are differentially algebraically independent, is called the **differential transcendence degree** of L over K. This important integer will be written diff.tr.d°L/K.

I.4. Take three differential fields $K \subset L \subset M$. The identity

$$\text{diff.tr.d°}M/K = \text{diff.tr.d°}M/L + \text{diff.tr.d°}L/K$$

will be used several times.

I.5. Remark. The most important feature of this communication is to show that many natural control problems have simple answers when employing the language of field theory.

[1] See [25,28]

II. WHAT IS AN INPUT-OUTPUT SYSTEM?

II.1. In algebraic geometry it is customary to work with a huge field, called the **universal domain**, which contains all the elements [41]. We will do the same here by considering an ordinary differential field Ω which is a **universal extension** [25] of the field $\mathbf{Q}$ of rational numbers.

II.2. Take $m + p$ elements $u = (u_1,\dots,u_m)$, $y = (y_1,\dots,y_p)$ in Ω. Assume moreover that $u_1,\dots,u_m$ are **differential indeterminates**, i.e., that they are algebraically differentially independent over $\mathbf{Q}$ [25]. Let $\mathbf{Q}\langle u\rangle \subset \Omega$ be the smallest differential field containing $\mathbf{Q}, u_1,\dots,u_m$. A typical element of $\mathbf{Q}\langle u\rangle$ is a rational expression of the form

$$\frac{u_1^{(3)}\dot{u}_2 - (u_1)^4}{7(u_2^{(2)})^5} .$$

II.3. Definition. To define a system with input u and output y amounts to saying that the components of y are differentially algebraic over $\mathbf{Q}\langle u\rangle$.

II.4. The preceding definition means that the components of u and y are related by a finite number of implicit differential equations. Let us emphasize once again that this fact is encountered in many physical and engineering case studies[2].

It should not be believed that such a definition of an I/O system is restricted to **algebraic differential equations** [25], i.e., to equations which are polynomial in the components of $u, \dot{u}, \dots, y, \dot{y}, \dots$. One can also tackle differentially algebraic coefficients and therefore all realistic case studies. In order to make this statement more concrete, let us consider the Josephson junction circuit described by [4]

$$\dot{y} = E - RI \sin y \qquad (m=0,\ p=1),$$

where E, R, I are device constants. It is easy to verify that the non-differential transcendence degree of $\mathbf{Q}\langle y, \sin y\rangle$ over $\mathbf{Q}$ is finite, and therefore that y is differentially algebraic over $\mathbf{Q}$.

II.5. Many times in engineering, I/O behaviours are defined via functional expansions, like Volterra series or generating series (see [13,32] and the references therein). Contrarily to our approach, such a viewpoint seems to our opinion quite unrealistic since it implies the knowledge of an infinite number of coefficients. Moreover, due to convergence properties, the

[2] See [3,4,17,27,33] for various examples.

adequation between physical systems and functional expansions is in general only local.

III. SOME BASIC PROPERTIES

1. Inversion[3]

III.1.1. For constant linear systems, the inversion problem is quite trivial when using the frequency domain approach. In the nonlinear situation, it is simple to understand only in the case of a one-dimensional control [18,37]. Everything becomes clear with differential algebra.

III.1.2. Definition. The **differential output rank** is the differential transcendence degree of the differential field $\mathbf{Q}\langle y\rangle$ over $\mathbf{Q}$.

III.1.3. The following is not difficult:
Proposition. For a constant linear system, the differential output rank turns out to be equal to the rank of the transfer matrix.

III.1.4. The next definition is consistent with linear systems.
Definition. A system is said to be **(differentially) left invertible** (resp. **right invertible**) if, and only if, its differential output rank is equal to the number of controls (resp. outputs).

III.1.5. Proposition. (i) If a system is right invertible, then there are no relations between the components of the output, which are independent of the control and of the state.

(ii) If a system is left invertible, it is possible to recover the control from the output by a finite set of equations.

Proof. (i) The first statement is just a rephrasing of the fact that the differential transcendence degree of $\mathbf{Q}(y)$ over $\mathbf{Q}$ is p. It is the nonlinear analogue of the linear independence of the rows of the transfer matrix.

(ii) By I.4, we may write

$$\text{diff.tr.d}^{\circ}\mathbf{Q}\langle u,y\rangle/\mathbf{Q} = \text{diff.tr.d}^{\circ}\mathbf{Q}\langle u,y\rangle/\mathbf{Q}\langle y\rangle + \text{diff.tr.d}^{\circ}\mathbf{Q}\langle y\rangle/\mathbf{Q} ,$$

where

- $\text{diff.tr.d}^{\circ}\mathbf{Q}\langle y\rangle/\mathbf{Q} = m$ by assumption,
- $\text{diff.tr.d}^{\circ}\mathbf{Q}\langle u,y\rangle/\mathbf{Q} = m$, since we are assuming m independent controls.

[3] See [10] for details and references.

Therefore

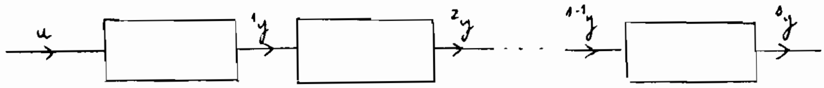
$$\text{diff.tr.d}^\circ Q\langle u,y\rangle/Q\langle y\rangle = 0.$$

This means that $u_1,\dots,u_m$ are differentially algebraic over $Q\langle y\rangle$.

III.1.6. Remark. It has been recently noticed [20,21] that the geometric approach for extending various important invariant integers in linear system theory to a nonlinear setting exhibits severe pathologies. Our differential output rank should be a good candidate for the right nonlinear analogue of the sum of the **zeros at infinity.**

2. Series connection

III.2.1.



In the above block diagram representing a series connection, the output of a system is the input of the next one. This is equivalent to a **tower**, i.e., an increasing sequence of differentially algebraic fields extensions:

$$Q\langle u\rangle \subset Q\langle u, {}^1y\rangle \subset Q\langle u, {}^1y, {}^2y\rangle \subset \ \dots\dots \subset Q\langle u, {}^1y, \dots, {}^sy\rangle$$

The notion of tower of fields is quite important in number theory.

III.2.2. Remark. This analysis shows that the problem of decomposing a given system into a series connection, the elements of which should be as "simple" as possible, is strongly related to a Galois theory for differential fields [25,28]. See [7] for another approach.

3. Exact model matching[*]

III.3.1. We are given two systems with the same input u and outputs y and z $= (z_1,\dots,z_r)$. The exact model matching problem consists of finding a system with input z and output y such that the diagram

$$\begin{tikzcd} u \arrow[rr] \arrow[dr] & & y \\ & z \arrow[ur] & \end{tikzcd}$$

commutes.

[*] Result recently obtained by Conte, Moog and Perdon [5].

III.3.2. Proposition. The exact model matching problem is soluble if, and only if, the differential transcendence degrees of $\mathbf{Q}\langle y,z\rangle$ and $\mathbf{Q}\langle z\rangle$ over $\mathbf{Q}$ are equal.

Proof. From

$$\text{diff.tr.d}^\circ\mathbf{Q}\langle y,z\rangle/\mathbf{Q} = \text{diff.tr.d}^\circ\mathbf{Q}\langle y,z\rangle/\mathbf{Q}\langle z\rangle + \text{diff.tr.d}^\circ\mathbf{Q}\langle z\rangle/\mathbf{Q},$$

we see that

$$\text{diff.tr.d}^\circ\mathbf{Q}\langle y,z\rangle/\mathbf{Q} = \text{diff.tr.d}^\circ\mathbf{Q}\langle z\rangle/\mathbf{Q}$$

is equivalent to

$$\text{diff.tr.d}^\circ\mathbf{Q}\langle y,z\rangle/\mathbf{Q}\langle z\rangle = 0.$$

III.3.3. Remarks. (i) As an exercise we invite the reader to verify that for constant linear systems the preceding criterion reduces to the usual rank condition for transfer matrices.

(ii) For another approach to the same problem, see [6].

IV. STATE AND REALIZATION

IV.1. Discussions on the concept of **state** can be found in the literature [23,40] at the beginning of the state space area, which has become the mainstay of control theory since the sixties. These considerations, which are quite satisfactory for linear systems[5], have not been further examined for nonlinear ones. This fact caused several difficulties when systems could not be described by the usual state space form. Methods such as singular perturbations had to be employed in order to apply results and techniques from state space theory[6].

IV.2. Here is a first non rigourous attempt to describe what a state should be. A state $x = (x_1,\dots,x_n)$ is a set of n elements in Ω such that $\dot{x}_i, i=1,\dots,n$, and $y, j=1,\dots,p$, depend on $x,u,\dot{u},u^{(2)},\dots$.

IV.3. In differential algebra, a **constant** is an element with derivative zero. All the constants in a given differential field form a subfield. Call C, $\mathbf{Q}\subset C$, the field of constants of $\mathbf{Q}\langle u,y\rangle$. Clearly, a non-differential

[5] See nevertheless Willems' criticism [39].

[6] For a most interesting example, see [27].

transcendence basis of Q<u,y> over C<u> satisfies the preceding properties if the dependence is algebraic. Such a choice, however, could bring some trouble like the one described in the next example. Take the "memoryless" system $y = \sin u$, $m=p=1$. A sound **minimal realization** should have dimension zero. But the non-differential transcendence degree of Q<u,sin u> over **Q**<u> is one and would therefore imply a state space realization of dimension one.

IV.4. The solution is given by **local differential algebra** [31], where differential algebra is supplemented so as to take into account initial conditions. We will just sketch some of the ideas here by using plain words. Take analytic control $u_i = \sum\limits_{\nu\geq 0} a_{i\nu}\, t^{\nu}$, $i=1,\dots,m$, and initial conditions such that the output $y_j = \sum\limits_{\nu\geq 0} b_{j\nu}\, t^{\nu}$, $j=1,\dots,p$, is also analytic. To take into account the analytic dependence [41] between the $u_i^{(\alpha)}$'s, $y_j^{(\beta)}$'s, $i=1,\dots,m$, $j=1,\dots,p,\alpha,\beta\geq 0$, we look at the **Krull dimension** [16] of $\mathbf{R}[[u_i^{(\alpha)}-a_{i\alpha},\ y_j^{(\beta)}-b_{j\beta}\mid \alpha,\beta\leq\nu]]$, where **R** is the field of real numbers. It can be shown that this dimension remains constant and equal to d when $\nu\geq\nu_0$. This integer d is called the **minimal dimension** of the system. When it is equal to the non-differential transcendence degree of **Q**<u,y> over C<u>, the system is said to be **algebraic.**

IV.5. Consider now, for simplicity's sake, an algebraic system of minimal dimension d. The **minimal state** will be a non-differential transcendence basis $q = (q_1,\dots,q_d)$ of **Q**<u,y> over C. The derivatives $\dot{q}_k, k=1,\dots,d$, and the outputs $y_j, j=1,\dots,p$, are algebraically dependent on $q,u,\dot{u},\dots$:

$$\begin{cases} F_k(\dot{q}_k,q,u,\bar{u},\dots,u^{(s)}) = 0, \\ \Phi_j(y_j,q,u,\bar{u},\dots,u^{(s)}) = 0. \end{cases}$$

This means that the usual state space form

$$\begin{cases} \dot{q}_k = f_k(q,u,\dots,u^{(s)}) \\ y_j = \phi_j(q,u,\dots,u^{(s)}) \end{cases}$$

might only be locally valid. Another major change with realization theory in the differential geomettric setting [20,22,36] concerns the transformation between two minimal states, which here depends on the control and its derivatives.

IV.6. To the best of our knowledge, the local validity of nonlinear state space equations has been completely overlooked in the control literature. It shows the difficulty of obtaining a global differential geometric realization with reasonably weak assumptions (see [22] and the references therein). Among circuit theorists [4,17], however, such problems are well-known, and we believe to have offered here for the first time a clear mathematical explanation.

IV.7. The intimate connection between control and state has already been noticed in the literature [34,38] and was dealt with by employing the language of fibered manifolds.

V. CONTROLLER AND OBSERVABILITY CANONICAL FORMS

V.1. There have been several attempts to generalize linear canonical forms to a nonlinear setting (see, e.g., [43] and the references therein), and also some applications to control problems (see, e.g., [24]). We will show here that two of these canonical forms can be obtained very easily thanks to our methods.

V.2. Take the usual state space form

$$\dot{x}_i = A_i(x_1,\dots,x_n,\ u,\dots,u^{(s)}), \qquad i=1,\dots,n,$$

such that $x_1,\dots,x_n$ are differentially algebraic over $\mathbf{Q}\langle u\rangle$. By applying the differential algebraic generlization of the theorem of the **primitive element** [25,28], there exists an element ξ such that $\mathbf{Q}\langle u,\xi\rangle = \mathbf{Q}\langle u,x_1,\dots,x_n\rangle$. As in IV.4, let d be the first integer such that $\xi^{(d+1)}$ is analytically dependent on $\xi,\dot{\xi},\dots,\xi^{(d)},u,\dot{u},\dots$:

$$\Phi(\xi^{(d+1)},\xi^{(d)},\dots,\xi,u,\dots,u^{(s)}) = 0.$$

It can be solved locally as

$$\xi^{(d+1)} = a(\xi,\dots,\xi^{(d)},u,\dots,u^{(s)}).$$

Set $q_i = \xi^{(i)}, i=0,\dots,d$. We obtain the following local state space form, which can be regarded as a generalization of the linear **controller canonical form**:

$$\begin{cases} \dot{q}_o = q_1 \\ \text{-------} \\ \dot{q}_{d-1} = q_d \\ \dot{q}_d = a(q_o,\ldots,q_d,u,\ldots,u^{(s)}) \;. \end{cases}$$

See [35] for another approach.

V.3. For the sake of simplicity, take a system with a one-dimensional output y, i.e., $p=1$. As before, let d be the least integer such that $y^{(d+1)}$ is analytically dependent on $y,\dot{y},\ldots,y^{(d)},u,\dot{u},\ldots$:

$$\psi(y^{(d+1)},y^{(d)},\ldots,y,u,\ldots,u^{(s)}) = 0.$$

It can be solved locally as

$$y^{(d+1)} = b(y,\ldots,y^{(d)},u,\ldots,u^{(s)}).$$

Set $q_i = y^{(i)}$, $i=o,\ldots,d$. We obtain the following local state space form which can be regarded as a generalization of the linear **observability canonical form** (compare with [26,42]):

$$\begin{cases} \dot{q}_o = q_1 \\ \text{-------} \\ \dot{q}_{d-1} = q_d \\ \dot{q}_d = b(q_o,\ldots,q_d,u,\ldots,u^{(s)}) \\ y = q_o \;. \end{cases}$$

REFERENCES

[1] N. BOURBAKI, Eléments d'histoire des mathématiques, 2e éd., Hermann, Paris, 1969.

[2] A. BACCIOTTI, Fondamenti geometrici della teoria della controllabilità, Pitagora Editrice, Bologna, 1986.

[3] L.O. CHUA, Device modeling via basic nonlinear circuit elements, IEEE Trans. Circuits Systems, 27, 1980,pp.1014-1044.

[4] L.O. CHUA, Dynamic nonlinear networks: state-of-the-art, IEEE Trans. Circuits Systems, 27, 1980, pp.1059-1087.

[5] G. CONTE, C.H. MOOG and A.M. PERDON, in preparation.

[6] M.D. DI BENEDETTO and A. ISIDORI, The matching of nonlinear models via dynamic state feedback, SIAM J. Control Optimiz., 24, 1986.

[7] M. FLIESS, Décomposition en cascade des systèmes automatiques et feuilletages invariants, Bull. Soc. Math. France, 113, 1985, pp.285-293.

[8] M. FLIESS, Some remarks on nonlinear invertibility and dynamic state-feedback, in "Theory and Applications of Nonlinear Control Systems", MTNS-85, C.I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp.115-121.

[9] M. FLIESS, Vers une nouvelle théorie du bouclage dynamique sur la sortie des systèmes non linéaires, in "Analysis and Optimization of Systems", Proc. Conf. Antibes, 1986, A. Bensoussan and J.L. Lions, eds., Lect. Notes Control Inform. Sci., 83, pp.293-299, Springer, Berlin, 1986.

[10] M. FLIESS, A note on the invertibility of nonlinear input-output differential systems, Systems Control Lett., 8, 1986.

[11] M. FLIESS, Esquisses pour une théorie des systèmes non linéaires en temps discret, Proc. Conf. Linear Nonlinear Math. Control Theory, Torino, July 1986. To appear in Rend. Semin. Mat. Univ. Politec. Torino.

[12] M. FLIESS and M. HAZEWINKEL, eds., Algebraic and Geometric Methods in Nonlinear Control Theory, Proc. Conf. Paris, 1985, Reidel, Dordrecht, 1986.

[13] M. FLIESS, M. LAMNABHI and F. LAMNABHI-LAGARRIGUE, An algebraic approach to nonlinear functional expansions, IEEE Trans. Circuits Systems, 30, 1983, pp.554-570.

[14] M. FLIESS and C. REUTENAUER, Théorie de Picard-Vessiot des systèmes réguliers (ou bilinéaires), in "Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Systèmes et le Traitement du Signal", I.D. Landau, ed., t.3, pp.557-581, C.N.R.S., Paris, 1983.

[15] J.-P. GAUTHIER, Structure des systèmes non-linéaires, C.N.R.S., Paris, 1984.

[16] H. GRAUERT and R. REMMERT, Analytische Stellenalgebren, Springer, Berlin, 1971.

[17] M. HASLER and J. NEIRYNCK, Circuits non linéaires, Presses Polytechniques Romandes, Lausanne, 1985.

[18] R.M. HIRSCHORN, Invertibility of nonlinear control systems, SIAM J. Control Optimiz., 17, 1979, pp.289-297.

[19] A. ISIDORI, Nonlinear Control Systems: An Introduction, Lect. Notes Control Inform. Sci., 72, Springer, Berlin, 1985.

[20] A. ISIDORI, Control of nonlinear systems via dynamic state-feedback, in [12], pp.121-145.

[21] A. ISIDORI and C.H. MOOG, On the nonlinear equivalent of the notion of transmission zeros, this volume.

[22] B. JAKUBCZYK, Realization theory for nonlinear systems; three approaches, in [12], pp.3-31.

[23] R.E. KALMAN, P.L. FALB and M.A. ARBIB, Topics in Mathematical System Theory, McGraw-Hill, New York, 1969.

[24] H. KELLER and H. FRITZ, Design of nonlinear observers by a two-step transformation, in [12], pp.89-98.

[25] E.R. KOLCHIN, Differential Algebra and Algebraic Groups, Academic Press, New York, 1973.

[26] A.J. KRENER and W. RESPONDEK, Nonlinear observers with linearizable error dynamics, SIAM J. Control Optimiz., 23, 1985, pp.197-216.

[27] J. LEVINE and P. ROUCHON, Disturbances rejection and integral control of aggregated nonlinear distillation models, in "Analysis and Optimization of Systems", Proc. Conf. Antibes, 1986, A. Bensoussan and J.L. Lions, eds., Lect. Notes Control Inform. Sci., 83, pp.699-714, Springer, Berlin, 1986.

[28] J.-F. POMMARET, Differential Galois Theory, Gordon and Breach, New York, 1983.

[29] J.-F. POMMARET, Géométrie différentielle algébrique et théorie du contrôle, C.R. Acad. Sci. Paris, I-302, 1986, pp.547-550.

[30] J.F. RITT, Differential Algebra, Amer. Math. Soc., New York, 1950.

[31] A. ROBINSON, Local differential algebra, Trans. Amer. Math. Soc., 97, 1960, pp.427-456.

[32] W.J. RUGH, Nonlinear System Theory. The Volterra/Wiener Approach, The Johns Hopkins University Press, Baltimore, 1981.

[33] S. SASTRY and P. VARAIYA, Hierarchical stability and alert state steering control of power systems, IEEE Trans. Circuits Systems, 27, 1980, pp.1102-1112.

[34] A.J. VAN DER SCHAFT, System theoretic description of physical systems, CWI Tract 3, Centrum voor Wiskunde en Informatica, Amsterdam, 1984.

[35] R. SOMMER, Control design for multivariable non-linear time-varying systems, Internat. J. Control, 31, 1980, pp.883-891.

[36] H.J. SUSSMANN, Lie brackets, real analyticity and geometric control, in "Differential Geometric Control Theory", R.W. Brockett, R.S. Millman and H.J. Sussmann, eds., Birkhäuser, Boston, 1983, pp.1-116.

[37] J. TSINIAS and N. KALOUPTSIDIS, Invertibility of nonlinear analytic single-input systems, IEEE Trans. Automat. Control, 28, 1983, pp.931-933.

[38] J.C. WILLEMS, System theoretic models for the analysis of physical systems, Ricerche Automatica, 10, 1979, pp.71-106.

[39] J.C. WILLEMS, From time series to linear systems, Automatica, to appear.

[40] L.A. ZADEH and C.A. DESOER, Linear System Theory, the State Space Approach, McGraw-Hill, New York, 1963.

[41] O. ZARISKI and P. SAMUEL, Commutative Algebra, vol.II, van Nostrand, Princeton, 1960.

[42] M. ZEITZ, Observability canonical (phase-variable) form for nonlinear time-variable systems, Internat. J. Systems Sci., 15, 1984, pp.949-958.

[43] M. ZEITZ, Canonical forms for nonlinear systems, in "Geometric Theory of Nonlinear Control Systems", Proc. Conf. Bierutowice, 1984, B. Jakubczyk, W. Respondek and K. Tchoń, eds., Technical University of Wrocław, 1985, pp.255-278.

ON THE NONLINEAR EQUIVALENT OF THE NOTION OF TRANSMISSION ZEROS

A.Isidori
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Roma, Italy

C.H.Moog
Laboratoire d'Automatique
E.N.S.M. -Unité Associée au C.N.R.S.
Nantes, France

Abstract. *The purpose of this paper is to show that three possible characterizations of the notion of "transmission zero", namely "pole" of the inverse system, zero-output-constrained dynamics and unobservable dynamics under certain state-feedback, which are equivalent for any invertible linear system, may have different analogues for nonlinear input-affine systems. It is also shown that some nonlinear versions of the so-called structure algorithm, proposed by Hirschorn and Singh, may be successfully used in this framework.*

1. Introduction.

The study of the nonlinear analogue of the notion of "transmission zero" has received little attention in the literature, despite of the relatively large amount of contributions in other areas, like disturbance-decoupling, noninteracting control, inversion, model matching, high-gain output feedback, in which at least for linear systems the notion of zero (more specifically, the latter being "left-half-plane" or not) plays an important role. In view of their contributions to the geometric understanding of disturbance decoupling in nonlinear systems, Krener and Isidori proposed in [1] a nonlinear equivalent of the notion of zero based upon the consideration of the "dynamics" associated with that part of the system which becomes unobservable after a disturbance-decoupling state-feedback is set up. This notion is worth being considered in such a setting (i.e. in order to examine the internal behavior of a disturbance-decoupled system, for instance in order to find whether or not this is stable), but as we shall see later on this is not the only aspect one has to deal with. For single-input single-output nonlinear systems, Byrnes and Isidori [2], and Marino [3], showed that the nonlinear analogue of the notion of zero proposed in [1] turns out to be equivalent to the idea of a "dynamics" of an inverse system as well as to that of "dynamics" under high-gain (stabilizing) output feedback. This equivalence is quite appealing because is exactly the one found in any (invertible, even multivariable) linear system. However, recent progresses in the study of multivariable nonlinear systems have shown that the approach to the notion of zero based exclusively, as in [1], on the analysis of a dynamics which is made unobservable under certain state-feedback is not complete, in the sense that some pathologies may occurr. For instance, Isidori has shown in [4] that this dynamics may be affected by addition of integrators on some input channel, fact that looks quite strange if viewed with a linear system background.

In this paper we show that the two other important aspects behind the notion of zero, namely that of a dynamics of the inverse system and that of a dynamics yielding zero output, in a multivariable nonlinear system may have non-equivalent analogues (and even both differing from the former, namely a dynamics which is made unobservable under a certain state-feedback). This shows that there are really three independent ways to approach the nonlinear analogue of the notion of zero and each one has, as we shall remark, its own good reasons to be considered.

In the second half of the paper, we show that the nonlinear versions of the so-called structure algorithm, developed by Hirschorn and later by Singh, under appropriate regularity assumptions lend themselves to very interesting applications in the present framework. More specifically, they provide means to compute the "dynamics" in question as well as to prove several related results.

The system we consider are described by equations of the form:

(1.1a) $\dot{x} = f(x) + g(x)u$

(1.1b) $y = h(x)$

where $x\varepsilon\mathbf{R}^n$, $u\varepsilon\mathbf{R}^m$, $y\varepsilon\mathbf{R}^p$. f and the m columns of g are analytic vector fields, h is an analytic mapping. Throughout most of the paper we shall assume p=m. In the first part of the following section we suppose the reader having some familiarity with the basic principles of controlled invariance for systems of the form (1.1): the appropriate background material can be found e.g. in [5].

2.Three different notions in one.

The purpose of this section is to point out that three important phenomena associated with the notion of "transmission zero", which are equivalent for any invertible linear system, may have in fact different meanings in a more general setting.

The first of these is related to the loss of observability under static state feedback. Let Δ^* denote the largest controlled invariant distribution contained in the kernel of (dh) and suppose Δ^* has constant dimension around x^e. Then, if α and β are such that $f+g\alpha$ and $g\beta$ make Δ^* invariant, then, it is well-known that the feedback $u=\alpha(x)+\beta(x)v$ makes the system maximally unobservable and, in fact, in a neighborhood of x^e, sets of indistinguishable states are integral submanifolds of Δ^*. If, in particular, x^e is an equilibrium point for f (i.e. $f(x^e)=0$), one can always find an α such that $\alpha(x^e)=0$ and this makes $f+g\alpha$ tangent to the integral submanifold of Δ^* passing through x^e. This motivates the following definition.

Definition 2.1. Let x^e be an equilibrium point of f and suppose Δ^* has constant dimension in a neighborhood U of x^e. Let α be such that $\alpha(x^e)=0$ and such that $f+g\alpha$ makes Δ^* invariant. Let N'_{x^e} denote the maximal (on U) integral submanifold of Δ^* through x^e. The vector field on N'_{x^e} defined as $(f+g\alpha)$ restricted to N'_{x^e} is said to be a local dynamics associated with maximal loss of observability (under feedback). ◊

Remark 2.1. In a linear system $(x^e=0)$ N'_{x^e} is exactly V*, the largest controlled invariant subspace contained in the kernel of C. If F is such that A+BF makes V* invariant, then the dynamics associated with the maximal loss of observability is that of the linear mapping defined as A+BF restricted to V*. It is well-known that if the system is square (i.e. same number of inputs and outputs) and invertible, the eigenvalues of this mapping coincide with the transmission zeros.◊

The second aspect we wish to consider consists in the analysis of the dynamics of (1.1) under the constraint that the output $y=h(x)$ be zero for all times. In more precise words, the idea of a dynamics constrained in such a way that $h(x)=0$ can be formulated as follows

Definition 2.2. Let x^e be an equilibrium point of f, let $h(x^e)=0$, suppose there exists a neighborhood U of x^e and a smooth submanifold N''_{x^e} of U containing x^e with the following properties:
(i) N''_{x^e} is contained in $h^{-1}(0)$;
(ii) there exists a feedback $u=\alpha(x)$, defined on U, such that $f^*=f+g\alpha$ is tangent to N''_{x^e};
(iii) N''_{x^e} is maximal (i.e. any submanifold of U through x^e such that (i) and (ii) are true is contained in N''_{x^e}).
The vector field of N''_{x^e} defined as the restriction of f^* to N''_{x^e} is said to be a local zero-output-constrained dynamics.◊

Remark 2.2. In a linear system ($x^e=0$) such an N''_{x^e} exists and coincides with V^*. Thus, the notions of zero-output-constrained dynamics and that of dynamics associated with maximal loss of observability coincide.◊

The third phenomenon we consider is related to the existence of inverse systems. Suppose the system (1.1) has the same number of inputs and outputs and is invertible (in the sense of [6]). Then an inverse system exists and can be described by equations of the form:

(2.1a) $\quad \dot{z} = F(z,y,y^{(1)},\dots,y^{(k)})$
(2.1b) $\quad u = G(z,y,y^{(1)},\dots,y^{(k)})$

where k is a suitable integer. An inverse system should be "generically" able to reproduce the input u of (1.1) on the basis of the knowledge of the output y and of the initial state x^o ("generically" here is to be understood as "for almost all inital states x^o and outputs y of (1.1)"). An inverse system of the form (2.1) is said to be *reduced* if the dimension of its dynamics (i.e. the dimension of z) is minimal over all inverse systems of the form (2.1).

Existence, uniqueness and construction of reduced inverse systems (for systems of the form (1.1)) are not yet fully understood. However, loosely speaking, it seems quite natural to regard the dynamics (2.1a) of a reduced inverse, as a *minimal* set of differential equations required to recover the input function u of (1.1) starting from the knowledge of its output function y and of its initial state x^o.

In a linear system, the dynamics of a reduced inverse has the form:

$$\dot{z} = Fz + G_o y + G_1 y^{(1)} + \dots G_k y^{(k)}$$

and F is a linear mapping whose eigenvalues coincides with the transmission zeros. Thus, again, to look at the spectrum of the dynamics of a reduced inverse is nothing else but an equivalent way to look at the transmission zeros.

The three concepts illustrated so far are no longer equivalent when the system is nonlinear, as we shall see in two simple examples. However, each one has its own interest in control theory. The dynamics associated with maximal loss of observability if found, for instance, as "internal" dynamics of a closed loop system in which a disturbance decoupling problem via static state feedback has been solved. The zero-output-constrained dynamics is found, again as "internal" dynamics of a closed loop system, when high-gain output feedback (or variable structure control) is used to force the output to stay close to zero (see [7]). Finally, the dynamics of a reduced inverse has clearly interest in the construction of inverses.

Intuitively, the difference between dynamics associated with maximal loss of observability and zero-output-constrained dynamics depends on the fact that in the former one looks at invariance (of a distribution) under the vector field $f+g\alpha$ and all the vector fields of $g\beta$, whereas in the latter one looks at invariance (of a submanifold) under the vector field $f+g\alpha$ alone. This is sometimes referred to as the difference between invariance under full control and invariance under singular control. On the other hand, the difference between zero-output-constrained dynamics and the dynamics of a reduced inverse is related to the fact that the output function $y(t)=0$ may be a singular value in the inversion problem. Both these differences appear in the examples that follow.

Example 2.1. Consider the system:

$$\begin{aligned}
\dot{x}_1 &= u_1\\
\dot{x}_2 &= x_4 + x_3 u_1\\
\dot{x}_3 &= \lambda x_3 + x_4\\
\dot{x}_4 &= u_2
\end{aligned}$$

$$\begin{aligned}
y_1 &= x_1\\
y_2 &= x_2
\end{aligned}$$

An easy computation shows that $\Delta^* = 0$ (for all x). Nevertheless, the zero-output-constrained dynamics exists and is given by:

$$\dot{x}_3 = \lambda x_3$$

It may be worth seeing that if $\lambda < 0$, i.e. if the zero-output-constrained dynamics is asymptotically stable, the whole system can be asymptotically stabilized via high-gain output feedback. A compensator doing this job is the one characterized by a transfer function of the form:

$$K(s) = K \begin{bmatrix} (s+z)/s & 0 \\ & \\ 0 & (s+z)/(1+Ts) \end{bmatrix}$$

where $K>0$ is large, $z>0$ and $T>0$ are small.◊

Example 2.2. Consider the system:

$$\dot{x}_1 = x_2 + u_1$$
$$\dot{x}_2 = x_2x_3 + x_4 + x_3u_1$$
$$\dot{x}_3 = u_2$$
$$\dot{x}_4 = x_3$$

$$y_1 = x_1$$
$$y_2 = x_2$$

In this case also we have $\Delta^* = 0$ (for all x); however, the constraint $y=0$ implies now $x=0$. In other words, no nontrivial zero-output-constrained dynamics exists. A reduced inverse is the one given by:

$$\dot{x}_3 = (- x_3 \ - x_3y_1{}^{(2)} + y_2{}^{(2)})/y_1{}^{(1)}$$

$$u_1 = y_1{}^{(1)} - y_2$$
$$u_2 = (- x_3 \ - x_3y_1{}^{(2)} + y_2{}^{(2)})/y_1{}^{(1)}$$

This reduced inverse has a 1-dimensional dynamics. Note also that the value $y_1{}^{(1)}=0$ is a singular value of this dynamics. ◊

3. The zero-output-constrained dynamics.

In this section we shall see that, under appropriate regularity assumptions, the so-called structure algorithm, ideated by Silverman [8] and then generalized by Hirschorn |9] in order to analyze system invertibility comes out in a most natural way when dealing with the zero-output constrained dynamics. The first stage of our study shall consists in the exploitation of some interesting features of the structure algorithm that perhaps are known but, as far as we know, haven't yet been explicitly formulated. To this end we shall revisit the algorithm in question from a slightly different perspective.

Let x^e be an isolated equilibrium state of (1.1) and let $h(x^e)=0$. Suppose the mapping h has constant rank, say s_0, around x^e. Then, locally the set $L_0=h^{-1}(0)$ is a smooth $(n-s_0)$-dimensional submanifold. Choose a coordinate chart (U,ϕ) around x^e in such a way that locally L_0 coincides with a slice of U. More precisely, let $x=(x_0,x_1)$ denote new coordinates around x^e, with $\dim(x_0)=s_0$, chosen in such a way that locally:

$$L_0 = \{x\varepsilon U : x_0=0\}$$

Let f and g be partitioned accordingly:

$$f(x) = \begin{bmatrix} f_0(x_0,x_1) \\ \\ f_1(x_0,x_1) \end{bmatrix} \qquad g(x) = \begin{bmatrix} g_0(x_0,x_1) \\ \\ g_1(x_0,x_1) \end{bmatrix}$$

The constraint $y(t)=0$ for all t clearly implies $x(t)\varepsilon L_0$ and this, in turn, implies $(f(x)+g(x)u)\varepsilon T_xL_0$. Since, at all $x\varepsilon L_0$, $T_xL_0 = \text{span}\{\partial/\partial x_1\}$, in the new coordinates this constraint becomes:

$$(3.1) \qquad f_0(0,x_1) + g_0(0,x_1)u = 0$$

for all x_1 and all u. If this equation can be solved for $u=u(x_1)$, then L_0 (around x^e) is clearly the set we were looking for. The feedback control $u=u(x_1)$ is such as to keep in L_0 the trajectory starting from any point of L_0. The vector field $f^*(x_1) = f_1(0,x_1) + g_1(0,x_1)u(x_1)$ characterizes the zero-output-constrained dynamics.

Consider the case where (3.1) cannot be solved for u and suppose the rank of $g_0(0,x_1)$ is constant, say r_0, around x^e (on L_0).Let $R_0(x_1)$ denote an $(s_0-r_0)xs_0$ matrix of analytic functions, of full rank at all x_1, such that:

$$R_0(x_1)g_0(0,x_1) = 0$$

Then (3.1) clearly implies:

$$\lambda_1(x_1)= R_0(x_1)f_0(0,x_1) = 0$$

along any trajectory that produces zero output. Note that $\lambda_1(x_1)$ is not identically zero (because otherwise (3.1) would be solvable for u). Suppose the mapping λ_1 has constant rank, say s_1, around x^e (note that $\lambda_1(x^e)=0$ because $f(x^e)=0$). Then, locally around x^e, the set $L_1=\lambda_1^{-1}(0)$ is a smooth $(n-s_0-s_1)$-dimensional submanifold. Choose local coordinates on L_0, $x_1 = (x_1',x_2)$ with $\dim(x_1')=s_1$, in such a way that, locally around x^e, $L_1=\{x\varepsilon U\colon x_0=0,x_1'=0\}$ and set:

$$f(x) = \begin{bmatrix} f_0(x_0,x_1',x_2) \\ f_1'(x_0,x_1',x_2) \\ f_2(x_0,x_1',x_2) \end{bmatrix} \qquad g(x) = \begin{bmatrix} g_0(x_0,x_1',x_2) \\ g_1'(x_0,x_1',x_2) \\ g_2(x_0,x_1',x_2) \end{bmatrix}$$

The constraint $y(t)=0$ now implies $x(t)\varepsilon L_1$ and this, in turn implies:

$$0 = \begin{bmatrix} f_0(0,0,x_2) \\ \\ f_1'(0,0,x_2) \end{bmatrix} + \begin{bmatrix} g_0(0,0,x_2) \\ \\ g_1'(0,0,x_2) \end{bmatrix} u$$

This equation is quite similar to (3.1) and from it one can pursue similar iterations.The reader familiar with Silverman-Hirschorn's structure algorithm will easily realize that the iterations in question are essentially the same as those considered in that algorithm. At the k-th stage one considers the equation:

$$0 = F_{k-1}(x_k) + G_{k-1}(x_k)u \tag{3.2}$$

where:

$$F_{k-1}(x_k) = \begin{bmatrix} f_0(0,\ldots,0,x_k) \\ \ldots \\ f'_{k-1}(0,\ldots,0,x_k) \end{bmatrix} \quad G_{k-1}(x_k) = \begin{bmatrix} g_0(0,\ldots,0,x_k) \\ \ldots \\ g'_{k-1}(0,\ldots,0,x_k) \end{bmatrix} \tag{3.3}$$

F_{k-1} and G_{k-1} have $\sigma_{k-1} = s_0+\ldots+s_{k-1}$ rows and x_k denotes coordinates on L_{k-1}, an $(n-\sigma_{k-1})$-dimensional submanifold. If the matrix G_{k-1} has constant rank r_{k-1} around x^e, one finds an $(\sigma_{k-1}-r_{k-1})x\sigma_{k-1}$ matrix $R_{k-1}(x_k)$ such that:

$$R_{k-1}(x_k)G_{k-1}(x_k) = 0$$

and considers the mapping

$$\lambda_k(x_k) = R_{k-1}(x_k)F_{k-1}(x_k) = 0$$

If the set $L_k = \lambda_k^{-1}(0)$ is a smooth $(n-\sigma_{k-1}-s_k)$-dimensional submanifold of L_{k-1}, one chooses new local coordinates and continues.

It is important to remark that the sets $L_0, L_1, \ldots$, the ranks $r_0, r_1, \ldots$, and even the regularity assumptions made at each iteration (constancy of the ranks of $G_{k-1}(x_k)$ around x^e on L_{k-1} and the fact that L_k is a smooth submanifold) depend intrinsically on the system and not on the particular choice of coordinates performed at each stage nor on that of $R_{k-1}(x_k)$. To clarify this point, note first that $G_{k-1}(x_k)$ is simply a submatrix of $g(0,\ldots,0,x_k)$. Moreover, note that the rows of $R_{k-1}(x_k)$ are a basis of the space of row vectors γ solving the linear equation $\gamma G_{k-1}(x_k) = 0$. Thus any other matrix $R'_{k-1}(x_k)$ such that $R'_{k-1}(x_k)G_{k-1}(x_k) = 0$ is related to $R_{k-1}(x_k)$ by an expression of the form:

$$R'_{k-1}(x_k) = T(x_k)R_{k-1}(x_k)$$

where $T(x_k)$ is a nonsingular matrix. From this we see immediately that the set $L_k=\lambda_k^{-1}(0)$ is always the same no matter what $R_{k-1}(x_k)$ is chosen.

If the said regularity assumptions are satisfied, the procedure terminates in at most n iterations. For, if at a certain stage λ_k is not identically zero on L_{k-1} and $L_k=\lambda_k^{-1}(0)$ is a smooth submanifold, then $\dim(L_k) < \dim(L_{k-1})$. If λ_k is identically zero on L_{k-1} we may still set, formally, $L_k=L_{k-1}$. Thus, the procedure terminates after at most $k^*<n$ iterations, where k^* is the least integer such that either one of the following cases occurs:

(i) $L_{k^*} = L_{k^*-1}$ (with $\dim(L_{k^*}) \neq 0$), or
(ii) $L_{k^*} = \{x^e\}$.

In either cases, the equation (3.2) can be solved for $u=u(x_k)$ in the neighborhood of x^e on L_{k^*} ; such a solution may not be unique (unless G_{k^*-1} has rank m). However, in case (ii), only those u which annihilate $G_{k^*-1}(x^e)u$ solve the equation in question.

We synthesize the discussion up to this point in two formal statements.

Definition 3.1. An isolated equilibrium x^e, such that $h(x^e)=0$, is said to be a regular point for the structure algorithm if, for each k>0, the set L_{k-1} is a smooth submanifold of the state space and the matrix $G_{k-1}(x_k)$ has constant rank around x^e for all $x_k \varepsilon L_{k-1}$.◊

Proposition 3.1. Let x^e be a regular point for the structure algorithm. Then locally (around x^e) the set L_{k^*} is the largest submanifold of $h^{-1}(0)$ on which the state of system (1.1) may evolve under suitable control. Any feedback $u=u(x_{k^*})$ solving the equation (3.2) (for $k=k^*$) is such as to keep the state of system (1.1) evolving on L_{k^*}.◊

From what we have seen, it is clear that *the zero-output-constrained dynamics* of (1.1) locally around x^e is described by the vector field of L_{k^*}:

$$f^*(x_{k^*}) = f_{k^*}(0,...,0,x_{k^*}) + g_{k^*}(0,...,0,x_{k^*})u(x_{k^*})$$

where $u(x_{k^*})$ is a solution of (3.2) (for $k=k^*$).

In the previous discussion, we haven't made any specific assumption on the value of the rank r_{k^*}. However, if the system is such that $r_{k^*}=m$, then more properties hold, as specified in the following Lemmas.

Lemma 3.1. Let x^e be a regular point for the structure algorithm. Then the following inequalities hold:
(i) $s_0 \leq m$
(ii) $s_1 \leq s_0 - r_0$
and, for all k such that $2 \leq k \leq k^*$:
(iii) $s_k \leq s_{k-1} - (r_{k-1} - r_{k-2})$
(iv) $r_k \leq r_{k-1} + s_k$
Thus, $r_k \leq m$ for all $k \leq k^*$. If $r_{k^*} = m$ then necessarily in all previous inequalities the equality sign holds.

Proof. The vector h(x) has m rows and this implies (i). The vector $R_0(x_1)f_1(0,x_1)$ has $s_0 - r_0$ rows and this implies (ii). At any $k \geq 2$, one can choose $R_{k-1}(x_k)$ in such a way that its first $\sigma_{k-2}-r_{k-2}$ rows have the form $[\ R_{k-2}((0,x_k))\ \ 0\ ]$. Thus, the first $\sigma_{k-2}-r_{k-2}$ rows of

$R_{k-1}(x_k)F_{k-1}(x_k)$ vanish identically on L_{k-1}. The other rows of this vector are exactly s_{k-1} - $(r_{k-1} - r_{k-2})$ and this implies (iii). Finally, G_k has s_k more rows than G_{k-1} and this implies (iv). The last statements are trivial consequences of (i)-(iv).◊

Lemma 3.2. Let x^e be a regular point for the structure algorithm. Suppose r_{k*} = m. Then at each stage of the structure algorithm one can choose $R_{k-1}(x_k)$ and a partition:

$$R_{k-1}(x_k) = \begin{bmatrix} R'_{k-1}(x_k) \\ \\ R''_{k-1}(x_k) \end{bmatrix}$$

where $R'_{k-1}(x_k)$ has σ_{k-2}-r_{k-2} rows and $R''_{k-1}(x_k)$ has exactly s_k rows. Moreover:

(i) $R'_{k-1}(x_k)F_{k-1}(x_k) = 0$ for all $x_k \varepsilon L_{k-1}$;

(ii) the mapping $c_k(x_k) = R''_{k-1}(x_k)F_{k-1}(x_k)$ has full rank s_k at x^e. ◊

4. The dynamics of a reduced inverse.

In this section we shortly outline how the inversion procedure developed by Singh [10],which consist of a modification of the structure algorithm, can be used in order to construct also a reduced inverse. The procedure in question is defined in the following way. Consider the mapping:

$$S_0(y,x) = h(x) - y$$

and set:

$$\dot{S}_0(y,y^{(1)},x,u) = (\partial S_0/\partial x)f(x) + (\partial S_0/\partial x)g(x)u + (\partial S_0/\partial y)y^{(1)}$$

Note that $\dot{S}_0(y,y^{(1)},x,u)$ is linear in u and it is possible to express it in the form:

$$\dot{S}_0(y,y^{(1)},x,u) = f_0(y,y^{(1)},x) + g_0(y,x)u$$

Let ρ_0 denote the rank of g_0 and set p_0=m, $p_1 = p_0 - \rho_0$. Let $K_0(y,x)$ be a $p_1 x p_0$ matrix of rank p_1 such that:

$$K_0(y,x)\, g_0(y,x) = 0$$

and set:

$$S_1(y,y^{(1)},x) = K_0(y,x)\, f_0(y,y^{(1)},x).$$

At the (k+1)-th stage, consider the mapping $S_k(y,...,y^{(k)},x)$ and set:

$$\dot{S}_k(y,\ldots,y^{(k)},y^{(k+1)},x,u) = (\partial S_k/\partial x)f(x) + (\partial S_k/\partial x)g(x)u + (\partial S_k/\partial y)y^{(1)} + \ldots$$
$$+ (\partial S_k/\partial y^{(k)})y^{(k+1)}$$

Note that $\dot{S}_k(y,\ldots,y^{(k)},y^{(k+1)},x,u)$ is linear in u and it is possible to express it in the form:

$$\dot{S}_k(y,\ldots,y^{(k)},y^{(k+1)},x,u) = f_k(y,\ldots,y^{(k+1)},x) + g_k(y,\ldots,y^{(k)},x)u$$

Set also:

$$F_k = \begin{bmatrix} F_{k-1} \\ \\ f_k \end{bmatrix} \qquad G_k = \begin{bmatrix} G_{k-1} \\ \\ g_k \end{bmatrix}$$

Let ρ_k denote the rank of G_k and set $p_{k+1} = p_k - (\rho_k - \rho_{k-1})$. Let:

$$[\ T_k(y,\ldots,y^{(k)},x) \qquad K_k(y,\ldots,y^{(k)},x)\ ]$$

be a matrix in which $T_k(y,\ldots,y^{(k)},x)$ is $p_{k+1}x(p_0 + \ldots + p_{k-1})$ and $K_k(y,\ldots,y^{(k)},x)$ is $p_{k+1}xp_k$ and has rank p_{k+1}, such that:

$$T_k(y,\ldots,y^{(k)},x)G_{k-1}(y,\ldots,y^{(k)},x) + K_k(y,\ldots,y^{(k)},x)g_k(y,\ldots,y^{(k)},x) = 0$$

and set:

$$S_{k+1}(y,\ldots,y^{(k+1)},x) = T_k(y,\ldots,y^{(k)},x)F_{k-1}(y,\ldots,y^{(k)},x) + K_k(y,\ldots,y^{(k)},x)f_k(y,\ldots,y^{(k+1)},x)$$

If at some k* the matrix G_{k*} has rank $r_{k*} = m$, then it is easy to conclude that the equation:

$$F_{k*}(y,\ldots,y^{(k*)},x) + G_{k*}(y,\ldots,y^{(k*-1)},x)u = 0 \tag{4.1}$$

is solvable in u (see [10]). Moreover, using arguments which are more or less similar to those used in order to prove Lemma 3.1, it is also possible to show that the jacobian matrix:

$$(\partial/\partial x) \begin{bmatrix} S_0(y,x) \\ \ldots \\ S_{k*-1}(y,\ldots,y^{(k*-1)},x) \end{bmatrix}$$

has rank $\mu = (p_0 + \ldots + p_{k*-1})$ (namely, equal to the number of its rows). Thus form the equation:

$$0 = \begin{bmatrix} S_0(y,x) \\ \ldots \\ S_{k^*-1}(y,\ldots,y^{(k^*-1)},x) \end{bmatrix}$$

one can recover μ components of x, expressed as functions of $y,\ldots,y^{(k^*-1)}$ and of the remaining n-μ components, noted z. From that and (4.1) one obtains a reduced inverse system, in the form:

$$\begin{aligned} \dot{z} &= F(z,y,y^{(1)},\ldots,y^{(k^*-1)}) \\ u &= G(z,y,y^{(1)},\ldots,y^{(k^*)}) \end{aligned}$$

Remark 4.1. The previous construction, essentially suggested by the work of Singh, shows how it is possible to construct a "reduced" inverse system. This inverse is defined for almost all output functions. If $y=y^{(1)}=\ldots=y^{(k^*)}=0$ and $x=x^e$ is a point of regularity for G_{k^*}, then the dynamics of this inverse, when driven by y=0, reduces to the zero-output-constrained dynamics discussed before. It is important to stress that this is no longer true when the said regularity assumption fails to hold, has shown by Example 2.2. ◊

5. Further remarks.

In the previous two sections we have shown how it is possible - under suitable regularity assumptions - to calculate zero-output constrained dynamics and inverse system dynamics. This, together with known methods of computing the distribution Δ^* (and a feedback making the latter invariant) completes in some sense the range of procedures needed to evaluate the three different types of dynamics described at the beginning. Moreover, the constructions outlined so far enable us to draw some interesting conclusions.The first of these is expressed in the following statement.

Lemma 5.1. Suppose x^e is a regular point for the structure algorithm and $r_{k^*}=m$. Then, at each $x \varepsilon L_{k^*}$, $\Delta^*(x)$ is a subspace of $T_x L_{k^*}$. As a consequence, the dimension of the zero-output-constrained dynamics is always larger than or equal to that of the dynamics associated to the maximal loss of observability.

Proof (sketch of). Recall that the annihilator Ω_* of Δ^* can be computed (via the so-called maximal controlled invariant distribution algorithm [11]), by means of a sequence Ω_k of codistributions approaching Ω_* in a finite ($\leq$ n) number of stages. Then, show, by induction, that the differentials of the entries of λ_k belong to Ω_k . ◊

In the previous section we have seen that, under suitable regularity and invertibility assumptions, the dynamics associated with the inversion problem reduces - when y is identically zero - to the zero-output-constrained dynamics. However, if y=0 is a singular value for the inversion problem, then the latter has a dimension which is possibly smaller than that found in a reduced inverse. This appears clearly from the comparison of Hrischorn's and Singh's algorithms (see also Example 2.2).

One may wish to examine whether or not in some cases the three types of dynamics coincides. One of these is described in the following statement.

Lemma 5.2. Suppose the decoupling matrix A(x) (see [11] for the notations) is nonsingular at x^e. Then dynamics associated with maximal loss of observability, zero-output-constrained dynamics and dynamics of reduced inverse (the latter being driven by y=0) coincide.

Proof (sketch of). If $A(x^e)$ is nonsingular, the noninteracting control problem is solvable around x^e. In this case the three type of constructions yielding the dynamics in question clearly coincide. ◊

Aknowledgement.

This paper was completed when C.H.Moog was visiting the Dipartimento di Informatica e Sistemistica dell'Università di Roma "La Sapienza", under the sponsorship of the Commission of the European Communities ("Stimulation Action").

References.

[1] A.J.Krener, A.Isidori: Nonlinear zero distributions, *19th IEEE Conf. Decision and Control*, (1980).

[2] C.Byrnes, A.Isidori: A frequency domain philosophy for nonlinear systems, with applications to stabilization and adaptive control, *23rd IEEE Conf. Decision and Control*, (1984), pp. 1569-1573.

[3] R.Marino: High-gain feedback in non-linear control systems, *Int. J. Control*, 42 (1985), pp. 1369-85.

[4] A.Isidori: Control of Nonlinear Systems via dynamic state-feedback, *Algebraic and Geometric Methos in Nonlinear Control Theory* (M.Hazewinkel and M.Fliess, eds.), Reidel (1986).

[5] A.Isidori: Nonlinear control systems: an introduction, *Lecture Notes in Control and Information Sciences*, Vol. 72, Springer Verlag (1985).

[6] M.Fliess: A note on the invertibility of nonlinear input-output differential systems, *Systems and Control Lett.*, (1986), to appear.

[7] C.Byrnes, A.Isidori: Global feedback stabilization of nonlinear systems, *24th IEEE Conf. Decision and Control*, (1985).

[8] L.M.Silverman: Inversion of multivariable linear systems, *IEEE Trans. Automatic Control*, AC-14 (1969), pp. 270-276.

[9] R.M.Hirschorn: Invertibility of multivariable nonlinear control systems, *IEEE Trans. Automatic Control* , AC-24 (1979), pp. 855-865.

[10] S.N.Singh: A modified algorithm for invertibility in nonlinear systems, *IEEE Trans. Automatic Control* , AC-26 (1981), pp. 595-598.

[11] A.J.Krener, A.Isidori, C.Gori-Giorgi, S.Monaco: Nonlinear decoupling via feedback: a differential-geometric approach, *IEEE Trans. Automatic Control* , AC-26 (1981), pp.331-345.

[12] J.Descusse, C.H.Moog: Decoupling with dynamic compensation for strong invertible affine nonlinear systems, *Int. J. of Control* , 42 (1985), pp. 1387-1398.

REALIZATIONS OF RECIPROCAL PROCESSES

Arthur J. Krener*
Department of Mathematics and
the Institute of Theoretical Dynamics
University of California
Davis, CA 95616
USA

Abstract We review the concept of a reciprocal process and show that a stationary Gaussian reciprocal process, which satisfies a certain technical assumption, can be realized by a linear stochastic differential equation with independent initial condition.

*Research supported in part by NSF under DMS-8601635.

1. Reciprocal Processes.

Suppose $x(t)$ is an n vector valued stochastic process where t ranges over a subset of the reals or the integers. The process $x(t)$ is called reciprocal (or quasi-Markov) if given any $\tau_0 \le \tau_1$ the values of the process within $[\tau_0, \tau_1]$ are independent of the values of the process outside of $[\tau_0, \tau_1]$ conditioned on $x(\tau_0)$ and $x(\tau_1)$.

In particular a Gaussian process $x(t)$ is reciprocal if for any $t_1, \ldots, t_k \le \tau_0 \le s_1, \ldots, s_m \le \tau_1 \le t_{k+1}, \ldots, t_l$ we have for $i = 1, \ldots, l$

$$\begin{aligned} &E(x(t_i)|x(\tau_0), x(\tau_1)) \\ (1.1a)\qquad &= E(x(t_i)|x(\tau_0), x(\tau_1), x(s_1), \ldots, x(s_m)) \end{aligned}$$

and for $j = 1, \ldots, m$

$$\begin{aligned} &E(x(s_j)|x(\tau_0),x(\tau_1)) \\ (1.1b)\quad & \\ &= E(x(s_j)|x(\tau_0),x(\tau_1),x(t_1),...x(t_l)). \end{aligned}$$

This definition was formulated by Serge Bernstein [1] as a generalization of the concept of a Markov process. Recall that a process x(t) is Markov if for any τ_0 the values of the process to the left of τ_0 are independent of the values to the right conditioned on $x(\tau_0)$. A Gaussian process x(t) is Markov if for any $t_1,...,t_k \le \tau_0 \le s_1,...,s_m$ we have for $i = 1,...,k$

$$(1.2a)\quad E(x(t_i)|x(\tau_0)) = E(x(t_i)|x(\tau_0),x(s_1),...x(s_m))$$

and for $j = 1,...m$

$$(1.2b)\quad E(x(s_j)|x(\tau_0)) = E(x(s_0)|x(\tau_0),x(t_1),...,x(t_k)).$$

It is easy to see that Markov processes are reciprocal but the converse is not true. Throughout this paper we will restrict our attention to zero-mean Gaussian processes, and often we shall further restrict our attention to stationary zero mean Gaussian processes. Because of the zero-mean Gaussian assumption, all the probablistic information about the process x(t) is contained in its covariance

$$(1.3)\quad R_x(t,s) = E(x(t)x^*(s))$$

where * denotes transpose. This is a nxn matrix valued function.

A process x(t) is nonsingular of order one if $R_x(\tau_0,\tau_0)$ is nonsingular for every τ_0. Such a process is Markov iff its covariance satisfies

$$(1.4) \quad R_x(t,s) = R_x(t,\tau_0)R_x^{-1}(\tau_0,\tau_0)R_x(\tau_0,s)$$

for any $t \le \tau_0 \le s$.

Let τ denote the ordered k-tuple $(\tau_0,\ldots,\tau_{k-1})$ where $\tau_0 \le \tau_1 \le\ldots\le \tau_k$. Define an k·n dimensional random vector $X(\tau)$ by

$$(1.5) \quad X(\tau) = \begin{pmatrix} x(\tau_0) \\ \cdot \\ \cdot \\ \cdot \\ x(\tau_{k-1}) \end{pmatrix}$$

A process x(t) is nonsingular of order k if for any $\tau_0 < \tau_1 <\ldots< \tau_{k-1}$ the covariance of the random vector $X(\tau)$ is positive definite.

Suppose x(t) is nonsingular of order 2, then x(t) is reciprocal iff its covariance satisfies

$$(1.6) \quad R_x(t,s) =$$

$$[R_x(t,\tau_0) \;\; R_x(t,\tau_1)] \begin{bmatrix} R_x(\tau_0,\tau_0) & R_x(\tau_0,\tau_1) \\ R_x(\tau_1,\tau_0) & R_x(\tau_1,\tau_1) \end{bmatrix}^{-1} \begin{bmatrix} R_x(\tau_0,s) \\ R_x(\tau_1,s) \end{bmatrix}$$

for all $t \le \tau_0 \le s \le \tau_1$ and for all $\tau_0 \le s \le \tau_1 \le t$.

If $\tau = (\tau_0,\tau_1)$ and $\sigma = (\sigma_0,\sigma_1)$ then we can define a partial ordering by $\tau \ge \sigma$ if $\tau_0 \le \sigma_0 \le \sigma_1 \le \tau_1$. Let $X(\tau)$ be defined by (1.5). The process x(t) is reciprocal if the process $X(\tau)$ is Markov relative to this partial ordering.

Mehr and McFadden [2] noted that reciprocal processes are <u>conditionally</u> <u>Markov</u>. If we condition an $x(\tau_1)$ then the conditional process is Markov to the left of τ_1 and if we condition on $x(\tau_0)$ the conditional process is Markov to the right of τ_0.

2. Examples. We review the classification of all one dimensional, stationary, Gaussian, reciprocal processes where $t \in \mathbb{R}$. This is due to Jamison [3], Chay [4] and Carmichael-Masse-Theodorescu [5]. Essentially there are only six families of such processes,

1a. Ornstein Uhlenbeck Processes
1b. Cosh Processes
1c. Sinh Processes
2. Slepian Processes
3a. Cosine Processes
3b. Shifted Cosine Processes

The Ornstein Uhlenbeck processes are the only ones that are Markov. They have covariance $R_x(t,s) = R_x(t-s)$ given by

$$(2.1a) \qquad R_x(t) = e^{-|At|} R_x(0)$$

Such processes have an infinite lifetime, i.e. they can be defined for all $t \in \mathbb{R}$. Of course one can restrict t to lie in some proper subset of $\mathbb{R}$. If $A \doteq 0$ then the process is constant with respect to t and hence singular of order two. If $R_x(0) = 0$ then the process is identically zero and singular of order one. Otherwise the process is nonsingular of every order $k \geq 0$.

The remaining one dimensional stationary Gaussian reciprocal processes are not Markov. A Cosh process has covariance

$$(2.1b)\qquad R_x(t) = \frac{\cosh A(T/2-t)}{\cosh AT/2} \; R_x(0)$$

where A, T > 0. A Cosh process has a finite lifetime because any covariance must satisfy the Cauchy-Schwartz inequality, $|R_x(t)| \le R_x(0)$. But $R_x(t)$ given by (2.1b) violates this for $t > T$. Since $R_x(T) = R_x(0)$, it is a cyclic process, $x(0) = x(T)$ a.s.

A Sinh process has covariance

$$(2.1c)\qquad R_x(t) = \frac{\sinh A(T/2-t)}{\sinh AT/2} \; R_x(0)$$

where A, T > 0. It also has a finite lifetime of length at most T. Since $R_x(T) = -R_x(0)$, it is an anticyclic process $x(0) = -x(T)$ a.s.

A Slepian process has covariance of the form

$$(2.2)\qquad R_x(t) = (1-2t/T)\; R_x(0)$$

where T > 0. Again it has a finite lifetime of length at most T. It also is anticyclic, $x(0) = -x(T)$ a.s.

A Cosine process has covariance

$$(2.3a)\qquad R_x(t) = (\cos At)\; R_x(0).$$

It has an infinite lifetime.

Since $R_x(t)$ is periodic with period $2T = 2\pi/A$, the process is also periodic $x(t) = x(t+2T)$ a.s. Furthermore, it is antiperiodic $x(t) = -x(t+T)$ a.s.

A <u>Shifted Cosine process</u> has covariance of the form

$$R_x(t) = \frac{\cos A(t+t_0)}{\cos At_0} R_x(0) \tag{2.3b}$$

where $0 < t_0 < \pi/2A$. It has a finite maximum lifetime $T = \pi/A-2t_0$ and it is anticyclic, $x(0) = x(-T)$ a.s.

The Cosh, Sinh, Slepian, Cosine and Shifted Cosine processes are all nonsingular of order two on any interval of length less than T. Since in each case, $x(t) = \pm\, x(t+T)$, they are singular of order two on intervals of length T. All of the above processes except for the Cosine processes are nonsingular of arbitrary order on any interval of length less than T. A Cosine process is singular of order 3. This means that the behavior of such a process is completely determined by its values $x(\tau_0)$ and $x(\tau_1)$ at two times where $\tau_1-\tau_0$ is not an integer multiple of T.

3. <u>Realization Theory</u>. It is well known [6] that if R(t) is a continuous covariance of stationary Gauss Markov process then R(t) is C^∞ and it satisfies a first order linear differential equation

$$\dot{R}(t) = AR(t) \tag{3.1}$$

for $t \geq 0$. Furthermore if B is an nxn matrix such that $BB^* = -(\dot{R}(0)+\dot{R}^*(0))$ then the process x(t) defined for $t \geq 0$ by the stochastic differential equation

$$dx = A\,x\,dt + Bdw \tag{3.2a}$$

$$x(0) \simeq N(0,R(0)), \tag{3.2b}$$

where w is standard m dimensional Wiener process independent of x(0), has covariance R(t). Note $\dot{R}(0) = \dot{R}(0^+)$.

In this section we shall show that certain continuous stationary Gaussian reciprocal covariances can be realized by second order linear stochastic differential equations driven by white Gaussian noise with independent initial conditions. This partially confirms a conjecture of ours made in [7].

The first step is to show that a continuous stationary Gaussian covariance $R(t,s) = R(t-s)$ must be C^∞. We did this in [7] but we shall repeat the proof here. We assume R(t) is defined for $|t| \leq T$ and is nonsingular of order two for $|t| < T$.

For a stationary reciprocal covariance, (1.6) becomes

$$R(t-s) = \tag{3.3}$$

$$[R(t-\tau_0)\ R(t-\tau_1)] \begin{bmatrix} R(0) & R(\tau_0-\tau_1) \\ R(\tau_1-\tau_0) & R(0) \end{bmatrix}^{-1} \begin{bmatrix} R(\tau_0-s) \\ R(\tau_1-s) \end{bmatrix}$$

and this holds for $\tau_0 \le s \le \tau_1$ and either $t \le \tau_0$ or $t \ge \tau_1$. Assume that $\tau_1 - \tau_0 < T$. If we integrate with respect to t over $[t_0 - \delta, \tau_0]$ where $0 < \delta < T - \tau_1 + \tau_0$, we obtain

$$\int_{\tau_0-\delta}^{\tau_0} R(t-s)dt = \int_{\tau_0-\delta-s}^{\tau_0-s} R(t)dt = [\delta I + o(\delta) \quad o(\delta)] \begin{bmatrix} R(\tau_0-s) \\ R(\tau_1-s) \end{bmatrix}$$

where $o(\delta)/\delta \to 0$ as $\delta \to 0$. If we integrate (3.1) with respect to t over $[\tau_1, \tau_1+\delta]$ we obtain

$$\int_{\tau_1}^{\tau_1+\delta} R(t-s)dt = \int_{\tau_1-s}^{\tau_1+\delta-s} R(t)dt = [o(\delta) \quad \delta I + o(\delta)] \begin{bmatrix} R(\tau_0-s) \\ R(\tau_1-s) \end{bmatrix}$$

Putting these together we have

$$(3.4) \quad \begin{bmatrix} \displaystyle\int_{\tau_0-\delta-s}^{\tau_0-s} R(t)dt \\ \displaystyle\int_{\tau_1-s}^{\tau_1+\delta-s} R(t)dt \end{bmatrix} \begin{bmatrix} \delta I + o(\delta) & o(\delta) \\ o(\delta) & \delta I + o(\delta) \end{bmatrix} \begin{bmatrix} R(\tau_0-s) \\ R(\tau_1-s) \end{bmatrix}$$

Since R(t) is C^0, the left side of (3.4) in C^1 in s for $s \epsilon [\tau_0, \tau_1]$. By this we mean the left (right) derivative exists and is continuous at $\tau_0 (\tau_1)$. For sufficiently small $\delta > 0$ the first matrix on the right is invertible hence we conclude that $R(\tau_0-s)$ and $R(\tau_1-s)$ are C^1 in $s \epsilon [\tau_0, \tau_1]$. Since τ_0 and τ_1 are arbitrary except that $0 < \tau_1 - \tau_0 < T$ we conclude R(t) is C^1 on [0,T]. By repeating the argument we conclude R(t) is C^∞ on [0,T). Since $R(-t) = R^*(t)$ it follows that R(t) is also C^∞ on (-T,0]. By continuity $R(0^+) = R(0^-)$. The left and right derivatives need not agree at 0, instead $-\dot{R}(0^-) = \dot{R}^*(0^+)$, $\ddot{R}(0^-) = \ddot{R}^*(0^+)$, etc. Henceforth we shall take $\dot{R}(0)$ as $\dot{R}(0^+)$, $\ddot{R}(0)$ as $\ddot{R}(0^+)$, etc.

The next step is to show that R(t) satisfies two second order matrix differential equations. Let $\tau_0 = s - \sigma$ and $\tau_1 = s + \sigma$ for $\sigma > 0$ then (3.3) becomes

$$(3.5a) \qquad R(t-s) = [R(t-s+\sigma) \ R(t-s-\sigma)] \begin{bmatrix} H_1(\sigma) \\ H_2(\sigma) \end{bmatrix}$$

where $H_1(\sigma)$ and $H_2(\sigma)$ are determined for $\sigma > 0$ by

$$(3.5b) \qquad \begin{bmatrix} R(0) & R^*(2\sigma) \\ R(2\sigma) & R(0) \end{bmatrix} \begin{bmatrix} H_1(\sigma) \\ H_2(\sigma) \end{bmatrix} = \begin{bmatrix} R^*(\sigma) \\ R(\sigma) \end{bmatrix}$$

since R(t) is assumed to be the covariance of a process which is nonsingular of order two. For convenience, we make a change of coordinates, $x_{new}(t) = R^{-1/2}(0)x_{old}(t)$ and thereby normalize $R(0) = I$. We would like to study the limit of $H_1(\sigma)$ and $H_2(\sigma)$ and their derivatives as $\sigma \to 0$. From (3.5b) we obtain for $\sigma > 0$

$$(3.6a) \qquad H_1(\sigma) = G^{-1}(\sigma) \ F(\sigma)$$

$$(3.6b) \qquad H_2(\sigma) = R(\sigma) - R(2\sigma) \ H_1(\sigma)$$

where $F(\sigma)$ and $G(\sigma)$ are C^∞ for $\sigma \geq 0$ and given by

$$(3.6b) \qquad F(\sigma) = R^*(\sigma) - R^*(2\sigma)\, R(\sigma)$$

$$(3.6c) \qquad G(\sigma) = I - R^*(2\sigma)\, R(2\sigma).$$

Since $F(0) = G(0) = 0$, (3.6a) is indeterminate at $\sigma = 0$. We define

$$\overline{F}(\sigma) = \begin{cases} F(\sigma)/\sigma & \sigma > 0 \\ \dot{F}(0) & \sigma = 0 \end{cases}$$

$$\overline{G}(\sigma) = \begin{cases} G(\sigma)/\sigma & \sigma > 0 \\ \dot{G}(0) & \sigma = 0 \ . \end{cases}$$

By repeated application of L'Hopital's rule it is easy to verify that $\overline{F}(\sigma)$ and $\overline{G}(\sigma)$ are C^∞ for $\sigma \geq 0$.

Henceforth we shall invoke the assumption that

$$(3.7) \qquad \overline{G}(0) = \dot{G}(0) = -2(\dot{R}(0) + \dot{R}^*(0))$$

is invertible. Rewriting (3.6a) we have for $\sigma > 0$

$$H_1(\sigma) = (G(\sigma)/\sigma)^{-1}(F(\sigma)/\sigma) = \overline{G}(\sigma)/\overline{F}(\sigma)$$

and hence $H_1(\sigma)$ has a C^∞ extension to $\sigma \geq 0$. Equation (3.6b) defines a C^∞ extension of $H_2(\sigma)$ to $\sigma \geq 0$. By straightforward differentiation of (3.6) we obtain

(3.8a) $H_1(0) = H_2(0) = \frac{1}{2} I$

(3.8b) $\dot{H}_1(0) = -\dot{H}_2(0) = \frac{1}{4} (\dot{R}(0)+\dot{R}^*(0))^{-1}(\ddot{R}(0)-\ddot{R}^*(0))$

(3.8c) $\ddot{H}_1(0) + \ddot{H}_2(0) = -\ddot{R}(0) -4\dot{R}(0)\dot{H}_1(0).$

We return to (3.5a) at s = 0 and differentiate twice with respect to σ at σ = 0 to obtain

$$\begin{aligned} 0 = {} & \ddot{R}(t)(H_1(0)+H_2(0)) \\ & + 2\dot{R}(t)(\dot{H}_1(0)-\dot{H}_2(0)) \\ & + R(t)(\ddot{H}_1(0)+\ddot{H}_2(0)) \end{aligned}$$

By utilizing (3.8) we obtain

(3.9a) $\ddot{R}(t) = -2\dot{R}(t)M^* + 2R(t)\ N^*$

where

(3.10a) $-2M^* = (\dot{R}(0) + \dot{R}^*(0))^{-1}\ (\ddot{R}(0) - \ddot{R}^*(0))$

(3.10b) $2N^* = \ddot{R}(0) + 2\dot{R}(0)M^*$

Equation (3.5a) is valid both for $t \geq s + \sigma$ and for $t \leq s + \sigma$. Since $s = \sigma = 0$ this implies that (3.9a) is valid for $t \epsilon [0,T]$ and for $t \epsilon (-T,0]$. The covariance $R(t) = R^*(-t)$ so $\dot{R}(t) = -\dot{R}^*(-t)$ and $\ddot{R}(t) = \ddot{R}^*(-t)$.

We transpose (3.9a) and substitute to obtain

$$(3.9b) \quad \ddot{R}(t) = 2M\dot{R}(t) + 2NR(t).$$

By adding and subtracting (3.9a,b) we obtain the following.

Theorem 1 Suppose R(t) is the continuous covariance of a stationary Gaussian reciprocal process defined on [0,T] and (WLOG) R(0) = I. Suppose that $\dot{R}(0) + \dot{R}^*(0)$ is invertible. Then R(t) is C^∞ on [0,T) and satisfies the differential equation

$$(3.11a) \quad \ddot{R}(t) = M\dot{R}(t) - \dot{R}(t)M^* + NR(t) + R(t)N^*$$

and the side constraint

$$(3.11b) \quad 0 = M\dot{R}(t) + \dot{R}(t)M^* + NR(t) - R(t)N^*$$

where M,N are defined by (3.10).

We now construct a process y(t) which realizes the stationary Gaussian reciprocal covariance R(t), under the assumption that $\dot{R}(0) + \dot{R}^*(0)$ is invertible. By the Cauchy-Schwartz inequality $R(0) - R^*(\sigma)R(\sigma)$ is monotone increasing for small $\sigma > 0$ hence $\dot{R}(0) + \dot{R}^*(0)$ is nonpositive definite. Since it is assumed to be invertible, it is negative definite and there exists an invertible nxn matrix B, such that

$$(3.12) \qquad B_1B_1{*} = -(\dot{R}(0) + \dot{R}{*}(0))$$

Let N and M be as above (3.10). Define a nxn symmetric matrix $\pi(t)$ as the solution of the matrix Riccati differential equation

$$(3.13a) \qquad \frac{d\pi}{dt} = 2N\dot{R}{*}(0) + 2\dot{R}(0)N{*} + 2M\pi(t) + 2\pi(t)M{*} + (\ddot{R}(0) + \pi(t))(B_1B_1{*})^{-1}(\ddot{R}{*}(0) + \pi(t))$$

$$(3.13b) \qquad \pi(0) = \dot{R}(0)\dot{R}{*}(0)$$

Let $B_2(t)$ be an nxn matrix defined by

$$(3.14) \qquad B_2(t) = -(\ddot{R}(0) + \pi(t))B_1{*}^{-1}$$

Finally we define a 2n dimensional process $x(t) = (x_1(t), x_2(t))$ by the stochastic differential equation

$$(3.15a) \qquad \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix} = \begin{pmatrix} 0 & I \\ 2N & 2M \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} dt + \begin{pmatrix} B_1 \\ B_2(t) \end{pmatrix} dw$$

$$(3.15b) \qquad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} R(0) \\ \dot{R}(0) \end{pmatrix} v \qquad\qquad v \simeq N(0,I)$$

where w is an n dimensional standard Wiener process independent of v.

Let

$$P(t,s) = \begin{bmatrix} P_{11}(t,s) & P_{12}(t,s) \\ P_{21}(t,s) & P_{22}(t,s) \end{bmatrix}$$

then P(t,s) satisfies for $T > t \geq s \geq 0$

$$\text{(3.16b)} \qquad \frac{\partial P}{\partial t}(t,s) = AP(t,s)$$

$$\text{(3.16b)} \qquad \frac{d}{dt} P(t,t) = AP(t,t) + P(t,t)A^* + B(t)B^*(t)$$

and

$$\text{(3.16c)} \qquad P(0,0) = \begin{pmatrix} R(0) & \dot{R}^*(0) \\ \dot{R}(0) & \dot{R}(0)\dot{R}^*(0) \end{pmatrix}$$

where

$$A = \begin{bmatrix} 0 & I \\ 2N & 2M \end{bmatrix} \qquad B(t) = \begin{bmatrix} B_1 \\ B_2(t) \end{bmatrix}$$

It is straightforward to verify that

$$\text{(3.17a)} \qquad P(t,t) = \begin{bmatrix} R(0) & \dot{R}^*(0) \\ \dot{R}(0) & \pi(t) \end{bmatrix}$$

satisfies (3.16b,c). From (3.16a) we obtain for $T > t \geq s \geq 0$,

$$(3.18b) \qquad \frac{\partial P_{11}}{\partial t}(t,s) = P_{21}(t,s)$$

$$(3.18c) \qquad \frac{\partial P_{21}}{\partial t}(t,s) = 2MP_{21}(t,s) + 2NP_{11}(t,s)$$

hence $P_{11}(t,s) = R(t-s)$ and $P_{21}(t,s) = \dot{R}(t-s)$. We have proved the following.

Theorem 2 Suppose $R(t)$ is the nxn continuous covariance of a stationary Gaussian reciprocal process defined on $[0,T]$ and (WLOG)$R(0) = I$. Suppose $\dot{R}(0) + \dot{R}^*(0)$ is invertible. Then $R(t)$ can be realized by a first order 2n dimensional linear stochastic differential equation (3.15a) driven by n dimensional white Gaussian noise with an independent initial condition (3.15b).

4. Conclusion In Sections One and Two we defined and gave examples of reciprocal processes. In Section Three we showed how certain stationary Gaussian reciprocal processes can be realized via stochastic differential equations. The condition that we required was that $\dot{R}(0) + \dot{R}^*(0)$ be invertible, but we believe that this technical condition can be dropped. We hope to prove this in the near future.

REFERENCES

1. Bernstein, S., Sur les laisons entre les grandeurs aleatoires, Proc. of Int. Cong. of Math., Zurich (1932) 288-309.

2. Mehr, C. B. and McFadden, J. A., Certain properties of Gaussian processes and their first passage times, J. Roy. Stat. Soc. B27 (1965) 505-522.

3. Jamison, D., Reciprocal processes: the stationary Gaussian case, Ann. Math. Stat. 41 (1970) 1624-1630.

4. Chay, S. C., On quasi-Markov random fields, J. Multivar. Anal. 2 (1972) 14-76.

5. Carmichael, J-P., Masse, T-C. and Theodorescu, R., Processus gaussiens stationaires reciproques sur un intervalle, C. R. Acad. Sc. Paris, Serie I 295 (1982) 291-293.

6. Doob, J. L., Stochastic Processes (Wiley, N.Y. 1953).

7. Krener, A. J., Reciprocal Processes and the Stochastic Realization Problem for Acausal Systems, in Modeling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, 197-211.

Modelling And Analysis of Distributed Systems : A Net Theoretic Approach

Sadatoshi Kumagai
Department of Electronic Engineering
Osaka University, suita
Osaka 565 JAPAN

0. Introduction

Industries today are confronted with an urgent need to deal with reduced life cycles of engineering products, which bring about trends toward total factory automation (FA) which includes CAD and production control. This implies greater necessity for flexible and maintainable control software tools for systems categorized as discrete event systems.

Methods for describing sequential control systems such as relay ladder diagram or procedural languages can not adequately adress all aspects of such systems. Especially, concurrent evolutions in finite resource sharing systems and resulting problems such as synchronization, starvation, and deadlock can not properly analized by these ad hoc techniques used to date. Drastic changes in system archtecture towards distributed configurations composed of separate sequential processing systems as seen in wide range of applications of computer embedded control systems exacerbate the difficulties. To resolve these difficulties, a formal approach based on solid theories is inevitable, and this should start with a sound mathematical model which can be used from specifications to analysis of the real time control. Of course there are previously a variety of theories of concurrent systems. Algebraic, Net-theoretic, and Temporal logic or Axiomatic approaches are fundamental ones among others and they reflect different perspectives of the phenomena of concurrent systems. The relations among them are examined in [1].

In this paper, we introduce Petri nets as a modelling tool of distributed concurrent systems in general and show some analytical properties obtained by a net-theoretic approach. The advantages of a net-theoretic approach exist in its structure preserving modelling ability and in theoretic accomplishments accumlated in the past two decades. A brief introduction of Petri nets and their properties related to control problems are presented in the next section. In contrast to Algebraic approach where recurrent substitutions and products of operators are naturally introduced , net theory lacks in operator viewpoint of input-output notion : The composition of nets has not been investigated in terms of the concatenation of the operations of their subnets. In Section 2, we propose a way of composition of nets which can be used effectively for modelling distributed systems in a hierarchial fashion. In Section 3, conditions for consistency and deadlock-free property of a composite net are investigated in terms of subnets and their interconnections. Net theory is by no means matured yet but there already exist several attempts to apply the formal net-specification methodology in real time systems. Some notable industrial applications of net theory are introduced in Section 4.

1. Petri nets:Definitions and Properties

The structure of a Petri net can be defined as a directed bipertite graph with two disjoint sets of nodes P and T, called a set of places (symbol:o) and a set of transitions (symbol:|),respectively. A Petri net is simply denoted by (P,T). Let U and Q are subsets of P and Q, respectively. A subnet denoted by (U,Q) is a net such that the connections between U and Q are defined as those in the original net (P,T). Places can be seen as conditions and transitions can be seen as events in various activity level of discrete event systems. Marked graphs are a subclass of Petri nets, where each place has exactly one incoming edge and exactly one outgoing edge. Marked graphs can express the concurrent evolution but cannot express the conflict of system being modelled. On the other hand, State machines are a subclass of Petri nets, where each transition has exactly one incoming edge and exactly one outgoing edge. State machines can express the conflict but not the concurrency. Let m and n denote the number of places and transitions in a Petri net, respectively. To each place p, we associate an nonnegative integer M(p), called a number of tokens on p. Marking $M \in N^m$ is defined as an nonnegative integer vector whose component M(p) equals to the number of tokens of place p. For a subnet U of P, M(U) denotes a submarking vector defined on U.${}^\bullet U$ denotes the set of all transitions t such that there exists an edge e:t->p, $p \in U$. ${}^\bullet U$ is called the set of input transitions of U. Similarly, $U^\bullet$ denotes the set of all transitions t such that there exists an edge e:p->t, $p \in U^\bullet$ and is called the set of output transitions of U. For a subset Q of T, the set of input places ${}^\bullet Q$ and the set of output places $Q^\bullet$ of Q are similarly defined. Now the dynamic behavior of Petri nets is stipulated by the following simple firing axiom. For a transition t, t is said to be firable at M if M(p)>0 for each $p \in t$. A firing of firable transition t at M is said to be legal and consists of the transformation of M to M' such that

$$\begin{array}{lll} M'(p)=M(p)+1 & : p \in t \text{ and } p \notin t & (1) \\ M'(p)=M(p)-1 & : p \in t \text{ and } p \notin t & (2) \\ M'(p)=M(p) & : \text{otherwise.} & (3) \end{array}$$

If there exists a legal sequence of firings that transform M_o to M, M is said to be reachable from M_o. $R(M_o)$ denotes a set of all markings reachable from M_o and is called the reachability set of M_o. For a transition t, t is said to be live at M_o if for any $M \in R(M_o)$, there exists $M' \in R(M)$ such that t is firable at M'. If there exists $M \in R(M_o)$ such that t is not firable at any $M' \in R(M)$ then t is said to be dead at M. If each t of T is live at M, the Petri net is said to be live at M. Conflicts between transitions are expressed such that firing one of them brings others ceased to be firable. The incidence matrix of a Petri net, $A = [a_{ij}]$, is an nxm matrix of integers, and its typical entry is given by

$$a_{ij} = a_{ij}^{+} - a_{ij}^{-}$$

where a_{ij}^{+} is equal to one if there exists an edge from transition i to the output place j, and is equal to zero otherwise. a_{ij}^{-} is equal to one if there exists an edge to transition i from the input place j, and is equal to zero otherwise.

The major advantage for using a net model comes from that it can afford a mathematical analysis and so a formal validation and verification for systems being modelled. The control and verification for such systems are basically reduced to solving reachability and liveness on the nets, respectively. For a given initial marking M_O, to decide whether N is live or not is called a liveness program and for given two marking M_O and M, to decide whether M is reachable from M_O or not is called a reachablity problem. Reachability problem has been proved decidable by Mayr[2]. On the other hand, liveness problem was proved to be equivalent to reachablity problem [3]. However, these two problems require at least exponential order of space and time [4]. For restrictive classes of Petri net such as marked graphs or state machines, we can obtain more amenable conditions for the verification. Let A be an incidence matrix of a Petri net N with initial marking M_O. The evolution of system modelled by N can be expressed by a marking transformation which obeys the following state equations.

$$M = M_O + A^T\Sigma \qquad (4)$$

where Σ is an integer vector, called a firing count vector, composed of numbers of appearance of corresponding transition in a legal firing sequence which transforms M_O to M. A^T is the transpose of A.

An nonnegative integer vector I_p which satisfies

$$A\ I_p = 0 \qquad (5)$$

is called a P-invariant. Multiplying P-invariant to both sides of equation (4), we obtain

$$I_p^T\ M = I_p^T\ M_O\ , \qquad (6)$$

that is, weighted sum of tokens on places correponding to non-zero elements of I_p is invariant through the transformation. Thus if $I_p^T M_O = 0$ for some P-invariant I_p, any output transition of a place corresponding to a non-zero element of I_p can not be made firable, i.e., the net N is not live. For marked graphs, a P-invariant I_p corresponds to a set of places which form a directed circuit. In[5], it was shown that $I_p^T M_O \neq 0$ is also a sufficient condition for a marked graph to be live at M_O. In other words, a marked graph is live if and only if there exists no token-free directed circuit.

If there exists a positive P-invariant, then number of tokens on each place cannot exceed some integer through any firings for any initial marking, and such net is said to be structually bounded. On the other hand, a net is said to be bounded, if , for a given initial marking M_O, number of tokens on each place cannot exceed some integer k. Especially, if k=1, then a net is said to be safe. Live and safeness are commonly required for the well-behavedness of systems. For a marked graph, the safeness is guaranteed if and only if, for each place p, there exists a directed circuit containing p with token sum equal to one. For a state machine N, it is also known that N is live and safe if and only if N is strongly connected and the token sum of all places is exactly one.

Considering equation (4), $I_p^TM = I_p^TM_0$ is known to be a necessary condition for M to be reachable from M_0. For given M_0 and M which satisfy equation (6), there might exist infinite integer solutions of $A^T\Sigma = M - M_0$ such that $\Sigma = \Sigma_0 + I_T$ where I_T is an nonnegative integer solution of $A^TI_T = 0$ and Σ_0 is a minimal nonnegative solution of $A^T\Sigma = M - M_0$. The difficulty in reachablity probrem is to verify the executability of these solutions, i.e., the existence of a legal firing sequence having this solution as the firing count vector. For marked graphs, it is known that it suffices to examine the executability of the unique minimum solution for the reachability. Specifically, for marked graphs, M is reachable from M_0 if and only if $I_p^TM = I_p^TM_0$ for any P-invariant I_p and each element of the minimun firing count vector corresponding to a transition on a token free directed circuit is equal to zero [6].

An nonnegative integer solution I^T of $A^TI_T = 0$ is called a T-invariant. If there exists a positive T-invariant, then the net is said to be consistent. If ther exists a T-invariant $I_T^T=[1,1,\cdots,1]^T$, then the net is said to be 1-consistent. 1-consistency guarantees the existence of an initial marking M_0 and a firing sequence which transforms M_0 back to M_0 and each transition appears just once in the sequence. This property is also required for a well-behaved system with cyclic evolution. A state machine is structually dual of a marked graph. There exist, however, no dual concepts on the dynamic properties between them. For a state machine we have no counter part of the conditions obtained for the reachablity of a marked graph. For a bounded net, the reachability problem can be solved fy finding a path in the finite reachability tree.

Several extensions of Petri nets have been proposed, e.g., coloured Petri nets, stochastic Petri nets, or timed Petri nets. These models naturally have a stronger modelling power for real time systems. The analysis, however, is much more difficult and requires further sophistication in net theory.

2. Modelling of distributed systems

Distributed configurations of systems are architectures extensively used for improving efficiency, flexibility , maintainability, and falt-tolerant ability in general large scale systems such as production systems, computer systems, and many other socio-economic managing systems. Design of these systems, however, requires much more attention on avoiding deadlock, conflict, or starvation. Petri nets as introduced in the preceeding section can be used for analysing these problems. On the other hand, when we adopt a net theoretic approach for designing such systems, it is necessary to structurize a net as a composition of components: It is necessary to devise a way of construction of total system with desired properties from well-examined components just as commonly done in electrical network synthesis. The discussions below might hopefully be a starting point for such investigations.

We define a component as a safe Petri net with prescribed input and output places as shown in Fig. 1.

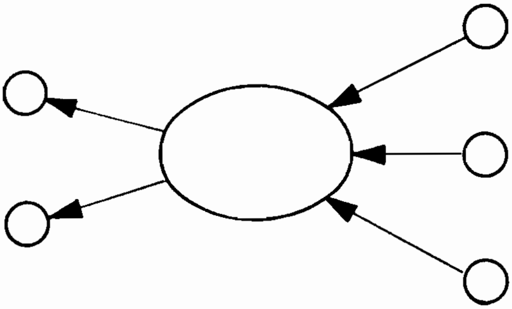


Fig.1 a component net

Tokens on input and output places represent the activating signals and the completion signals of events in the subnet, respectively. Fig. 2 models a inverter with input place S_{in} and output place S_{out} [7]. An edge -• denotes a self-loop $\rightleftarrows$. At this marking, the state {p2,p3} can change to the state [pl,p4]. The completion of the firings signal out a token on S_{out} and make the net to be dead. Formally, we can define a component of composite net as follows.

Definition 1. A component of a composite net is defined as a net N_k such that

$$N_k := (P_k, T_k, P_k(IN), P_k(OUT)), \quad k=1,2,\dots,n$$

where

P_k: a set of internal places,
T_k: a set of transitions,
$P_k(IN)$: a set of input places,
$P_k(OUT)$: a set of output places.

The incidence matrix A_k of N_k can be written as

$$A_k = \begin{array}{cccccc} & P_k & & P_k(IN) & & P_k(OUT) \\ T_k & [A_k^0 & : & A_k^- & : & A_k^+ \;] \end{array} \qquad (7)$$

Non-zero elements in $A_k^-(A_k^+)$ are all negative (positive) respectively.

The well-behavedness of a component with an initial marking on internal places is defined as follows.

Definition 2. N_k is called a well-behaved component if it satisfies the following properties.
(1) Net N_k^0 obtained from N_k by removing all input and output places (and the corresponding edges) is live and safe for a given initial marking $M_0(P_k)$.

(2) $N_k{}^0$ is 1-consistent, i.e., for $e_k{}^T=[1,...,1]^T$,

$$(A_k{}^0)^T e_k=0 \qquad (8)$$

where $A_k{}^0$ is an incidence matrix of $N_k{}^0$.
(3) With internal marking $M_0(P_k)$, the substraction of a token from a input place can only happen simultaneously for all input places after they all have positive tokens. Moreover, this is the only events that can happen at $M_0(P_k)$.
(4) Before a token has been added to all output places $P_k(OUT)$, no subtraction of a token from $P_k(IN)$ can occur.
(5) Unless all of $P_k(IN)$ become marked again after finishing the addition of a token to all of $P_k(OUT)$, net N_k comes to be dead at, say, $M(P_K)$. If all of $P_k(IN)$ become marked again, then N_k has a property (3) at $M(P_k)$ instead of $M_0(P_k)$.

Remark Well-behavedness of a component depends on the net structure and the initial marking. Property (2) gurantees the cyclic internal evolution. For general nets, it is difficult to verify properties (3), (4), and (5). We would assume that the structure of a component is simple enough that the well-behavednes can easily be verified. A net as shown in Fig. 2 is a simple example of a well-behaved component.

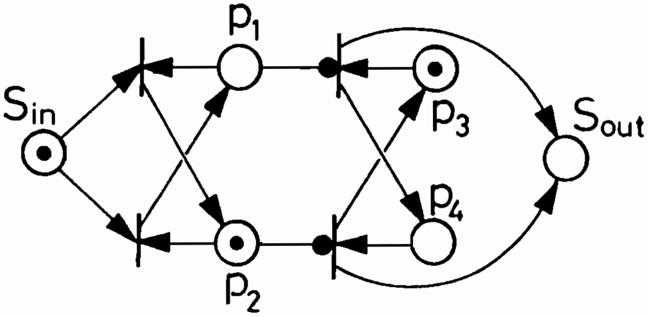


Fig.2 an inveter

Now the composition of components can be seen as the fusion of corresponding nets by identifying each input and output places.

Definition 3. For components N_i and N_j, the composition is defined by identifying a subset of output places of N_i as a subset of input places of N_j and vice vasa.
Let F_{ij} be defined as

$$F_{ij}(p,q)= \begin{cases} 1: p \in P_i(OUT) \text{ and } q \in P_j(IN) \text{ are identified.} \\ 0: \text{otherwise} \end{cases}$$

Then the connection matrix F of N_i and N_j can be written as

$$F = \begin{array}{c|ccc|} & P_i(\mathrm{OUT}) & P_j(\mathrm{OUT}) & P^*(\mathrm{OUT}) \\ P_i(\mathrm{IN}) & 0 & F_{ij} & F_{*i} \\ P_j(\mathrm{IN}) & F_{ij} & 0 & \\ P^*(\mathrm{IN}) & F_{i*} & F_{j*} & 0 \end{array}$$

where P^*(IN) and P^*(OUT) are the set of output places and input places of a composite net, respectively, which in turn can be seen as input places and output places of outside world (environment), respectively.

The incidence matrix A of a composite net formed by n number of components can be calculated from each incidence matrix A_k and their connection matrices as defined in (9), i.e.,

<u>Step 1.</u> Let $\tilde{A}$ be

$$\tilde{A} = \begin{array}{c|cccc|ccccc|} & & & & & P_1(\mathrm{OUT}) & P_2(\mathrm{OUT}) & \cdots & P_n(\mathrm{OUT}) & P^*(\mathrm{OUT}) \\ T_1 & A_1^0 & & & & A_1^+ & A_1^-F_{21} & \cdots & A_1^-F_{n1} & A_1^-F_{*1} \\ T_2 & & A_2^0 & & & A_2^-F_{12} & A_2^+ & \cdots & A_2^-F_{n2} & A_2^-F_{*2} \\ & & & & & & & & & \\ T_n & & & & A_n^0 & A_n^-F_{1n} & & & A_n^+ & A_n^-F_{*n} \end{array}$$

<u>Step 2.</u> If

$F_{ij}(k,l)=1$, $i<j$,

then add $[0,\ldots,0,[A_j]_l{}^T,0,\ldots,0]^T$ to the column $[P_i(\mathrm{OUT})]_k$ of $\tilde{A}$ and delete the column $[P_j(\mathrm{OUT})]_l$. The resulting matrix is the incidence matrix.

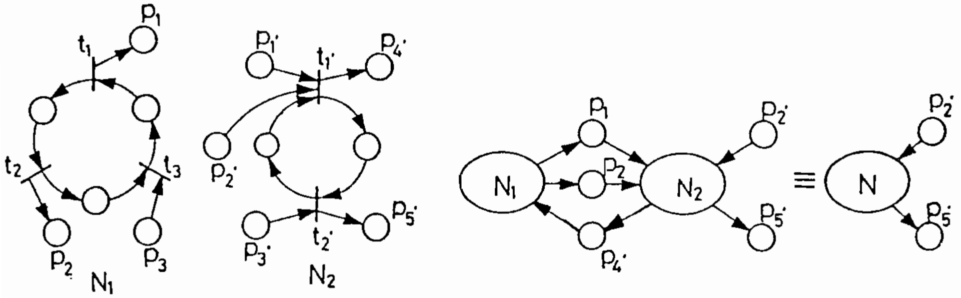


Fig. 3 an example of composite net

Example 1. Let N_1 and N_2 be components as shown in Fig. 3. By identifying $P_1=P_1'$, $P_2=P_3'$, and $P_4'=p_3$, the connection matrix F is expressed as

$$F = \begin{array}{ll|cc:cc:c|}
 & & \multicolumn{2}{c}{P_1(\mathrm{OUT})} & \multicolumn{2}{c}{P_2(\mathrm{OUT})} & P^*(\mathrm{OUT}) \\
 & & 1 & 2 & 4' & 5' & \\
P_1(\mathrm{IN}) & 3 & 0 & 0 & 1 & 0 & 0 \\ \hdashline
 & 1' & 1 & 0 & 0 & 0 & 0 \\
P_2(\mathrm{IN}) & 2' & 0 & 0 & 0 & 0 & 1 \\
 & 3' & 0 & 1 & 0 & 0 & 0 \\ \hdashline
P^*(\mathrm{IN}) & & 0 & 0 & 0 & 1 & \\
\end{array}$$

The incidence matrices of N_1 and N_2 are, respectively,

$$A_1 = [\quad A_1^0 \qquad A_1^- \qquad A_1^+ \quad]$$

$$= \begin{array}{c|ccc:c:cc|}
 & P_1 & P_2 & P_3 & P_1(\mathrm{IN}) & \multicolumn{2}{c}{P_1(\mathrm{OUT})} \\
t_1 & 1 & & -1 & -1 & 0 & 0 \\
t_2 & -1 & 1 & & 0 & 0 & 1 \\
t_3 & & -1 & 1 & 0 & 1 & 0 \\
\end{array}$$

$$A_2 = [\quad A_2^0 \qquad A_2^- \qquad A_2^+ \quad]$$

$$= \begin{array}{c|cc:ccc:cc|}
t_1 & 1 & -1 & -1 & -1 & 0 & 1 & 0 \\
t_2 & -1 & 1 & 0 & 0 & -1 & 0 & 1 \\
\end{array}$$

Thus the incidence matrix of the composite net can be written as

$$A = \begin{array}{c|c:c:c:c:c|}
 & P_1 & P_2 & P_1(\mathrm{OUT}) & P_2(\mathrm{OUT}) & P^*(\mathrm{OUT}) \\
t_1 & & & & & \\
t_2 & A_1^0 & & A_1^+ & A_1^-F_{21} & A_1^-F_{*1} \\
t_3 & & & & & \\ \hdashline
t_1 & & & & & \\
t_2 & & A_2^0 & A_2^-F_{12} & A_2^+ & A_2^-F_{*2} \\
\end{array}$$

$$= \begin{array}{|ccc:cc:cc:cc:c|}
1 & 0 & -1 & & & 0 & 0 & -1 & 0 & 0 \\
-1 & 1 & 0 & \multicolumn{2}{c:}{0} & 0 & 1 & 0 & 0 & 0 \\
0 & -1 & 1 & & & 1 & 0 & 0 & 0 & 0 \\ \hdashline
 & & & 1 & -1 & -1 & 0 & 1 & 0 & -1 \\
 & 0 & & -1 & 1 & 0 & -1 & 0 & 1 & 0 \\
\end{array}$$

3. Analysis of distributed systems

A composite net as described in the preceeding section can be seen itself as another component with more detailed internal specifications if it also satisfies the properties 1 through 5 of Definition 2 with respect to $P^*(IN)$ and $P^*(OUT)$. According to the design stage, we can choose appropriate level of refinement by resolving a component into more precise subcomponents. For continuing this process in a hierarchical fashion, it is important to verify which property can be preserve through the composition. Let N be a composite net of n well-behaved components and let N^0 be a net obtained from N by removing all $P^*(IN)$ and $P^*(OUT)$. Then the incidence matrix A^0 of N^0 can be written as

$$A^{0} = \begin{array}{cccc} & & & P_c \\ \hline A_1 & & & \\ & A_2 & & A_c \\ & & A_n & \end{array}$$

where P_c is a subset of I/O places of components used for interconnections. Since each component is 1-consistent,

$$(A_i{}^0)^T e_i = 0, \qquad i = 1,\dots,n$$

for $e_i{}^T = [1,\dots,1]^T$. For $e^T = [e_1{}^T, e_2{}^T, \dots, e_n{}^T]$

$$(A^0)^T e = 0$$

if and only if $(A_c)^T e = 0$. Thus we obtain

Theorem 1. A composite net is 1-consistent if and only if, for each place of P_c, the number of incoming edges is equal to the number of outgoing edges.

Next we consider the preservation of liveness and safeness in N^0. Properties (3), (4), and (5) imply that each live and safe subnet $N_i{}^0$ can equivalently be contracted to a single transition with respect to the I/O behavior. Let $\hat{N}^0$ be a contracted net of N^0 with P_c as a set of places and each component $N_i{}^0$ as a transition. It is obvious that N^0 is live and safe if and only if $\hat{N}^0$ is live and safe. The preservation of these properties in N^0 thus depends on the way of interconnection and the initial token distribution on P_c. Here we consider two restrictive structures of connection: we assume $\hat{N}^0$ to be a class of marked graph decomposable nets (MGD-net) or state machine decomposable nets (SMD-net) defined as follows. Let N = (P,T) be a Petri net.

Definition 4. For a subset T_p of T, a subnet composed $({}^{\bullet}T_p \cup T_p^{\bullet}, T_p)$ is called a T-closed subnet of N and denoted as $\langle T_p \rangle$. Similarly, for a subset P_t of P, a subnet composed of (P_t, $^\bullet P_t U P_t^\bullet$) is called a P-closed subnet of N and denoted as $<P_t>$.

Definition 5. If there exists a mutually disjoint subdivision T_1, T_2,..., T_k of T such that $N = <T_1>U<T_2>U...U<T_k>$ and each $<T_i>$ is a strongly connected marked graph , then N is called MGD-net.

Similarly, if there exists a mutually disjoint subdivision P_1, P_2,..., P_k of P such that $N = <P_1>U<P_2>U...U<P_k>$ and each $<P_i>$ is a strongly connected state machine, then N is a SMD-net.

SMD-net is a structually dual of MGD-net, i.e., SMD-net is obtained by replacing each place and transition of MGD-net as a transition and place, respectively, and inverting the direction of all edges of MGD-net.

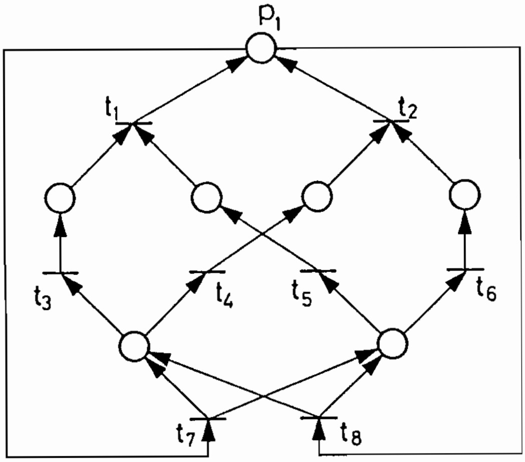


Fig.4 a MGD-net

A net as shown in Fig.4 is an example of MGD-net with marked graph components N_1=<{ t_1 , t_3 , t_5 , t_7 }> and N_2=<{ t_2 , t_4 , t_6 , t_8 }>.

Putting a token on P_1, it can be seen that both N_1 and N_2 are live and safe marked graphs. However, by firing t_7 -> t_3 -> t_6 , the net becomes dead.

The convenient way of verifying the liveness and safeness of $\hat{N}^0$ which is assumed to be a MGD-net or SMG-net is to find out the conditions that guarantee the same properties from that of marked graph or state machine components because the liveness and safeness can be verified very easily for these classes of Petri net as stated in Section 3.

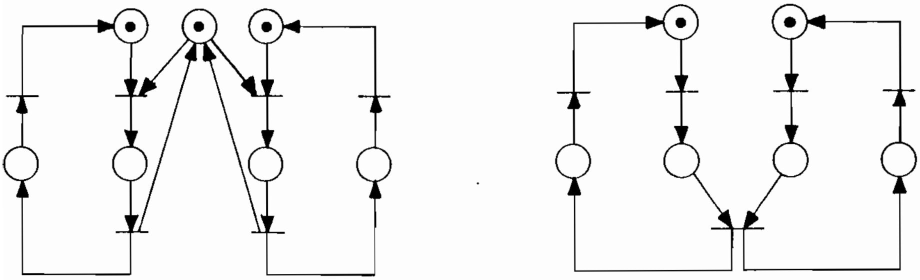
Definition 6. A MGD-net (P,T) is said to be complete if the following condition is satisfied, i.e., for any mapping $A : P \rightarrow P^{\bullet}$, there exists at least one T-closed strongly connected marked graph $\langle T_p \rangle$ such that $A(p) \in T_p$ for any place $p \in \langle t_p \rangle$.

Definition 7. A SMD-net (P,T) is said to be complete if the following condition is satisfied, i.e., for any mapping $B : T \rightarrow {}^{\bullet}T$, there exists at least one P-closed strongly connected state machine $\langle P_t \rangle$ such that $B(t) \in P_t$ for any transition $t \in \langle P_t \rangle$.

The following result was originally prooved by Hack [8] for SMD-net. For the dual case the proof can be done similarly.

Theorem 2. A complete MGD-net and a complete SMD-net are live and safe if there exists a decomposition such that each component (strongly connected marked graph or strongly connected state machine, respectively) is live and safe.

Note that MGD-net as shown in Fig.4 is not complete. Examples of complete MGD- and SMD-net are shown in Fig.5. It can easily be verified that they are both live and safe.



(a) MGD-net

(b) SMD-net

Fig.5 complete MGD- and SMD-net

By this theorem, if the connection of components is restricted so that $\hat{N}_0$ to be a MGD- or SMD-net, the live and safeness of composite net can be assured by easy inspection.

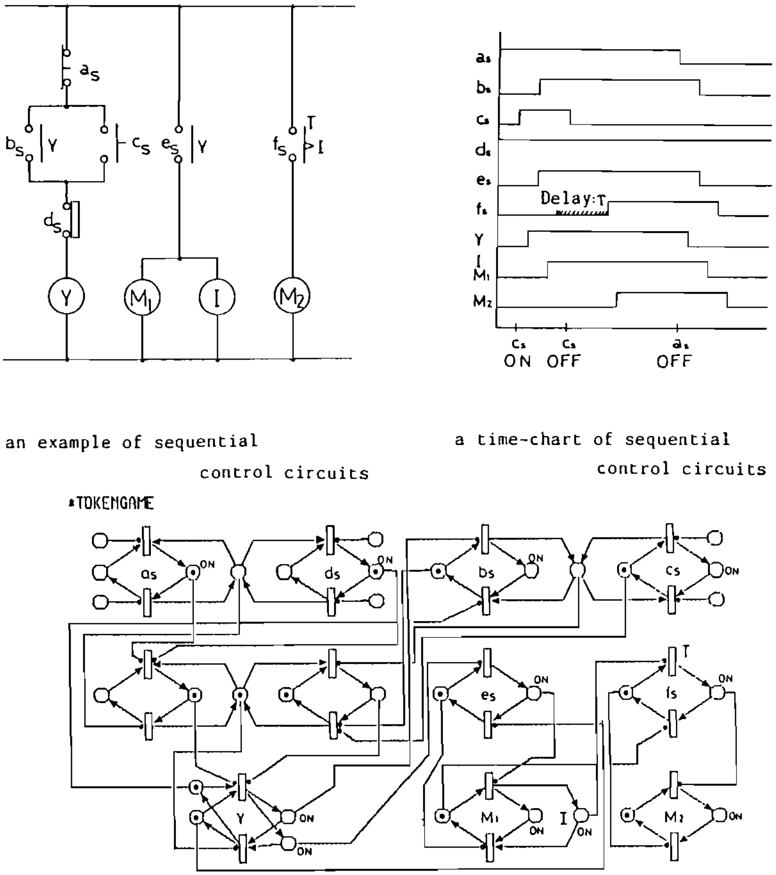
4. Applications of Net-specification Methodology

Industrial applications of Petri nets have increased in numbers and scale from the beginning of 1980's. In Europe, a group in LAAS-

CNRS, Toulouse, France, has been conducting a project called SECOIA [9]. The aim of SECOIA is integrating different levels of Flexible Manufacturing Systems (FMS) such as local machine control, coordination of subsets, monitoring and real-time shop scheduling, and planning/product mix evaluation and management. The integration of these levels by LAN is not sufficient because if the programmation of each level is done by different tools, the global interlevel communication will suffer from a side effect and it will be difficult to verify, maintain and modify it.

SECOIA consistently adopts a net-based methodology as a formal mathematical tool for the global specification , design, and structural verification. They use the concept of common place for describing handshake procedure in distributed applications and also use a concept of module with I/O places which can equivalently be substituted by other module. In France, another group in SYSECA, Saint-Cloud Cedex, has been developing ESPRIT Project "For-Me-Too" [10]. In this project, the way of fusion and substitution of component nets are inplemented so that it can afford a global validation of large real time systems. In England, a group in Plessey Electronic Systems Research Ltd., Ramsey, is developing and implementing CAD and control systems of databases for query in military communication system called Project PTARMIGAN [11]. By using Petri net for hierarchically describing each database structure and the control sequence flow, the flexibility and maintainability in change assessments are greatly improved. The validation of processes can be done through reachability analysis.

In Japan, a Petri net-based station controller named SCR for flexible and maintainable sequence control has been developed for a factory automation system by Hitachi Ltd., Kawasaki, [12]. As far as we know, this is the first commercial product based on Petri net. In SCR, safe Petri net, called C-net, which augments the control functions by adding Predicates onto each transition, is installed and is used for specification, simulation, and real time control of coordinating robots or parts assembly station, e.g.. A group in Mitsubishi Electric Co., Itami, [13] has adopted a Petri net as a language for describing specifications to cope with frequent changes in design stage of large scale systems such as power plants. A Petri net-based concurrent system simulator, called PCSS, has been developed in Osaka University [14]. In PCSS, controllers and controlled objects are both modelled by Petri nets. By prescribing a time delay at a transition, PCSS can simulate the concurrent behavior of a real system. Fig. 6(a) shows a Petri net model in PCSS of a relay ladder diagram. PCSS simulates the swiching sequence and outputs the time chart as shown in Fig. 6(b) . Note that a delay T is prescribed at the switch fs.



an example of sequential control circuits

a time-chart of sequential control circuits

Petri net models of sequential control circuits

Fig. 6 an example of simulation by PCSS

5. Conclusion

Net theory provides a formal approach to analysis and description of concurrent systems. Specifically, flexiblity and maintainability of control softwares for such systems are improved by Petri net-based descriptions. A clear seperation of events and conditions due to Petri net modelling improves system comprehension in various activity levels and opens up a new system viewpoint.

The hierarchical nature of Petri net-description reflects on a set of top-down design methodology as seen in industrial applications. Verification and validation can be done in mathematical way by examining,e.g., the liveness, safeness, and reachability on the nets. Invariants play a central role in structural analysis such as boundedness or consistency which in turn can validate the mutual exclusion in critical section or cyclic stational motion, respectively.

Synthesis aspect of a net theoretic approach is, however, presently far from satisfactory stage. Fusion and substitution of component nets should be based on more concrete equivalence notion of nets. Also, a formal way of global analysis via properties of components and their interconnections as briefly introduced here should be established to this end.

References

1. B.T.Denvir et al. edited: The analysis of concurrent systems, Lecture notes in computer science 207, Springer-Verlag(i983)
2. E.W.Mayr: An algorithm for general Petri net reachability problem, Proc. of the 13th ann. ACM Symp. on Theory of Computing, Milwaukee, USA(1981)
3. M.H.Hack: The recursive equivalence of the reachability problem and the liveness problem for Petri nets and vector addition system, Project MAC, MIT, Cambridge, USA(1974)
4.R.Lipton: The reachability problem requires exponential space, Report 62, Yale University, New Haven, USA(1976)
5. F.Commoner et al.: Marked directed graphs, J. of Computer and System Science, 5, 511/523 (1971)
6.T.Murata: Circuit theoretic analysis and synthesis of marked graphs, IEEE Trans. on Circuit and Systems, CAS-24-7, 400/405 (1977)
7. J.Grabowski: On the analysis of switching circuit by means of Petri nets, Electronishe Informationsverarbeitung und Kybernetik, EIK 14, 611/617 (1978)
8. M.H.Hack: Analysis of production schemata by Petri nets, M.S. Thesis, TR-94, Project MAC, MIT, Cambridge, USA(1972)
9. R. Valette et al.: Putting Petri nets to work for controlling flexible manufacturing systems, Proc. of the IEEE International Symposium on Circuits and Systems, Kyoto, Japan(1985)
10. E.Sibille et al.: Nets as formalisms, methods and tools for specification of large real time systems, Proc. of the 7th European Workshop on Application and Theory of Petri nets, Oxford, England(1986)
11. R. Threadgold: Application of Petri nets to the maintenance of a large software system, Proc. of the 7th European Workshop on Application and Theory of Petri nets, Oxford, England(1986)
12. T.Murata et al.:A Petri net-based controller for flexible and maintainable sequence contril and its applications in factory automation, IEEE Trans. on Industrial Electronics, IE-33, 1, 1/8(1986)
13. Y.Nakamura et al.: The reachability of complementary-place Petri nets, Proc. of the IEEE International Symposium on Circuits and Systems, San Jose, California, USA(1986)
14. T.Tsuji,S.Kumagai,S.Kodama, and T.Yamada: Modeling and verification of sequential control systems by Petri nets, Proc. of the IEEE International Symposium on Circuits and Systems, San Jose, California, USA(1986)

SET-VALUED CALCULUS IN PROBLEMS OF ADAPTIVE CONTROL

A. B. Kurzhanski
IIASA

Introduction

This paper deals with *feedback control* for a linear nonstationary system whose objective is to reach a preassigned set in the state space while satisfying a certain *state constraint*. The state constraint to be fulfilled cannot be predicted in advance being governed by a second "uncertain" system, with its state space variable unknown and available only on the basis of observations. It is assumed that there is no statistical data for the uncertain parameters of the second system the only information on these being the knowledge of some constraints on their admissible values. Therefore the state constraint to be satisfied by the basic system may be specified only through an *adaptive procedure* of "guaranteed estimation" and the objective of the basic process is to *adapt to this constraint.*

The problems considered in the paper are motivated by some typical applied processes in environmental, technological, economical studies and related topics.

The techniques used for the solution are based on nonlinear analysis for *set-valued maps.* They also serve to illustrate the relevance of set-valued calculus to

- problems of control in devising solutions for the "guaranteed filtering and extrapolation" problems
- constructing set-valued feedback control strategies,
- duality theory for systems with set-valued state space variables,
- approximation techniques for control problems with set-valued solutions, etc.

The research in the field of control and estimation for uncertain systems (in a deterministic setting), in differential games and also in set-valued calculus, that motivated this paper, is mostly due to the publications of [1-10].

1. The Uncertain System

Consider a system modelled by a linear-convex differential inclusion

$$\dot{q} \in A(t)q + \mathbf{P}(t) \tag{1.1}$$

$$t \in \mathbf{T} = \{t : t_0 \leq t \leq t_1\} ,$$

where $q \in \mathbf{R}^n$, $A(t)$ is a continuous matrix function $(A : \mathbf{T} \longrightarrow \mathbf{R}^{n \times m})$, $\mathbf{P}(t)$ is a continuous multivalued map from T into the set *conv* $\mathbf{R}^n$ of convex compact subsets of $\mathbf{R}^n$. (Here $\mathbf{R}^n$ will stand for the n-dimensional vector space and $\mathbf{R}^{m \times n}$ for the space of $m \times n$ - matrices.)

The function $\mathbf{P}(t)$ reflects the *uncertainty in the specification of the system inputs.* The initial state $q(t_0) = q^{(0)}$ is also taken to be unknown in advance. Namely,

$$q^{(0)} \in Q^{(0)} \tag{1.2}$$

with the set $Q^{(0)} \in$ *conv* $\mathbf{R}^n$ being given.

An isolated trajectory of (1.1) generated by point $q^{(\tau)} = q[\tau]$ will be further denoted as $q[t] = q(t , \tau , q^{(\tau)})$, while the set of all solutions to (1.1) that start at $q^{(\tau)}$ will be denoted as $Q(t , \tau , q^{(\tau)})$.

We also assume

$$Q(t , \tau , Q^{(\tau)}) = \bigcup\{Q(t , \tau , q^{(\tau)}) \mid q^{(\tau)} \in Q^{(\tau)}\} .$$

The sets $Q(t , t_0 , q^{(0)})$,$Q(t , t_0 , Q^{(0)})$ are therefore the *attainability domains* for (1.1) (from $q(t_0) = q^{(0)}$ and $Q^{(0)}$ respectively).

It is known that the multivalued function

$$Q[t] = Q(t , t_0 , Q^{(0)})$$

satisfies the "funnel equation", [11]

$$\lim_{\sigma \dashrightarrow 0} \sigma^{-1} h(Q[t + \sigma] , (E + A(t)\sigma) \; Q[t] + \mathbf{P}(t) \, \sigma) = 0 \tag{1.3}$$

where

$$h(Q' , Q'') = \max\{h^+(Q' , Q'') , h^-(Q'' , Q')\} ,$$

$$h^+(Q' , Q'') = \max_p \min_q \{ || \, p - q \, || \mid p \in Q' , q \in Q'' \} ,$$

$$h^+(Q' , Q'') = h^-(Q'' , Q')$$

is the *Hausdorff distance* between $Q' \in$ *conv* $\mathbf{R}^n$, $Q'' \in$ *conv* $\mathbf{R}^n$ [12].

Let us now assume that there is some *additional information* on the system (1.1), (1.2). Namely, this information arrives through an *equation of observations*

$$y \in G(t) \; q(t) + \mathbf{R}(t) \tag{1.4}$$

where $y \in \mathbf{R}^m$, $G(t)$ is continuous $(G : \mathbf{R}^n \longrightarrow \mathbf{R}^m)$ and the set-valued function $\mathbf{R}(t)$ from T into *conv* $\mathbf{R}^m$ reflects the presence of "noise" in the observations. The realization $y_\tau(\sigma) = y(\tau + \sigma)$, $t_0 - \tau \leq \sigma \leq 0$, of the observation y being given, it is possible to construct an "informational domain" Q_τ $(\bullet , t_0 , Q^{(0)} \mid y_\tau(\bullet))$ of all trajectories consistent with (1.1)-(1.3) and with the given realization $y_\tau(\bullet)$. The cross-section $\mathbf{Q}$ $(\tau , t_0 , Q^{(0)})$ of this set is the "generalized state" of the "total" system (1.1), (1.2), (1.4), (for convenience we further omit an explicit indication of $y_\tau(\bullet)$ taking it to be fixed).

Clearly, for $\tau' \leq \tau''$ we have $\mathbf{Q}(\tau'' , t_0 , Q^{(0)}) = \mathbf{Q}(\tau'' , \tau' , \mathbf{Q}$ $(\tau' , t_0 , Q^{(0)}))$

The map $\mathbf{Q}(\tau'' , t_0 , Q^{(0)}) = \mathbf{Q}[\tau]$ thus satisfies a *semigroup property* and defines a *generalized dynamic system.* The function $\mathbf{Q}$ $[\tau]$ also satisfies a more complicated version of the funnel equation (1.3), [3].

$$\lim_{\sigma \longrightarrow o} \sigma^{-1} h(\mathbf{Q}[\tau + \sigma] , (E + A(\tau)\sigma) \mathbf{Q}[\tau] + \mathbf{P}(\tau)\sigma)) \cap \mathbf{Y}[\tau + \sigma]) = 0$$

$$\mathbf{Q}[t_0] = Q^{(0)} \tag{1.5}$$

where

$$\mathbf{Y}[\tau] = \{q : G(\tau)\ q \in y(\tau) - \mathbf{R}(\tau)\}$$

is taken to be such that its support function

$$\rho(l \mid \mathbf{Y}\ [\tau]\ ) = sup\{(l\ , y) \mid y \in \mathbf{Y}[\tau]\}\ .$$

is continuously differentiable in l and τ. The latter property is true if $\rho(l \mid Y\ [\tau]\ )$ and $y(\tau)$ are continuously differentiable in the respective variables. This in turn is ensured if the measurement $y(t)$ is generated due to equation

$$y(t) = G(t)\ x(t) + \xi(t)\ , \xi(t) \in \mathbf{R}(t)$$

by continuously differentiable functions $\xi(t)$ and $G(t)$.

Consider the inclusion

$$\dot{q}_L \in (A(t) - L(t)\ G(t))\ q_L + L(t)\ (y(t) - \mathbf{R}(t)) + \mathbf{P}(t) \tag{1.7}$$

$$q_L(t_0) = q_L^{(0)}\ , q_L^{(0)} \in Q^{(0)}$$

whose attainability domain is

$$Q_L\ (t\ , t_0\ , Q^{(0)}) = Q_L\ [t]$$

Lemma 1.1 [13,14] *The following relation is true*

$$\bigcap Q_L\ (t\ , t_{(0)}\ , Q^{(0)}) = \mathbf{Q}\ (t\ , t_0\ , Q^{(0)}) = \mathbf{Q}\ [t]\ , \tag{1.8}$$

where the intersection is taken at all continuous matrix-valued functions $L(t)$ *with values* $L \in \mathbf{R}^{n \times m}$.

The last Lemma allows to decouple the calculation of $\mathbf{Q}[t]$ into the calculation of sets $Q_L[t]$ governed by "ordinary" differential inclusions of type (1.7).

According to [11] each of the multivalued functions $Q_L\ [t]$ satisfies a respective funnel equation

$$\lim_{\sigma \longrightarrow 0} \sigma^{-1} h(Q_L[\tau + \sigma] , (E + \sigma(A(\tau) - L(\tau)\ G(\tau))\ \mathbf{Q}_L[\tau] + \tag{1.9}$$

$$+ L(\tau)(y(\tau) - \mathbf{R}(\tau))\ \sigma + \mathbf{P}(\tau)\sigma) = 0$$

$$Q_L\ [t_0] = Q^{(0)}\ .$$

Hence from (1.8) it follows that the solution to (2.5) may be decoupled into the solutions of equations (1.9). The latter relations allow for a respective difference scheme.

2. An Inverse Problem

Assume that a square-integrable function $y_{t_1}(\sigma \mid \tau) = y(t_1 + \sigma)$, $\tau - t_1 \leq \sigma \leq 0$ and a set $N \in \text{conv } \mathbf{R}^n$ are given. Denote $\mathbf{W}(\tau , t_1 , N)$ to be the variety of all points $w \in \mathbf{R}^n$ for each of which there exists a solution $q(t , \tau , w)$ that satisfies (1.1), (1.4) for $t \in |\tau , t_1|$, and $q(t_1 , \tau , w) \in N$.

We observe that $W(\tau , t_1 , N)$ is of the same nature as $\mathbf{Q}(t , t_0 , Q^{(0)})$ except that it should be treated in backward time.

Hence, we will have to deal with the solutions to the inclusions

$$\dot{q} \in A(t)q + \mathbf{P}(t) , \quad t \leq t_1 , \tag{2.1}$$

$$t \in T , q(t_1) = q^{(1)} , q^{(1)} \in N$$

with isolated trajectories $q(t , t_1 , q^{(1)})$ that satisfy the restriction

$$q(t) \in \mathbf{Y}(t) \; \forall \; t \in T \tag{2.2}$$

Following Lemma 1.1, we have a similar

Lemma 2.1. The following equality is true

$$W(t , t_1 , N) = \bigcap_L W_L \, (t , t_1 , N) \tag{2.3}$$

the intersection being taken over all continuous matrix-valued functions $L(t)$ with $L \in \mathbf{R}^{m \times n}$, and $W_L(t , t_1 , N)$ is the assembly of all solutions to the inclusion

$$\dot{w}_L \in (A(t) - L(t) \; G(t))w_L + L(t)(y(t) - \mathbf{R}(t)) + \mathbf{P}(t) , \tag{2.4}$$

$$w(t_1) \in N$$

Lemma 2.2 Each of the realizations $W_L \, (t , t_1 , N) = W_L|t|$ may be achieved as a solution to the funnel equation

$$\lim_{\sigma \longrightarrow +0} \sigma^{-1} \, h(W(t - \sigma) , (E - \sigma(A(t) - L(t) \; G(t))) \; W(t) -$$

$$- L(t)(y(t) - \mathbf{R}(t)) \, \sigma - \mathbf{P}(t)\sigma) = 0$$

$$W(t_1) = \mathbf{N}$$

The uncertain system and inverse problem of the above will play an essential part in the formulation and the solution of the *adaptive control problem* discussed in this paper.

3. The Adaptive Control Problem

Consider a control process governed by the equation

$$\frac{dp}{dt} = C(t)p + u , t \in T \tag{3.1}$$

where $p \in \mathbf{R}^n$, $C(t)$ is a continuous matrix function ($C : \mathbf{R}^n \longrightarrow \mathbf{R}^n$) and u is restricted by the inclusion

$$u \in \mathbf{V}(t)$$

where $\mathbf{V}(t)$ is a continuous multivalued map from $\mathbf{T}$ into conv $\mathbf{R}$.

The basic problem considered in this paper is to devise a feedback control law that would allow the system to *adapt* to an *uncertain state constraint.*

Assume that an uncertain system (1.1), (1.2), (1.4) is given and a *state constraint* is defined by a continuous multivalued map

$$K(t)\ (\mathbf{K} : \mathbf{T} \longrightarrow \text{conv } \mathbf{R}^n)$$

The objective of the control process for system (3.1) will be to satisfy the constraint

$$p(t) + q(t) \in \mathbf{K}(t)\ , \forall\, t \in T\ , \tag{3.2}$$

and also a terminal inclusion

$$p(t_1) \in \mathbf{M}\ , \mathbf{M} \in \text{conv } \mathbf{R}^n \tag{3.3}$$

The principal difficulty is here caused by the fact that vector $q(t)$ of (3.2) is unknown and that the information on its values is confined to the inclusion

$$q(t) \in \mathbf{Q}(t\ , t_0\ , Q^{(0)})$$

Therefore the total state constraint on p at instant t will actually be

$$p(t) + \mathbf{Q}(t\ , t_0\ , Q^{(0)}) \in K(t) \tag{3.4}$$

where the realization

$$\mathbf{Q}|t| = \mathbf{Q}(t\ , t_0\ , Q^{(0)})$$

cannot be predicted in advance, being governed by the uncertainty

$$\omega_t\,(\bullet) = \{q(t_0)\ , \xi_t\,(\bullet)\ , v_t\,(\bullet)\}$$

Here the notation $f_t\,(\bullet)$ stands for

$$f_t\,(\sigma) = f(t + \sigma)\ , t_0 - t \leq \sigma \leq t\ .$$

In order to pose the *adaptive control* problem it is necessary to introduce the notion of the *state (the position)* of the overall system (3.1)-(3.3).

The *position* of the system (3.1)-(3.3) will be defined as the triplet

$$\{t\ , p\ , y_t(\bullet)\}$$

Hence the solution to the problem will be sought for in the *class of multivalued strategies*

$$V = U(t\ , p\ , y_t(\bullet))$$

with $U \in$ *conv* $\mathbf{R}^n$ and with the dependence of U upon $t\ , p\ , y_t(\bullet)$ being such that the joint system

$$\dot{p} \in C(t)p + \mathbf{U}(t\ , p\ , y_t(\bullet)) \tag{3.5}$$

$$\dot{q} \in A(t)q + \mathbf{P}(t) \tag{3.6}$$

$$y - Gq \in \mathbf{R}(t) \tag{3.7}$$

The specific triplet $\omega_\tau^{o\,*}(\bullet)$ should satisfy the inclusions

$$q^* \in \mathbf{Q}|\tau| \, , \, v_\tau^*(\bullet) \in \mathbf{P}_\tau^o(\bullet) \, , \, \xi_\tau^*(\bullet) \in \mathbf{R}_\tau^o(\bullet)$$

A triplet of this kind will be further referred to as an *admissible triplet*, i.e.

$$\omega_\tau^*(\bullet) \in \Omega_\tau(\bullet) \tag{4.1}$$

where

$$\Omega_\tau^o(\bullet) = \mathbf{Q}|\tau| \times \mathbf{P}_\tau^o(\bullet) \times \mathbf{R}_\tau^o(\bullet)$$

and as indicated above

$$\mathbf{P}_\tau^o(\bullet) = \{v_\tau^o(\bullet) : v(t) \in \mathbf{P}(t) \, , \tau \le t \le t_1\}$$

$$\mathbf{R}_\tau^o(\bullet) = \{\xi_\tau^o(\bullet) : \xi(t) \in \mathbf{R}(t) \, , \tau \le t \le t_1\}$$

Now obviously it will be possible to devise a related prolongation for the set-valued function $\mathbf{Q}^*|t|$ from $|t_0 \, , \tau|$ onto the interval $(\tau \, , t_1|$ in the form of a realization

$$\mathbf{Q}^*|t| = \mathbf{Q}(t \, , \tau \, , \mathbf{Q}(t \, , t_0 \, , Q^{(0)} \mid y_\tau^*(\bullet)) \mid y_\tau^{o\,*}(\bullet))$$

According to |7| and to the statements of § 1 of this paper, the multivalued map $\mathbf{Q}^*[\bullet|$ may be specified through the system

$$\dot q \in (A(t) - L(t)G(t))q + \mathbf{P}(t) + L(t)(y^* - \mathbf{R}(t))$$

$$\dot q^* = A(t) \, q^* + v^*(t)$$

$$y^* = G(t) \, q^* + \xi^*(t) \, ,$$

$$q^*(\tau) = q_\tau^* \, , \, q(\tau) = q_\tau$$

or, in equivalent form, through the system

$$\dot z^* \in (A(t) - L(t)G(t) \, z^* + (\mathbf{P}(t) - v^*(t)) - L(t)(\mathbf{R}(t) - \xi^*(t)) \tag{4.2}$$

$$z^*(\tau) = q_\tau - q_\tau^*$$

where

$$z^*(t) = q(t) - q^*(t) \qquad \tau \le t \le t_1$$

Denote $Z_L^*(\bullet \, , \tau \, , Z^*[\tau])$ to be the set of all solutions to (4.2) that start from $Z^*[\tau]$ at instant τ.

What follows from |13,14| is

Lemma 4.1. The prolongation $\mathbf{Q}_\tau^{o\,*}[\bullet]$ generated by $\omega_\tau^{o\,*}(\bullet)$ may be given by the relation

$$\mathbf{Q}_\tau^{o\,*}|\bullet| = \bigcap_L \left[q^*(\bullet \, , \tau \, , q_\tau^*) + Z_L^*(\bullet \, , \tau \, , Q^*|\tau| - q_\tau^*)\right] \tag{4.3}$$

over all constant matrices $L \in \mathbf{R}^{m \times n}$.

It is not difficult to observe that the following relation is true

Lemma 4.2. The union of all possible cross sections $\mathbf{Q}^* \, |t_1|$ of the prolongation $\mathbf{Q}_\tau^{o\,*}|\bullet|$ of $\mathbf{Q}^*[\tau]$ (over all triplets $\omega_\tau^*(\bullet)$ that satisfy (4.1)), is a convex compact set - the attainability domain $Q(t_1 \, , \tau \, , \mathbf{Q}^*[\tau])$ at time t_1 for the inclusion (1.1), starting from $\{\tau \, , \mathbf{Q}^*[\tau]\}$. Namely has a solution for any

$$p(t_0) = p^0 \in \mathbf{R}^n \ , \ q(t_0) = q^0 \in \mathbf{R}^{n.}$$

For the solution to (3.5)-(3.7) to exist, in the sense that (3.5) - (3.7) are satisfied for almost all $t \in [t_o \ , t_1]$, it suffices that $\mathbf{U}(t \ , p \ , y_t(\bullet))$ is a convex compact valued map, measurable in t and upper semicontinuous in $\{p \ , y_t(\bullet)\} \in \mathbf{R}^n \times \mathbb{L}_2 \ (t_0 \ , t)$, and that $\mathbf{P}(t)$, $\mathbf{R}(t)$ are of convex compact values and measurable in t, [8]. A strategy $\mathbf{U}(t \ , p \ , y_t(\bullet))$ that ensures the existence of a solution to (3.5) - (3.7) will be further referred to as an *admissible strategy.*

The Basic Problem

With mapping $K(t)$ and set M being given, specify a feedback control strategy

$$\mathbf{U} = \mathbf{U}(t \ , p \ , y_t(\bullet))$$

that would ensure the inclusions (3.2), (3.3) whatever is the realization $q(t)$ of the system (3.6), with $q(t_o) \in Q^{(0)}$ and set $Q^{(0)}$ given.

Thus the control problem is to adapt the process $p(t)$ to the uncertain state constraint:

$$p(t) \in \mathbf{K}(t) \stackrel{.}{-} \mathbf{Q}(t \ , t_o \ , \mathbf{Q}^{(0)})$$

where $\mathbf{Q}(t \ , t_0 \ , Q^{(0)})$ is achieved through a *guaranteed estimation process* for the system (3.6), (3.7) and $K \stackrel{.}{-} Q$ stands for the geometrical (Minkowski) difference of sets K , Q $(K \stackrel{.}{-} Q = \{p : p + Q \subseteq K\})$

The *information* on the *basic system* (3.1) is *complete* since the exact value of the vector p is assumed to be available.

We shall now proceed with the formal solution schemes for constructing the desired strategy

$$\mathbf{U} = \mathbf{U}(t \ , p \ , y_t(\bullet)) \ .$$

4. The Extrapolation Problem

Assume that at instant τ a realization $y_\tau^*(\bullet)$ is given and therefore, a set $\mathbf{Q}^*[\tau] = \mathbf{Q}(t \ , t_0 \ , Q^{(0)} \mid y_\tau^*(\bullet))$ is available. (From now on we will start to vary $y_\tau(\bullet)$ and will therefore include $y_\tau(\bullet)$ into the respective notations, substituting $\mathbf{Q}(\tau \ , t_0 \ , Q^{(0)})$ for $\mathbf{Q}(t \ , t_0 \ , Q^{(0)} \mid y_\tau(\bullet))$.

Suppose that the realization $y_T^*(\bullet)$ may be prolongated onto the interval $(\tau \ , t_1]$ in the form of a possible future measurement $y_\tau^*(\bullet)$ generated by a triplet

$$\omega_\tau^{o\,*}(\bullet) = \{q^* \ , v_\tau^{o\,*}(\bullet) \ , \xi_\tau^{o\,*}(\bullet)\}$$

where our further notation will be taken in the form $\psi_t^o(\sigma) = \psi(t + \sigma)$, $o < \sigma \leq t_1 - t$, so that the upper zero index would assign the respective element $\psi_t^o(\bullet)$ to the interval $(t \ , t_1]$. For a multivalued map $\Psi(t)$ the notation is similar $\Psi_t^o(\sigma) = \Psi(t + \sigma)$, $0 < \sigma \leq t_1 - t$.

For $t + \Delta t$ the element $y_t(\bullet)$ to be compared with $y_{t + \Delta t}(\bullet)$ should be modified to $y_t^\Delta(\bullet)$ which will be defined for $[t_0 \ , t + \Delta t]$ and such that

$$y_t^\Delta \ (\sigma) = \begin{cases} y \ (t + \sigma) & t_0 - t \leq \sigma \leq 0 \\ y \ (t) \ , & 0 < \sigma \leq \Delta t \end{cases}$$

$$\bigcup \{ \mathbf{Q}^*[t_1] \mid \omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^o(\bullet)\} = Q(t_1 , \tau , \mathbf{Q}^*[\tau]) ,$$

The schemes of the above allow to construct a solution procedure for the *basic problem*.

5. The Solution Scheme

Suppose that the position (the "state") of the overall system is given as

$$\{\tau , p , y_\tau(\bullet)\}$$

or in equivalent form as

$$\{\tau , p , \mathbf{Q}[\tau]\}$$

where

$$\mathbf{Q}[\tau] = \mathbf{Q}(\tau , t_0 , Q^{(0)} \mid y_\tau(\bullet))$$

A possible prolongation for $\mathbf{Q}[\tau]$ onto $(\tau , t_1]$ is the multivalued function $\mathbf{Q}_\tau^{o\,*}[\bullet]$ generated due to a possible "future" measurement $y_\tau^{o\,*}(\bullet)$ (which is uniquely defined by a triplet

$$\omega_\tau^{o\,*}(\bullet) = \{q^* , v_\tau^{o\,*}(\bullet) , \xi_\tau^{o\,*}(\bullet)\} , \omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^{o\,*}(\bullet))$$

Returning to an inverse problem of the type described in § 2, (except that system (2.1) is changed to (4.1) and sets $\mathbf{N}$, $\mathbf{Y}(t)$ to $\mathbf{M}$ and $\mathbf{K}(t) \dot{-} \mathbf{Q}^*[\tau]$, respectively), we observe that the set

$$W(\tau , t_1 \, M , \mathbf{Q}[\tau] \mid \omega_\tau^{o\,*}(\bullet)) = W(\tau , t_1 \, \mathbf{M} , \bullet \mid \omega_\tau^{o\,*}(\bullet))$$

consists of states $\{\tau , p\}$ such that for each of these there exists and "open-loop" control $u(t)$ that steers $\{\tau , p\}$ into M under the constraints

$$u(t) \in \mathbf{V}(t) , \quad p(t) + \mathbf{Q}[t] \in \mathbf{K}(t)$$
$$\tau \leq t \leq t_1$$

In view of Lemma 2.1 we come to

Lemma 5.1. The set $W(\tau , t_1 \, \mathbf{M} , \mathbf{Q}[\tau] \mid \omega_\tau^{o\,*}(\bullet))$ *may be described as*

$$W(\tau , t_1 , \mathbf{M} , \mathbf{Q}[\tau] \mid \omega_\tau^{o\,*}(\bullet)) =$$
$$= \bigcap \{ W_L(\tau , t_1 , M , \mathbf{Q}[\tau] \mid \omega_\tau^{o\,*}(\bullet)) \mid L_\tau(\bullet)\} \tag{5.1}$$

the intersection being taken over all continuous $(n \times n)$- matrix-valued functions $L(t)$ defined for $[\tau , t_1]$.

Here $W_L[\tau] = W(\tau , t_1 , \mathbf{M} , \mathbf{Q}[\tau] \mid \omega_\tau^{o\,*}(\bullet)) = W(\tau , t_1 , \mathbf{M} , \bullet \mid \omega_\tau^{o\,*}(\bullet))$
is the solution set to the system

$$\dot{w}_L \in (C(t) - L(t)) \, w_L + L(K(t) \dot{-} \mathbf{Q}^*[\tau]) + \mathbf{V}(t) \tag{5.2}$$
$$w_L(t_1) \in \mathbf{M}$$

or to the funnel equation

$$\lim_{\sigma \longrightarrow o} \sigma^{-1} h^+ (W[t - \sigma] - L\mathbf{Q}[t]\sigma , (E - \sigma(C(t) - L(t)) \, W[t] - L\mathbf{K}(t)\sigma - \mathbf{V}(t)\sigma) = 0 \tag{5.3}$$

$$W_L[t_1] = \mathbf{M}$$

The next step is to construct a set $W(\tau , t_1 , \mathbf{M} , \bullet)$ of such states $\{\tau , p\}$ that for every possible prolongation $\mathbf{Q}^*[t]$ (generated by $\omega_\tau^{o\,*}(\bullet)$) there exists an "open-loop" control $u(t)$ that steers $\{\tau , p\}$ into $\mathbf{M}$ under the constraints (5.1).

Lemma 5.2. The set $W(\tau , t_1 , \mathbf{M} ,\bullet)$ may be described as

$$W(\tau , t_1 ,\mathbf{M} , \bullet) = \bigcap \{ \; W(\tau , t_1 ,M , \bullet \mid \omega_\tau^{o\,*}(\bullet)) \mid \omega_\tau^{o\,*} \in \Omega_\tau(\bullet)\}$$

over all admissible triplets $\omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^o(\bullet)$

The graph of each of the multivalued maps $W_\tau^{o\,*}[\bullet]$ over the interval $[\tau , t_1]$ is closed, with convex cross-sections $W^*[t] = W(t , t_1 , M , \bullet \mid \omega_\tau^{o\,*}(\bullet))$, [7]. Therefore we come to

Lemma 5.3. The graph of the multivalued map $W_\tau[\bullet]$ is a closed set with convex cross-sections $\mathbf{W}[t] = W(t , t_1 , M , \bullet)$, $t \in [\tau , t_1]$.

With $W[\tau]$ given, the regular *extremal strategy* that follows the scheme of [1,3] is constructed through the relation

$$U(\tau , p , y_\tau(\bullet) = \begin{cases} \mathbf{V}(\tau) & \text{if } p \in \mathbf{W}[\tau] \\ \partial\rho(l \mid \mathbf{V}(\tau)) \;, \; l \in \partial\, d(p \;, \; W[\tau]) \;, & \text{if } p \,\overline{\in}\, \mathbf{W}[\tau] \end{cases} \tag{5.4}$$

where

$$d(p \;, \; W[\tau]) = \min\{|| \; p - w || \mid w \in \mathbf{W}[\tau]\}$$

is the Euclidean distance from p to $W[\tau]$, and $\partial f(l)$ is the *subdifferential* of the function f at point l.

For the function $\psi(p) = d(p \;, \; W)$, the subdifferential

$$\partial\, \Psi(p) = \partial d(p \;, \; W)$$

consists of a single point $w^* = \arg\min \{||p - w \;|| \;|w \in W[\tau]\}$,

The *regular extremal strategy* of (5.4) yields the solution to the basic problem under some *additional assumptions.*

Consider the support function

$$\rho(l \mid W(\tau , t_1 , \mathbf{M} , \bullet \mid \omega_\tau^{o\,*}(\bullet)))$$

and further on, the function

$$f(l \mid \tau , t_1 , M , \mathbf{Q}[\tau]) = f(l \mid \tau , t_1 , M \;, \bullet) =$$

$$=\inf \{\rho(l \mid \; W(\tau , t_1 , \mathbf{M} , \bullet \mid \omega_\tau^{o\,*}(\bullet))) \mid \omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^o(\bullet)\}$$

Lemma 5.4. The function $f(l \mid \tau , t_1 , M , \bullet)$ is a closed positively homogeneous function.

Assumption 5.1. Whatever the realization $\mathbf{Q}[\tau]$, the following relation is true

$$f(l \mid \tau , t_1 , \mathbf{M} , \bullet) = f^{**}(l \mid \tau , t_1 , \mathbf{M} , \bullet) > -\infty \tag{5.5}$$

where $f^{**}(l \mid \tau , t_1 , M , \bullet)$ is the second conjugate to $f(l \mid \tau , t_1 , \mathbf{M} , \bullet)$ in the variable l.

The second conjugate ([15]) to a function $f(l)$ is defined as $(f^*)^*(l)$ where $f^*(p) = \sup\{(p , l) - f(l) \mid l \in \mathbb{R}^n\}$

In other words, Assumption 5.1 requires that $f(l \mid \tau , t_1 , \mathbf{M} , \bullet)$ would be convex and lower semi-continuous in l.

This yields

$$f(l \mid \tau , t_1 , \mathbf{M} , \bullet) = \rho(l \mid W(\tau , t_1 , \mathbf{M} , \bullet))$$

Hence, under Assumption 5.1, the support function $\rho(l \mid \mathbf{W}(\tau , t_1 , \mathbf{M} , \bullet))$ of the intersection of sets $\mathbf{W}(\tau , t_1 , \mathbf{M} , \bullet) \mid \omega_\tau^{o\,*}(\bullet))$ (over $\omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^o(\bullet)$) should coincide with

$$\inf\{\rho(l \mid W(\tau , t_1 , M , \bullet \mid \omega_\tau^{o\,*}(\bullet)) \mid \omega_\tau^{o\,*}(\bullet) \in \Omega\}$$

This is a requirement which does not hold in the general case where the support function of the intersection of sets requires an *infimal convolution* of the respective supports rather than their infimum, |15|.

Lemma 5.5. Under Assumption 5.1., the multivalued map $W_\tau^o|\bullet|$ has a closed graph with convex compact cross-sections $W|t| = W(t , t_1 , M , \bullet)$.

Lemma 5.6. Under Assumption 5.1., the strategy $\mathbf{U}(\tau , p , y_\tau(\bullet))$ of (5.4) is an admissible strategy.

Theorem 5.2. Suppose the vector $p^0 = p(t_0)$ and the set $Q(t_0) = Q^{(0)}$ are such that Assumption 5.1 is true and that

$$p^0 \in W(t_0\, t_1 , \mathbf{M} , Q^{(0)})$$

Then the respective strategy $U(t , p , y_t(\bullet))$ of (5.4) will ensure the restrictions (3.2), (3.3) whatever are the solutions to the inclusions (3.5)-(3.7).

The regular case described here does not cover all the possible situations that may arise in the basic problem. We will therefore give a short description of two other "extremal" cases for the solution.

6. The "Blunt" Solution

Consider the attainability domain $Q(t , t_0 , Q^{(0)})$ for system (1.1) in the absence of any state constraints.

Assumption 6.1. The set $\mathbf{S}(t) = \mathbf{K}(t) \dot{-} \mathbf{Q}(t , t_0 , Q^{(0)}) \neq \emptyset$ for any $t \in |t_0 , t_1|$.

Denote $W_b|t| = W_b(\tau , t_1 , M)$ to be the solution of an inverse problem of the type given in § 2 - the set of all states $p_\tau = p(\tau)$ of system (3.1) such that for each of these there exists an open-loop control $u(t)$ $(u_\tau^o(\bullet) \in \mathbf{V}_\tau^o(\bullet))$ that ensures the inclusions

$$p(t_1 , \tau , p_\tau) \in \mathbf{M} \tag{6.1}$$

$$p(t , \tau , p_\tau) \in Q(t , \tau , \mathbf{Q}(\tau , t_0 , Q^{(0)})) , \quad \tau \le t \le t_1$$

Denote the "blunt" strategy to be

$$U_2(t , p) = \begin{cases} \mathbf{V}(t) & \text{if } p \in W_b(t , t_1 , M) \\ \partial\rho(l \mid \mathbf{V}(t)) \ , \ l \in \partial\, d(p , W_b|\tau|) & \text{if } p \,\overline{\in}\, W_b(t , t_1 , M) \end{cases} \tag{6.2}$$

Lemma 6.1. The strategy $U_b(t\ ,\ p)$ ensures the solution to the inclusion

$$p \in C(t)\, p + \mathbf{U}_b(t\ ,\ p)\ ,\ t_0 \leq t \leq t_1 \tag{6.3}$$

for any initial state $p(t_0) = p^0$.

The solution is here understood in the sense of Caratheodory [9].

Theorem 6.1. Under Assumption 6.1 suppose $p(t_0) \in W(t_0\ ,\ t_1\ ,\ \mathbf{M})$. Then the strategy $\mathbf{U}_b(t\ ,\ p)$ of (6.2) ensures that any solution $p(t\ ,\ t_0\ ,\ p^0)$ to the differential inclusion (6.3) would satisfy the restrictions (6.1).

The "blunt" solution does not require any on-line measurements for the uncertain system (1.1). It implements an "open-loop" feedback solution under a given state constraint and it may work only if the sets $\mathbf{S}(t)$ are nonvoid, which is a rather strong restriction on the parameters of the problem.

7. The General Approach

The general approach leads to a complicated scheme that follows the constructions of [2], [3] and [7].

Suppose a set $\mathbf{Q}(\tau)$ is given and

$$\mathbf{Q}(\bullet\ ,\ t_0\ ,\ \mathbf{Q}[\tau]\ ,\ \omega_\tau^{o\,*}(\bullet))\ ,\ \omega_\tau^{o\,*}(\bullet) \in \Omega_\tau^o(\bullet)\ ,$$

are the possible realizations of the informational sets (due to possible "future" measurements).

The sequence of operations is as follows. Divide the interval $[\tau\ ,\ t_1]$ into s subintervals

$$\tau = t^0\ ,\ t^1\ ,\dots,\ t^s = t_1\ ,$$
$$\max \mid t^i - t^{i-1} \mid = \epsilon_s$$

For the interval $(t^s\ ,\ t_1]$ find the set

$$W_s(t^{s-1}\ ,\ t_1\ ,\ \mathbf{M}\ ,\ \mathbf{Q}[t^{s-1}] \mid \omega_{t^{s-1}}^{o\,*}(\bullet))\ .$$

Take

$$\mathbf{W}_s(t^{s-1}\ ,\ t_1\ ,\ \mathbf{M}) = \bigcap\ \{\bigcap\ W_s(t^{s-1}\ ,\ t_1\ ,\ \mathbf{M}\ ,\ \mathbf{Q}[t^{s-1}] \mid \omega_{t^{s-1}}^{o\,*}(\bullet)) \mid$$
$$\mid \omega_{t^{s-1}}^{o\,*}(\bullet)) \in \Omega_{t^{s-1}}^{o}(\bullet)\} \mid \mathbf{Q}[t^{s-1}] = \mathbf{Q}(t^{s-1}\ ,\ t_0\ ,\ Q^{(0)} \mid y_\tau^*(\bullet)) : \omega_{t^{s-1}}^{*}(\bullet) \in \Omega_{t^{s-1}}(\bullet)\}$$

Repeat this procedure for $(t^{s-2}\ ,\ t^{s-1}]$, taking $\mathbf{W}_s(t^{s-1}\ ,\ t_1\ ,\ M)$ instead of M.

In a similar way continue to repeat this procedure for $(t^{s-3}\ ,\ t^{s-2}]$ taking $\mathbf{W}_s(t^{s-2}\ ,\ t^{s-1}\ ,\ \mathbf{W}_s(t^{s-1}\ ,\ t_1\ ,\ M)$ instead of M and so on, finally arriving at

$$\mathbf{W}_s(\tau\ ,\ t_1\ ,\ \mathbf{M}) = \mathbf{W}_s(\tau\ ,\ t^1\ ,\ \mathbf{W}_s\ (t^1\ ,\ t^2\ ,\ \dots\mathbf{W}_s(t^{s-1}\ ,\ t_1\ ,\ \mathbf{M}))\dots)$$

Under rather conventional conditions with $s \longrightarrow \infty\ ,\ \epsilon_s \longrightarrow 0$, the set $\mathbf{W}_s(\tau\ ,\ t_1\ ,\ \mathbf{M})$ will converge

$$\mathbf{W}_s(\tau\ ,\ t_1\ ,\ \mathbf{M}) \longrightarrow \mathbf{W}(\tau\ ,\ t_1\ ,\ \mathbf{M})$$
$$s \longrightarrow \infty\ ,\ \epsilon_s \longrightarrow 0$$

in the Hausdorff metric, and the set-valued function $\mathbf{W} = \mathbf{W}(\tau , t_1 , \mathbf{M})$ may then serve as a basis for a strategy similar to $\mathrm{U}(t , p , y_t(\bullet))$. The detailed treatment of this situation will be the subject of another paper.

A final remark is that the numerical implementation of this scheme requires an appropriate *approximation theory for set-valued maps.* Therefore an approximative scheme that traces the basic solutions in terms of ellipsoidal valued functions seems to be a relevant subject for investigation.

References

[1] Krasovskii, N.N. *The Control of a Dynamic System,* Nauka, Moscow, 1986.

[2] Pontriagin, L.S. *Linear Differential Games of Pursuit,* Mat. Sbornik, 112 (154) No.3(7), 1980.

[3] Krasovskii, N.N. and Subbotin, A.I. *Positional Differential Games,* Nauka, Moscow, 1976.

[4] Varaiya, P. *On the Existence of Solutions to a Differential Game,* Siam J. Control, 5, No.1, 1967.

[5] Friedman, A. *Differential Games,* New York, Wiley, 1971.

[6] Schweppe, F.C. *Uncertain Dynamic Systems,* Prentice Hall Inc., Englewood Cliffs, N.J., 1973.

[7] Kurzhanski, A.B. *Control and Observation Under Uncertainty Conditions,* Nauka, Moscow, 1978.

[8] Kurzhanski, A.B., Nikonov, O.I. *On Adaptive Processes of Guaranteed Control,* Izvestia Akad. Nauk SSSR, Engineering Cybernetics, No.4, 1986.

[9] Aubin, J.-P., Cellina, A. *Differential Inclusions,* Springer Verlag, Heidelberg, 1984.

[10] Leitman, G., Corless, M. *Adaptive Control for Uncertain Dynamical Systems,* Dynamic Systems and Microphysics, Control Theory and Mechanics, Acad. Press. Inc., 1984.

[11] Panasjuk, A.I. and Panasjuk, V.I. Zametki, Vol. 27, No.3, 1980.

[12] Kuratowski, K. Topologie, Vol I (Warsaw 1948) and Topologie, Vol. II (Warsaw 1950)

[13] Kurzhanski, A.B. and Filippova, T.F. *On the Analytical Description of the Set of Viable Solutions of a Control System,* Differencialniye Uravneniya (Differential Equations), No.8, 1987.

[14] Kurzhanski, A.B. *On the Analytical Description of the Pencil of Viable Trajectories of a Differential System,* Sov. Math. Doklady, Vol.33, No.2, 1986

[15] Rockafellar, R.T. *Convex Analysis,* Princeton Univ. Press, 1979.

Filtering and Control for Wide Bandwidth Noise and 'Nearly' Linear Systems

H.J. Kushner
Lefschetz Center for Dynamical Systems
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

and

W. Runggaldier
Institute of Analysis
University of Padua
Padua, Italy

Abstract

Typically, modern stochastic control theory uses ideal white noise driven systems (Itô equations), and if the observed data is corrupted by noise, that noise is usually assumed to be 'white Gaussian'. If the models are linear, a Kalman-Bucy filter is then used to estimate the state, and a control based on this estimate is computed. Actually, the noise processes are rarely 'white', and the system is only approximated in some serse by a diffusion. But, owing to lack of 'computable' alternatives, one still uses the above procedure. Then the 'filter' estimates and associated control might be quite far from being optimal. We examine the sense in which such estimates and/or control are useful, in order to justify the the use of the commonly used procedure. For the filtering problem where the signal is a 'near' Gauss-Markov process and the observation noise is wide band, it is shown that the usual filter is 'nearly optimal' with respect to a very natural class of alternative data processors. The asymptotic (in time and bandwidth) problem is treated, as is the conditional Gaussian case. Similar results are obtained for the combined filtering and control problem, where it is shown that good controls for the 'ideal' model are also good for the actual physical model, with respect to a natural class of alternative controls, for control over a finite time interval and the average cost per unit time problem.

The paper is an outline of some of the work reported in [9].

I. Introduction

Typical models in modern control and filtering theory are of the following type, where $W(\cdot)$ are standard Wiener processes, $u(\cdot)$ is a control, and b_z, σ_z, etc., are appropriate functions. We let $z(\cdot)$ denote a reference signal, $x(\cdot)$ the control system, $\bar{y}(\cdot)$ the noise corrupted observation and $r_T(u)$ and $\gamma(u)$ the cost functions.

$$dz = b_z(z)dt + \sigma_z(z)dW_z \tag{1.1}$$

$$dx = b_x(x,u)dt + \sigma_x(x)dW_x \tag{1.2}$$

$$dy = h(x,z)dt + dW_y \tag{1.3}$$

$$r_T(u) = \int_0^T E\ k(x(s),\ z(s),\ u(s))ds \tag{1.4}$$

$$r(u) = \overline{\lim_{T\to\infty}}\ r_T(u)/T\ . \tag{1.5}$$

The actual physical system, which we denote by $z^\epsilon(\cdot)$, $x^\epsilon(\cdot)$, $\bar{y}^\epsilon(\cdot)$ is not of the form (1.1) - (1.3). The reference signal $z^\epsilon(\cdot)$ might be only approximately representable by (1.1), and the noise in the control and observation system would rarely be 'white'. But, via some approximation or identification procedure, one chooses a model of the form (1.1) - (1.3), then computes a good control for that model, and then applies this control to the actual physical system. One must question the value of the filter output and the determined control when applied to the 'physical' problem.

The filter output might not be even nearly optimal for use in making estimates of $z^\epsilon(\cdot)$, and the control (based on the filter outputs) will rarely be 'nearly optimal'. Such questions are basic to the relevence of much theoretical work. We will deal with these questions here, when the approximating system (1.1), (1.2) is linear - for which a fairly complete theory can be obtained.

Owing to the usual lack of 'near optimality' (when applied to the physical system) of the filter and control which is obtained by using (1.1) - (1.3), one should ask the question: with respect to which alternative filters (called 'data processors' below) or controls for the physical system are the chosen ones nearly optimal? It turns out that this alternative class controls is quite large and quite reasonable. The basic mathematical techniques used here are those of the theory of weak convergence of probability measures [1], [3], [4], a technique which is quite useful for problems in the approximation of random processes [1], [5] - [8], [12], [13].

When the ideal model is linear - one would usually use the Kalman-Bucy filter appropriate for the ideal model, but whose input is the physical observation. Obviously, the filter does not usually yield the conditional distribution of the $z^\epsilon(t)$ given the data $y^\epsilon(s)$, $s \leq t$. In Section 2, we discuss some counter examples to illustrate the sort of difficulties which arise in such approximations, and in Section 3 the approximation theorem is given, together with the class of alternative data processors. Section 4 concerns the average filter error per unit time - or the errors for large time. The combined filtering and control problem is dealt with in Sections 5 and 6. The optimal control for (1.1) - (1.3) will be nearly optimal for the physical system - in comparison with a large class of alternative controls. The symbol ⇒ denotes weak convergence. A fuller development appears in [9], together with the conditional Gaussian case and a treatment of certain non-linear observations. For the weak convergence, we work with the space $D^k[0,\infty)$, the space of R^k-valued functions which are right continuous and have left-hand limits, and endowed with the Skorohod topology. (See [1], [3], [4].) Reference [2] deals with similar approximations for the non-linear filtering problem, and reference [10] concerns the approximation problem for the non-linear control problem. Here, owing to the linearity, we can do both 'approximate' control and filtering simultaneously. The models and results are formulated so that the paper is not burdened with a large amount of weak convergence theory. There are extensions in many directions: discrete parameter problems, impulsive control, etc.

2. Linear Filtering: Preliminaries

Consider the following filtering problem: For each $\epsilon > 0$, $z^\epsilon(\cdot)$ is a *signal* process, $\xi_y^\epsilon(\cdot)$ is a 'wide-bandwidth' observation noise, and the two are mutually independent. The actual observation process is:

$$\dot{y}^\epsilon(t) = H_z z^\epsilon(t) + \xi_y^\epsilon(t),\ y^\epsilon(0) = 0\ . \tag{2.1}$$

All 'noise' processes are assumed to be right continuous and have left-hand limits. Define $y^\epsilon(t) = \int_0^t \dot{y}^\epsilon(s)ds$ and $W_y^\epsilon(t) = \int_0^t \xi_y^\epsilon(s)ds$. Let $z(\cdot)$ satisfy (for matrices A_z , etc.)

$$dz = A_z z dt + B_z dW_z, \tag{2.2}$$

Since $\xi_y^\epsilon(\cdot)$ is to be 'nearly' white noise, and $z^\epsilon(\cdot)$ 'nearly' a Gauss-Markov diffusion, let

$$(z^\epsilon(\cdot), W_y^\epsilon(\cdot)) \Rightarrow (z(\cdot), W_y(\cdot)) \quad \text{as} \quad \epsilon \to 0\ , \tag{2.3}$$

where $W_y(\cdot)$ is a non-degenerate Wiener process. The $W_z(\cdot)$ and $W_y(\cdot)$ must be independent. Also $y^\epsilon(\cdot) \Rightarrow y(\cdot)$, where

$$dy = H_z z dt + dW_y\ ,\ y(0) = 0\ . \tag{2.4}$$

The actual physical system is, of course, 'fixed' and corresponds to some small $\epsilon > 0$. The use of weak convergence here is just a way of embedding the *actual data* in a sequence - so that an approximation method can be used. The approximation of the values of expectations of functions of $z^\epsilon(\cdot)$, conditioned on the data $y^\epsilon(\cdot)$ is not easy in general. Furthermore, we cannot restrict ourselves to Gaussian noise, since it itself is only an approximation to the physical processes.

For (2.2), (2.4), the filter equations are

$$d\hat{z} = A_z \hat{z}\,dt + Q(t)\,[dy - H_z \hat{z}\,dt] \tag{2.5}$$

$$Q(t) = \Sigma(t)H_z' R_0^{-1}$$

$$\dot{\Sigma} = A_z\Sigma + \Sigma A_z' + B_z B_z' - \Sigma H' R_0^{-1} H\Sigma\ , \tag{2.6}$$

where R_0 = covariance matrix of observation 'noise' $W_y(t)$, which we set to I, unless mentioned otherwise. In practice, with signal $z^\epsilon(\cdot)$ and noise $\xi_y^\epsilon(\cdot)$, one normally uses (2.6) and (2.5_{WB}):

$$\dot{\hat{z}}^\epsilon = A_z \hat{z}^\epsilon + Q(t)\,[\dot{\bar{y}}^\epsilon - H_z \hat{z}^\epsilon]\ . \tag{2.5_{WB}}$$

This system is not necessarily even a nearly optimal filter for the physical observation. But, as will be seen, it makes a great deal of sense and is quite appropriate in a specific but important way.

Some illustrations will illustrate the problems that we must contend with, particularly concerning the *possible lack of continuity in the optimal estimators* as the noise bandwidth goes to ∞. Let (X_n,Y_n) be bounded real-valued random variables which converge in distribution to (X,Y). Generally $E(X_n|Y_n) \not\to E(X|Y)$. For example, let $X_n = X$, $Y_n = X/n$. Next, let $Z_n = Z_n(Y)$, where Y is a random varible and $(Z_n,Y) \Rightarrow (Z,Y)$. Then Z is *not* necessarily a function of Y, and might even be independent of Y, as illustrated by the following:

Let Y be uniformly distributed on $[0,1]$. Define $Z_n = nY$ for $0 \leq Y < 1/n$ and, in general, define $Z_n = (nY - k)$ on $k/n \leq Y < (k+1)/n$, $k = 0,1,\ldots,n-1$. Then $(Z_n,Y) \Rightarrow (Z,Y)$ where Z is independent of Y, and both Z and Y are uniformly distributed on $[0,1]$. Clearly $E(Z_n|Y) \not\to E(Z|Y)$ in any sense.

Even though $W_y^\epsilon(\cdot) \Rightarrow W_y(\cdot)$, a non-degenerate Wiener process, $y^\epsilon(\cdot)$ might contain a *great deal* more information about $z^\epsilon(\cdot)$ than $y(\cdot)$ does about $z(\cdot)$. See [9] for an example where as $\epsilon \to 0$, we can calculate $z^\epsilon(t)$ nearly exactly from the data $y^\epsilon(\cdot)$. In general we have

Let $(X_n,Y_n) \Rightarrow (X,Y)$ (X_n*-real valued,* Y_n *with values in* R^s). *Then*

$$\overline{\lim_n}\ E[X_n - E(X_n|Y_n)]^2 \leq E[X - E(X|Y)]^2\ . \tag{2.7}$$

In the above examples, the inequality is strict. The examples do caution us to take considerable care in dealing with information processing with wide bandwidth noise disturbances.

3. The 'Approximately Optimal' Linear Filtering Problem

For the ideal filtering problem (2.2), (2.4), the optimal decisions are functions of $\hat{z}(\cdot)$, $\Sigma(\cdot)$, since these completely determine the conditional distribution. There are no functions of the data which give better estimates. This is *not so* with estimates based on $\Sigma(\cdot)$, $\hat{z}^{\epsilon}(\cdot)$ for the system $z^{\epsilon}(\cdot)$, $y^{\epsilon}(\cdot)$. We now define a class of functions of the observed data $y^{\epsilon}(\cdot)$ with respect to which functions of $\hat{z}^{\epsilon}(\cdot)$, $\Sigma(\cdot)$ *are* 'nearly optimal' for small $\epsilon > 0$. We need to specify both a criterion of comparison; i.e., a cost function. Although we use one particular cost function, the general idea and possible extensions should be clear .

Let $\mathcal{D}$ denote the class of measurable functions on $C[0,\infty]$, the space of real valued continuous functions on $[0,\infty)$ (with the topology of uniform convergence on bounded intervals), which are continuous w.p.1 relative to Wiener measure (hence, with repect to the measure of $y(\cdot)$). Let $\mathcal{D}_t$ denote the subclass which depends only on the function values up to time t. For arbitrary $F(\cdot) \in \mathcal{D}$ or in $\mathcal{D}_t$, we will use $F(y^{\epsilon}(\cdot))$ as an *alternative estimator* of a functional of $z^{\epsilon}(\cdot)$. The class is quite large.

First, note that $\mathcal{D}$ contains all continuous functions and that the $\hat{z}(\cdot)$ of (2.5) can be written as a continuous function of the integral of the driving force $y(\cdot)$. Thus, continuous functions of $\hat{z}^{\epsilon}(\cdot)$ are admissible estimators. Many important functionals are only continuous w.p.1 (relative to Wiener measure). Let $\tau(x(\cdot))$ denote the first time that a closed set A with a piecewise differential boundary is reached by $x(\cdot)$. Then the function with values $T \cap \tau(x(\cdot))$ is in $\mathcal{D}_T$ for any $T < \infty$. Thus, our alternative estimators can involve stopping times. This is essential in sequential decision problems, since there the cost function involves first entrance times of a function of $y(\cdot)$ into a decision set.

$\mathcal{D}$ and $\mathcal{D}_t$ do not contain 'wild' functions such as those involving differentiation. We consider $\mathcal{D}$ and $\mathcal{D}_t$ as a class of *data processors.* It seems to contain a large enough class for practical applications when the corrupting noise is 'white'.

We now state the 'model' 'robustness' or 'approximation' result. For a function $q(z)$, we write (P_t^{ϵ},q) for the integral of $q(z)$ with respect to the *Gaussian distribution* with mean $\hat{z}^{\epsilon}(t)$ and covariance $\Sigma(t)$ - the *ersatz conditional measure* of $z^{\epsilon}(\cdot)$.

The theorem states that (for a small ϵ) the ersatz conditional distribution is 'nearly optimal' with respect to a specific (but broad) class of alternative estimators. The alternative class includes those that make sense to use when the corrupting noise is white. If the noise is wide band, then it might not make sense to exploit its detailed structure and use other 'better' estimators. Doing so might, in practical cases, cause processing errors and other (unmodelled) noise effects.

Theorem 3.1. *Assume the conditions on* $z^{\epsilon}(\cdot)$, $W_y^{\epsilon}(\cdot)$ *of Section* 2. *Then* $(\hat{z}^{\epsilon}(\cdot), z^{\epsilon}(\cdot), W_y^{\epsilon}(\cdot)) \Rightarrow (\hat{z}(\cdot), z(\cdot), W_y(\cdot))$. *Let* $F(\cdot) \in \mathcal{D}_t$ *be bounded, and* $q(\cdot)$ *bounded continuous and real valued. Then (the limits all exist)*

$$\lim_{\epsilon} E[q(z^{\epsilon}(t)) - F(y^{\epsilon}(\cdot))]^2 \tag{3.1}$$
$$\geqslant \lim_{\epsilon} E[q(z^{\epsilon}(t)) - (P_t^{\epsilon},q)]^2.$$

Remark. The assertion concerning the weak convergence is necessary, since we need to know that the limit of the cited ϵ-triple represents a true filtering problem. The result would not make sense if only 2 out of the 3 components converged.

Proof. By the weak convergence and the w.p.1 continuity of $F(\cdot)$,

$$\left[q(z^{\epsilon}(t)), F(y^{\epsilon}(\cdot)), (P_t^{\epsilon},q)\right] \Rightarrow \left[q(z(t)), F(y(\cdot)), (P_t,q)\right],$$

where $(P_t,q) = \int q(z)dN(\hat{z}(t), \Sigma(t);dz)$, and $N(\hat{z},\Sigma;\cdot)$ is the normal distribution with mean $\hat{z}$ and covariance Σ. Thus, the left and right sides of (3.1) converge to, respectively,

$$E[q(z(t)) - F(y(\cdot))]^2, \quad E\left[q(z(t)) - E[q(z(t))\,|\,y(s), s \leqslant t]\right]^2. \tag{3.2}$$

Since the conditional expectation is the optimal estimator, the second expression is no greater than the first. This yields the theorem.

Q.E.D.

4. Filtering the Large Time Problem (*Large* t, *small* ϵ)

The filtering system often operates over a very long time interval. For the model (2.2), (2.4), or with (2.6), (2.5_{WB}), one would then use the stationary filter. But with the system $y^{\epsilon}(\cdot)$, $z^{\epsilon}(\cdot)$, *two limits are involved* since both $t \to \infty$ and $\epsilon \to 0$, and it is important that the results *not depend on how* $t \to \infty$ and $\epsilon \to 0$, and that the use of the stationary limit filter is justified. We make some additional assumptions.

C4.1. A_z *is stable,* (A_z,H_z) *is observable and* (A_z,B_z) *controllable.*

C4.2. $\xi_y^{\epsilon}(t)$ *takes the form* $\xi_y^{\epsilon}(t) = \xi_y(t/\epsilon^2)/\epsilon$, *where* $\xi_y(\cdot)$ *is a second order stationary process with integrable covariance function* $R(\cdot)$. Also, if $t_{\epsilon} \to \infty$ as $\epsilon \to 0$, then $W_y^{\epsilon}(t_{\epsilon}+\cdot) - W_y^{\epsilon}(t_{\epsilon}) \Rightarrow W_y(\cdot)$.

Remark. The model (C4.2) is a common way of modelling wide bandwidth noise, and is used to simplify a calculation below, and to avoid the details involved with other models. It can be extended in many ways. We also make the rather unrestrictive assumption that the initial time is not important and that the $z^{\epsilon}(\cdot)$ processes do not explode:

C4.3. *If* $\{z^{\epsilon}(t_{\epsilon})\}$ *converges weakly to a random variable* $z(0)$ *as* $\epsilon \to 0$, *then* $z^\epsilon(t_\epsilon + \cdot) \Rightarrow z(\cdot)$ *with initial condition* $z(0)$. *Also*

$$\sup_{\epsilon,t} E|z^\epsilon(t)|^2 < \infty.$$

Consistency. In order that $\hat{z}(\cdot), \Sigma(\cdot)$, be a filter for $z(\cdot), y(\cdot)$, it is necessary that the *initial conditions* be *consistent*. Let $N(\hat{z},\Sigma;A)$ denote the probability that the normal random variable (with mean $\hat{z}$, and covariance Σ) takes values in the set A. By *consistency*, we mean that $P\{z(0) \in A|\hat{z}(0), \Sigma(0)\} = N(\hat{z}(0), \Sigma(0);A)$. One cannot choose the initial (random) conditions arbitrarily. It should be obvious that if $\Sigma(0) = \bar{\Sigma}$ and $(z(0), \hat{z}(0))$ are the *stationary* random variables for (stable) (2.2) and (2.5), then the initial conditions are consistent.

The question of consistency arises because when we study the asymptotics as $t \to \infty$ and $\epsilon \to 0$, we will start the filter at some large t_ϵ and do not know a-priori what the *limits* of $(\hat{z}^\epsilon(t), z^\epsilon(t))$ are. The initial condition of the limit equations must be consistent for the problem to make sense. Fortunately, they will be consistent.

Theorem 4.1. *Assume the conditions of Section* 2 *and* (C4.1) - (C4.3). *Let* $q(\cdot)$ *be bounded and continuous and let* $F(\cdot) \in \mathcal{D}_t$. *Define* $y^\epsilon(s) = 0$, *for* $s \leq 0$ *and define* $y^\epsilon(-\infty,t,\cdot)$ *to be the 'reversed' function - with values* $(0 \leq \tau < \infty)$ $y^\epsilon(-\infty,t;\tau) = y^\epsilon(t-\tau)$. *Then, if* $t_\epsilon \to \infty$ *as* $\epsilon \to 0$,

$$\{z^\epsilon(t_\epsilon + \cdot), \hat{z}^\epsilon(t_\epsilon + \cdot), W_y^\epsilon(t_\epsilon + \cdot) - W_y^\epsilon(t_\epsilon)\} \Rightarrow \tag{4.1}$$
$$(z(\cdot), \hat{z}(\cdot), W_y(\cdot))$$

satisfying (2.3), (2.5), *and* $z(\cdot), \hat{z}(\cdot)$ *are stationary. Also* (3.1) *holds in the form*

$$\lim_{\epsilon,t} E\,[q(z^\epsilon(t)) - F(y^\epsilon(-\infty,t;\cdot))]^2 \tag{4.2}$$
$$\geq \lim_{\epsilon,t} E[q(z^\epsilon(t)) - (P_t^\epsilon,q)]^2 .$$

The limit of (P_t^ϵ,q) *is the expectation with respect to the stationary* $(\hat{z}(\cdot), \bar{\Sigma})$ *system.*

Proof. Suppose that $\{\hat{z}^\epsilon(t), \epsilon > 0, t < \infty\}$ is tight. Then, by the hypothesis, $\{\hat{z}^\epsilon(t), z^\epsilon(t), \epsilon > 0, t < \infty\}$ is tight and each subsequence of $\{z^\epsilon(t_\epsilon+\cdot), \hat{z}^\epsilon(t_\epsilon+\cdot), W_y^\epsilon(t_\epsilon+\cdot) - W_y^\epsilon(t_\epsilon), t_\epsilon < \infty, \epsilon > 0\}$ has a weakly convergent subsequence with limit satisfying (2.2), (2.5). Choose a weakly convergent subsequence (with $t_\epsilon \to \infty$) also indexed by ϵ and with limit denoted by $z(\cdot), \hat{z}(\cdot), W_y(\cdot)$. Suppose, for the moment, that $z(\cdot), \hat{z}(\cdot)$ is stationary. (Clearly, $\Sigma(t) \to \bar{\Sigma}$ as $t \to \infty$.) If all limits are stationary, then the subsequence is irrelevant since the stationary solution is unique. Also, since the initial conditions of $\hat{z}(\cdot)$ and $z(\cdot)$ are consistent (owing to the stationarity), $(\hat{z}(\cdot), \bar{\Sigma})$ is the optimal filter for $y(\cdot), z(\cdot)$. Inequality (4.2) is a consequence of this and the weak convergence.

We next prove tightness of $\{\hat{z}^\epsilon(t),\ \epsilon > 0,\ t < \infty\}$, and then the stationarity will be proved. We have

$$\dot{\hat{z}}^\epsilon = [A_z - Q(t)H_z]\hat{z}^\epsilon + Q(t)\,\xi(t/\epsilon^2)/\epsilon + Q(t)H_z\ z^\epsilon(t)\ . \tag{4.3}$$

Let $\Phi(t,\tau)$ denote the fundamental matrix for $[A_z - Q(t)H_z]$. There are $K < \infty$, $\lambda > 0$ such that $|\Phi(t,\tau)| \leqslant K \exp - \lambda(t-\tau)$. We have

$$\hat{z}^\epsilon(t) = \Phi(t,0)z^\epsilon(t) + \int_0^t \Phi(t,\tau)\ Q(\tau)\,\xi(\tau/{}_\epsilon{}^2)d\tau/\epsilon$$

$$+ \int_0^t \Phi(t,\tau)Q(\tau)Hz^\epsilon(\tau)d\tau\ .$$

A straightforward calculation using (C4.2 - C4.3) and the change of variable $\tau/\epsilon^2 \to \tau$ in the first integral yields

$$E\ |\hat{z}^\epsilon(t)|^2 \leqslant \text{constant}\ (1 + E|z^\epsilon(0)|^2)\ ,$$

giving the desired tightness.

To prove the stationarity of the limit of any weakly convergent subsequence, we need only show stationarity of the limit values $(z(0),\ \hat{z}(0))$ of the $(z^\epsilon(t_\epsilon),\ \hat{z}^\epsilon(t_\epsilon))$. For this, we use a 'shifting' argument. Fix $T > 0$ and take a weakly convergent subsequence of (indexed also by ϵ, and with $t \overset{\epsilon}{\rightarrow} \infty$)

$$\{z^\epsilon(t_\epsilon+\cdot),\ \ \hat{z}^\epsilon(t_\epsilon+\cdot),\ \ W_y^\epsilon(t_\epsilon+\cdot) - W_y^\epsilon(t_\epsilon),\ \ z^\epsilon(t_\epsilon-T+\cdot),\ z^\epsilon(t_\epsilon-T+\cdot),$$

$$W_y^\epsilon(t_\epsilon-T+\cdot) - W_y^\epsilon(t_\epsilon-T)\}$$

with limit denoted by $\{z(\cdot),\ \hat{z}(\cdot),\ W_y(\cdot),\ z_T(\cdot),\ \hat{z}_T(\cdot),\ W_{y,T}(\cdot)\}$. We have $\hat{z}_T(T) = \hat{z}(0)$ and $z_T(T) = z(0)$. We do not yet know what $\hat{z}_T(0)$ or $z_T(0)$ are - but, *uniformly in* T, they belong to a tight set, owing to the tightness of $\{\hat{z}^\epsilon(t),\ \epsilon > 0,\ t < \infty\}$. Write (where $W_{z,T}(\cdot)$ 'drives' the equation for dz_T)

$$z(0) = z_T(T) = (\exp\ A_zT)z_T(0) + \int_0^T \exp\ A_z(T-\tau)\cdot B_z dW_{z,T}(\tau)$$

$$\hat{z}(0) = \hat{z}_T(T) = (\exp\ [A_z - Q(\infty)H_z]T)\hat{z}_T(0)$$

$$+ \int_0^T \exp\ [A_z - Q(\infty)H_z](T-\tau)\cdot(dW_{y,T}(\tau) + H_z z_T(\tau)d\tau)$$

Since T is arbitrary and the set of all possible $\{z_T(0)\}$ is tight, the stability of A_z and $(A_z - Q(\infty)H_z)$ implies that $z(0)$ is the stationary random variable, hence $z(\cdot)$ is stationary. Similarly, the pair $(z(\cdot), \hat{z}(\cdot))$ is stationary.

Q.E.D.

5. The Filtering and Control Problem: Finite Time Case.

As seen in the previous sections, the use of the Kalman-Bucy filter for the wide bandwidth observation noise and 'near Gauss-Markov' signal might be far from optimal, but it is 'nearly optimal' with respect to a large and reasonable class of alternative data processors. For the combined filtering and control case, the control system will be driven by wide bandwidth noise as well. Suppose that one obtains a control based on the usual ideal white noise driven limit model. This control will be a function of the outputs of the filters, and one must question the value of applying this to the actual 'wide bandwidth noise' system.

$$dz = A_z z\, dt + B_z dW_z \, . \tag{5.1}$$

Define the control system (for constant matrices A_x, D_x, B_x, H_x) by

$$\dot{x}^\epsilon = A_x x^\epsilon + D_x u + B_x \xi_x^\epsilon , \tag{5.2}$$

and let the observations be $\dot{y}^\epsilon(\cdot)$, where

$$\begin{bmatrix} \dot{y}_z^\epsilon \\ \dot{y}_x^\epsilon \end{bmatrix} \equiv \dot{y}^\epsilon = \begin{bmatrix} H_z z \\ H_x x \end{bmatrix} + \xi^\epsilon , \quad y^\epsilon \in R^g , \; y^\epsilon(0) = 0 , \tag{5.3}$$

where the three processes $\int_0^t \xi^\epsilon(s)ds \equiv W^\epsilon(t)$, $\int_0^t \xi_x^\epsilon(s)ds \equiv W_x^\epsilon(t)$ and $z^\epsilon(\cdot)$ are mutually independent, and $W^\epsilon(\cdot) \Rightarrow W(\cdot)$, $W_x^\epsilon(\cdot) \Rightarrow W_x(\cdot)$, standard Wiener processes. Thus $\xi_x^\epsilon(\cdot)$ and $\xi^\epsilon(\cdot)$ are wide bandwidth noise processes.

Define the filters and limit system:

$$\begin{bmatrix} \dot{\hat{x}}^\epsilon \\ \dot{\hat{z}}^\epsilon \end{bmatrix} = \begin{bmatrix} A_x \hat{x}^\epsilon \\ A_z \hat{z}^\epsilon \end{bmatrix} + \begin{bmatrix} D_x u \\ 0 \end{bmatrix} + Q(t) \left[\dot{y}^\epsilon - \begin{bmatrix} H_x \hat{x}^\epsilon \\ H_z \hat{z}^\epsilon \end{bmatrix} \right] , \tag{5.4}$$

$$dy = \begin{bmatrix} H_x x \\ H_z z \end{bmatrix} dt + dW = H \begin{bmatrix} x \\ z \end{bmatrix} + dw \tag{5.5}$$

$$d\begin{bmatrix} \hat{x} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} A_x\hat{x} \\ A_z\hat{z} \end{bmatrix}dt + \begin{bmatrix} D_x u \\ 0 \end{bmatrix} d(t) + Q(t)\left[dy - \begin{bmatrix} H_x\hat{x} \\ H_z\hat{z} \end{bmatrix}dt\right], \tag{5.6}$$

$$dx = A_x x dt + D_x u dt + B_x dW_x , \tag{5.7}$$

with the obvious associated Ricatti equation for the conditional covariance $\Sigma(\cdot)$ of $(x(\cdot),\ z(\cdot))$. Here $Q(t) = \Sigma(t)H'[\mathrm{cov}\ W(1)]^{-1}$. Equation (5.4) will be the filter for $(x^{\epsilon}(\cdot),\ z^{\epsilon}(\cdot))$ with data $y^{\epsilon}(\cdot)$, and (5.6) is the filter for (5.5), (5.7). The cost functions for the control problem are

$$R^{\epsilon}(u) = \int_0^T E\ r(x^{\epsilon}(t),\ z^{\epsilon}(t),\ u(t))dt, \tag{5.8}$$

$$R(u) = \int_0^T E\ r(x(t),\ z(t),\ u(t))dt, \tag{5.9}$$

for bounded and continuous $r(\cdot,\cdot,\cdot)$, and some $T < \infty$.

The controls take values in a compact set U, and we let (see related definition of $\mathfrak{D}$ and $\mathfrak{D}_t$ in Section 3) $\mathfrak{U}$ denote the set of U-valued measurable (ω,t) functions on $C^g[0,\infty) \times [0,\infty)$ which are continuous w.p.1. relative to Wiener measure. Let $\mathfrak{U}_t$ denote the subclass which depends only on the function values up to time t. We view functions in $\mathfrak{U}$ as the data dependent controls with value $u(y(\cdot),t)$ at time t and data $y(\cdot)$. Let $\bar{\mathfrak{U}}$ denote the subclass of functions $u(\cdot,\cdot) \in \mathfrak{U}$ such that $u(\cdot,t) \in \mathfrak{U}_t$ for all t and with the use of control $u(y^{\epsilon}(\cdot),\cdot)$ (resp., $u(y(\cdot),\cdot)$), (5.2) and (5.4) (resp., (5.6), (5.7)) has a unique solution in the sense of distibutions. These $u(y^{\epsilon}(\cdot),\cdot)$ and $u(y(\cdot),\cdot)$ are the admissible controls.

Commonly, one uses the model (5.5) to (5.7) to get a (nearly) optimal control for cost (5.9). This control would, in practice, actually be applied to the 'physical' system (5.2), (5.4), with actual cost function (5.9). Although such controls would normally not be 'nearly' optimal in any strict sense for the physical system, they are 'nearly optimal with respect to a useful class of *comparison controls.*

Straightforward weak convergence arguments (using only the assumed weak convergence of the 'driving $W^{\epsilon}(\cdot)$, $W_x^{\epsilon}(\cdot)$ processes', and the uniqueness of the limit) can be used to prove Theorem 5.1. Let $\mathcal{M}$ denote the class of U-valued continuous functions $u(\cdot,\cdot,\cdot)$ such that with use of control with value $u(\hat{x}(t),\ \hat{z}(t),t)$ at time t, (5.6), (5.7), has a unique (weak sense) solution. Let $\mathcal{M}_0$ denote the subclass of controls (stationary controls) which do not depend on t (for use in the next section). Let $u(y^{\epsilon},\cdot)$, $\bar{u}^{\delta}(\hat{x}^{\epsilon},\hat{z}^{\epsilon},\cdot)$ and $\bar{u}^{\delta}(\hat{x},\hat{z},\cdot)$ denote the controls with values $u(y^{\epsilon}(\cdot),t)$ $\bar{u}^{\delta}(\hat{x}^{\epsilon}(t),\hat{z}^{\epsilon}(t),t)$ and $\bar{u}^{\delta}(\hat{x}(t),\hat{z}(t),t)$ at time t.

Theorem 5.1. *Assume the conditions above in this section. For* $\delta > 0$, *let there exist a control* $\bar{u}^{\delta}(\cdot,\cdot)$ *in* M *which is* δ*-optimal for* (5.6), (5.7), (5.9), *with respect to controls in* $\bar{\mathcal{H}}$. *Then, for any* $u(\cdot,\cdot) \in \bar{\mathcal{H}}$,

$$\varliminf_{\epsilon} R^{\epsilon}(u(y^{\epsilon},\cdot)) \geq \lim_{\epsilon} R^{\epsilon}(\bar{u}^{\delta}(\hat{x}^{\epsilon}, \hat{z}^{\epsilon}, \cdot)) - \delta \tag{5.10}$$
$$= R(\bar{u}^{\delta}(\hat{x}, \hat{z}, \cdot)) - \delta .$$

6. Filtering and Control: The Large Time Case.

For the combined filtering and control analog of the large time and bandwidth problem of Section 4, we use the assumptions:

C6.1. $\begin{bmatrix} A_x & 0 \\ 0 & A_z \end{bmatrix} \equiv A$ *is stable,* $\left[A;H_x,H_y\right]$ *is observable and* $\begin{bmatrix} A, & \begin{matrix} B_x \\ B_z \end{matrix} \end{bmatrix}$

controllable.

C6.2. $\xi^{\epsilon}(\cdot)$ *satisfies* (C4.2).

The cost functions are

$$\gamma^{\epsilon}(u) = \varlimsup_{T} \frac{1}{T} \int_0^T E\ r(z^{\epsilon}(t),x^{\epsilon}(t),u(t))dt \tag{6.1}$$

$$\gamma(u) = \varlimsup_{T} \frac{1}{T} \int_0^T E\ r(z(t),x^{\epsilon}(t),u(t))dt \tag{6.2}$$

We adapt the point of view of [10, Section 6] and assume that the system can be Markovianized. This is incorporated in the following assumption. This greatly facilitates dealing with the weak convergence on the infinite interval. The Skorohod topology gives 'decreasing' importance to the values of the processes as t increases - but it is the values at 'large' t that determine the cost γ^{ϵ} or γ. The problem is avoided by working with the invariant measures for the $\{\xi^{\epsilon}(\cdot), x^{\epsilon}(\cdot),...\}$ processes.

C6.3. *For each* $\epsilon > 0$, *there is a random process* $\psi^{\epsilon}(\cdot)$ *such that* $\{\psi^{\epsilon}(t), t < \infty\}$ *is tight and for each* $u(\cdot) \in M_0$, (M_0 *defined above Theorem* 5.1) $X^{\epsilon}(\cdot) \equiv \{x^{\epsilon}(\cdot),z^{\epsilon}(\cdot),\hat{x}^{\epsilon}(\cdot),\hat{z}^{\epsilon}(\cdot),\psi^{\epsilon}(\cdot),\xi^{\epsilon}(\cdot),\xi^{\epsilon}_{x}(\cdot)\}$ *is a right continuous homogeneous Markov-Feller process (with left hand limits).*

Remark. If $z^{\epsilon}(\cdot)$ satisfies $\dot{z}^{\epsilon} = A_z z^{\epsilon} + \xi^{\epsilon}_{z}$, then the assumption (C6.3) holds if the driving noises $(\xi^{\epsilon}_{x}(\cdot),\xi^{\epsilon}_{y}(\cdot),\xi^{\epsilon}_{z}(\cdot))$ satisfy (C6.3) and (C6.1), (C6.2) hold; i.e., if the noises $\xi^{\epsilon}_{z}(\cdot)$ and $\xi^{\epsilon}(\cdot)$ can be written as functions of a suitable Markov process. Let

$u(\hat{x}^\epsilon,\hat{z}^\epsilon)$ and $u(\hat{x},\hat{z})$ (and similarly for u^δ) denote controls with values $u(\hat{x}^\epsilon(t),\hat{z}^\epsilon(t))$ and $u(\hat{x}(t),\hat{z}(t))$ at time t.

Theorem 6.1. *Assume the conditions of Theorem* 5.1 *and* (C6.1) - (C6.3). *Let* $\xi^\epsilon(\cdot)$ *and* $\xi^\epsilon_x(\cdot)$ satisfy (C4.2) *and let* $z^\epsilon(\cdot)$ *satisfy* (C4.3). *For* $\delta > 0$, *let there be a* δ*-optimal control* $\bar{u}^\delta(\cdot,\cdot) \in M_0$ *for the system* (5.1), (5.6), (5.7), *and cost* (6.2), *and for which* (5.1), (5.6), (5.7) *has a unique invariant measure. Then, for* $u(\cdot,\cdot) \in M_0$

$$\underline{\lim}_{\epsilon}\ \gamma^\epsilon(u(\hat{x}^\epsilon,\hat{z}^\epsilon)) \geq \lim_{\epsilon} \gamma^\epsilon(\bar{u}^\delta(\hat{x}^\epsilon,\hat{z}^\epsilon)) - \delta \tag{6.3}$$

$$= \gamma(\bar{u}^\delta(\hat{x},\hat{z})) - \delta .$$

Remark. Various extensions of the class of admissible controls for which the same proof works are discussed in [9].

Proof. Fix $u(\cdot,\cdot) \in M_0$. Define the 'averaged transition measure'

$$P^\epsilon_T(\cdot) = \frac{1}{T} E\int_0 P\{X^\epsilon(t) \in \cdot \,|X^\epsilon(0)\}dt,$$

where the expectation E is over the possibly random initial conditions, and $X^\epsilon(\cdot)$ is the process corresponding to the use of $u(\hat{x}^\epsilon(\cdot),\ \hat{z}^\epsilon(\cdot))$. By the hypothesis, $\{P^\epsilon_T(\cdot), T \geq 0\}$ is tight. Also (writing $X = (x,z,\hat{x},\hat{z})$)

$$\gamma^\epsilon(u(\hat{x}^\epsilon,\hat{z}^\epsilon)) = \overline{\lim_{T}} \int r(x,z,u(\hat{x},\hat{z}))\ P^\epsilon_T(dX) . \tag{6.4}$$

Let $T^\epsilon_n \to \infty$ be a sequence such that it attains the limit $\overline{\lim_{T}}$, and for which $P^\epsilon_{T^\epsilon_n}(\cdot)$ converges weakly to a measure, which we denote by $P^\epsilon(\cdot)$. The $P^\epsilon(\cdot)$ is an invariant measure for $X^\epsilon(\cdot)$. Also, by construction of $P^\epsilon(\cdot)$,

$$\gamma^\epsilon(u(\hat{x}^\epsilon,\hat{z}^\epsilon)) = \int\gamma(x,z,u(\hat{x},\hat{z}))P^\epsilon(dX) .$$

Let $(x^\epsilon_0(\cdot),z^\epsilon_0(\cdot),\hat{x}^\epsilon_0(\cdot),\hat{z}^\epsilon_0(\cdot))$ denote the first four components of the *stationary Markov-Feller* $X^\epsilon(\cdot)$*-process* associated with the invariant measure $P^\epsilon(\cdot)$. By our hypotheses (see the argument in Section 4) $\{x^\epsilon_0(\cdot),z^\epsilon_0(\cdot),\hat{x}^\epsilon_0(\cdot),\ \hat{z}^\epsilon_0(\cdot)\}$ converges weakly to a limit $(x_0(\cdot),z_0(\cdot),\hat{x}_0(\cdot),\hat{z}_0(\cdot))$ satisfying (5.7), (5.1), (5.6). Also, the limit must be stationary, since the $(x^\epsilon_0(\cdot),\ldots,\hat{z}^\epsilon_0(\cdot))$ is for each ϵ. Let $\mu^u(\cdot)$ denote the invariant measure associated with *this stationary limit.* Then

$$\gamma^\epsilon(u(\hat{x}^\epsilon,\hat{z}^\epsilon)) \to \gamma(u(\hat{x},\hat{z})) = \int r(x,z,u(\hat{x},\hat{z}))\ \mu^u(dxdzd\hat{x}d\hat{z}) .$$

By a similar argument, it can be shown that

$$\gamma(\bar{u}^{\delta}(\hat{x},\hat{z})) \equiv \int r(x,z,\bar{u}^{\delta}(\hat{x},\hat{z}))\ \mu^{\bar{u}^{\delta}}\ (dxdzd\hat{x}d\hat{z})$$
$$= \lim_{\epsilon}\ \gamma^{\epsilon}(\bar{u}^{\delta}(\hat{x}^{\epsilon},\hat{z}^{\epsilon}))\ .$$

(The uniqueness of the invariant measure $\mu^{\bar{u}^{\delta}}(\cdot)$ is used here). Inequality (6.3) now follows from the δ-optimality of $\bar{u}^{\delta}(\cdot)$. Q.E.D.

References

[1] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, M.I.T. Press, Cambridge, U.S.A., 1984.

[2] H.J. Kushner and Hai Huang, "Approximate and Limit Results for Nonlinear Filters with Wide Bandwidth Observation Noise", Stochastics, Feb., 1986.

[3] P. Billingsley, *Convergence of Probability Measures*, 1968, Wiley, New York.

[4] T.G. Kurtz, *Approximation of Population Processes*, 1981, Vol. 36 in CBMS-NSF Regional Conf. Series in Appl. Math, Soc. for Ind. and Appl. Math, Phila.

[5] A. Benveniste, "Design of Monostep and Multistep Adaptive Algorithms for the Tracking of Time Varying Systems," Proc., 23 Conf. on Dec. and Control, 1984, IEEE Publications, New York.

[6] M. El-Ansary and H. Khalil, "On the Interplay of Singular Perturbations and Wide-band Stochastic Fluctuations", SIAM J. on Control, *24*, 1986, 83-98.

[7] H.J. Kushner and Hai Huang, "Averaging Methods for the Asymptotic Analysis of Learning and Adaptive Systems with Small Adjustment Rate", SIAM J. on Control and Optim., *19*, (1981), 635-650.

[8] H. Kushner, "Jump Diffusion Approximations for Ordinary Differential Equations with Wideband Random Right Hand Sides", SIAM J. on Control and Optimization, *17*, 1979, 729-744.

[9] H. J. Kushner and W. Runggaldier, "Filtering and Control for Wide Bandwidth Noise and 'Nearly' Linear Systems", LCDS Rept. #86-8, 1986, Brown Univ.; to appear in IEEE Trans. on Aut. Control.

[10] H. Kushner and W. Runggaldier, "Nearly Optimal State Feedback Controls for Stochastic Systems with Wideband Noise Disturbances", to appear SIAM J. on Control and Optimization. Also, LCDS Rept. #85-23, 1985, Brown Univ.

[11] A.V. Skorohod, "Limit Theorems for Stochastic Processes", Theory of Probability and Its Applications, *1*, 1956, 262-290.

[12] H.J. Kushner, "Diffusion Approximations to Output Processes of Nonlinear Systems with Wide-band Inputs, and Applications", IEEE Trans. on Inf. Theory, *IT-26*, 1980, 715-725.

[13] G.B. Blankenship and G.C. Papanicolaou, "Stability and Control of Stochastic Systems with Wide Band Noise Disturbances", SIAM J. Appl. Math *34*, 1978, 437-476.

CONVERGENCE,CYCLING OR STRANGE MOTION IN THE ADAPTIVE SYNTHESIS OF NEURONS

E.Labos
Semmelweis Medical School,1st Dept.of Anatomy,Budapest,Hungary

1.INTRODUCTION

Among the various learning procedures (see in Nilsson, 1965; Mendel and Fu; Kohonen, 1978; Minsky and Papert, 1969; Fukushima, 1981) perceptrons (Rosenblatt, 1958, 1962; Widrow, 1963) represent a class of machine, where machine efficiency is based on so called perceptron convergence theorems (PCDTs; Novikoff, 1962; Minsky and Papert, 1969). These kinds of theorems predict that the initially naive machine (laymachine) will metamorphose via a finite number of steps into a trained machine. The training itself can be made completely automatic.

The only important prerequisite for perceptron learning is that of linear separability, since the theory is applied mainly to formal neurons, threshold gates or linearly separable truth functions (McCulloch and Pitts, 1943; Winder, 1968, 1969; Lewis and Coates, 1967; Muroga, 1971).

In the actual formulations of PCOTs the initial state, the sequence of inputs, as well as the actual form of linear separation are regarded as indifferent with respect of successful learning because of the special formulations of the theorems.

This paper attempts to extend the PCDT to an optional formal neuron, i.e. an arbitrary kind of linear separability. At the same time conditions of cycling in the separable case also will be given. Such an example was published by Labos (1984). Until now a perceptron cycling theorem (PCYT; see in Minsky and Papert, 1965) was formulated for the nonseparable case only. The nontrivial aspect is here the lack of divergence when a non-teachable object is tried to be trained. At last conditions of bounded, aperiodic, non-convergent behaviour will be formulated.

2.PROPERTIES OF FORMAL NEURONS RELEVANT TO ADAPTIVE SYNTHESIS.

DEFINITION:A truth-function is a formal neuron (synonymous with linearly separable truth or switching function, threshold gate, McCulloch-Pitts neuron; McCulloch and Pitts, 1943; Dertouzos, 1965; Muroga, 1971; Lewis and Coates, 1967; Hu-Sze-Tsu, 1965; Sheng, 1969; Labos, 1984; Loe and Goto, 1986) if (1) it is defined on B^n set of binary vectors of n components; (2) the values are of the B^1 set; (3) if there exists a $h \in R^n$ Euclidean vector and T real number - called threshold - so that for each $p \in B^n$ the $f(p) \in B^1$ values are computable as follows:

$$f(p)=u(ph-T) \qquad (1)$$

where $ph=\sum_{i=1}^{n} p_i h_i$ and $u(r)=1$ if $r>0$ and $u(r)=0$ otherwise.

CODING CONVENTIONS:Vectors of B^n and functions of $B^n \longrightarrow B^1$ are identified by decimal codes: e.g. $(1,0,1)\sim 101 \sim 5$. Concerning the functions at first the arguments are listed by increasing code numbers. The corresponding list of f(0), f(1),... is a 0/1 tape. Its decimal equivalent is the code of the function. E.g. $f_{186}\sim 10111010$ with 3 input lines means that f((0,0,0))=f(0)=1 or f(5)=0, f(7)=0 etc..

The $h\in R^n$ vector is called a separating vector, its coordinates are called synaptic weights. The U set of p vectors for which f(p)=1 holds is called support or true vector set, while the complementary part W is the set of false vectors.Obviously:$U\cap W=\emptyset$ and $U\cup W=B^n$.

The realization of an f truth-function is rarely possible since a system of linear inequalities of 2^n relations have to be solved of the following forms:

$$ph-T>0 \qquad \text{or} \qquad ph-T\leq 0 \qquad (2)$$

The number of (h,T) realizations is infinite.It is always possible with integer h_i-s and with non-zero i.e. $T>0$ or $T<0$ thresholds.

The value of ph (where $p\in B^n$ and $h\in R^n$) is called the effect of p input or answer to the p question. The T threshold separates these effects into the Wh and Uh sets. For any particular realization there exists a maximal effect-number of false input vectors, M and also a minimum value of possible effects for true vectors:

$$M = \max Wh = \max_{p\in W}\{ph\} \text{ and } \quad m = \min Uh = \min_{p\in U}\{ph\} \qquad (3)$$

These numbers will be called lower(M) and upper(m) gateposts or margines. The positive value, g=m-M is called the gap of the neuron at realization $\langle h,T\rangle$. The threshold T can be chosen freely in $[M,m)$ if h is fixed. The interval (M,m) is either completely negative or completely positive. Otherwise the answer to $0\in B^n$ input would be ambigous.

The Chow-parameters (Chow, 1961; Elgot, 1961; Winder, 1969) are computable as follows. Take all true vectors and sum these as real and not as logical vectors to get a new vector. Its components are the Chow-parameters, which together with the number of true vectors exactly identify the neuron. For example for f_{186} $U=\{0,2,3,4,6\}$. Thus the parameter array is $f_{186}\sim(5;2,3,1)$. If the Chow-vector is doubled and the number of true vectors is subtracted from each coordinate, a vector is obtained which very often can be used as an h separating vector. In the example

2.(2,3,1)-(5,5,5)=(-1,+1,-3). The ordering and sign of these $2s-s_i$ numbers represents the order and sign of synaptic weigths (h_i-s).

3.THE LINEAR SEPARATION AS A LEARNING AUTOMATON

The input space of the learning machine applied to an optional formal neuron is B^n . The state space $S\subset R^n$ is determined only after the initial state of the student neuron was fixed. It is an unnecessary complication to say that the state space is a set of $B^n\longrightarrow B^1$ neurons and that the training modifies this functions. However, it is emphasized that the states represent functions.

The law of state transitions is as follows:

$$s_{k+1}=F(s_k,p_k)=\begin{cases} s_k+rp_k & \text{if } t(p_k)\neq z_k(p_k) \\ s_k & \text{if } t(p_k)=z_k(p_k) \end{cases} \qquad (4)$$

At the first condition errors emerge while in the 2nd case the transitions are mute.The correction factor is $r>0$,s_k is the actual state vector and p_k is the actual input vector or question.

If $t\in B^n\rightarrow B^1$ is the fixed teacher-function, which represents a reference in the teaching machine or adaptive synthesis procedure, and $z_k\in B^n\rightarrow B^1$ is the threshold gate determined by the s_k R^n vector at a previously fixed T thtreshold, then:

$$s_{k+1}=s_k+r(t(p_k)-z_k(p_k))p_k=s_k+rdp_k \quad \text{where } d=-1,1 \text{ or } 0 \qquad (5)$$

Thus errors and the corresponding corrections occur if the pattern of the two responses are either (0,1) or (1,0). However, the 'student' remains unchanged if the patterns are either (0,0) or (1,1).

4.THE STOP CONDITION

Generate a complete sequence of inputs without repetitions in some lexicographic order. If 2^n consecutive no error conditions emerge then this indicates the arrival at the desired learned state which is a not necessarily unique fixed point of the learning machine.

5.IS THE STATE-SPACE FINITE OR NOT?

The $\{s_k\}$ R^n space is not necessarily finite. However, the M_n function space which they represent ($M_n \subset B^n \rightarrow B^1$) is finite.E.g.at n=3 104 threshold gates occur among the 256 switching functions. Each is represented by a nonempty and most often unbounded convex set of states.

6.HOW TO CHANGE THE THRESHOLD?

This problem is related to the case of zero input. In numerous perceptron algorithms no specification is given for this case, since the threshold is arbitrarily made equal to zero.

If $p_k \in B^n$,$p_k=0$ and any of the previous error conditions appears, then the correction used for non-zero inputs is ineffective:

$$s_{k+1}=s_k+rp_k=s_k+r.0=s_k \qquad (6)$$

Therefore if $p_k=0$ and error occurs, then a threshold modification is necessary. Otherwise this $p_k=0$ input will block the convergence.

At the same time such an additional part to the machine is sufficient to avoid this kind of error, because the sign of the threshold is dependent solely on the response to zero-vector. If t(0)=1 then $T<0$ must hold and if t(0)=0 then $T \geq 0$ should be satisfied. Furthermore, if a threshold gate can be designed with a zero threshold, then it can be realized by a positive T as well. The inverse is not true. For example, the 'and-gates' need positive definite thresholds.

Consequently, the algorithm may be supplemented by an examination of the input: whether it is zero or not. Thus starting with an arbitrary non-zero T threshold, it is sufficient to examine the incorrectness of $z_k(0)$ compared to t(0) and if it is erronous, simply change sign. It is the easiest to do this at the beginning. This then would eliminate the threshold-problem.

If this part is left out and zero-vectors occur in the p_k question-sequence, then cycling may occur.

7.THE ERROR CONDITIONS.

Suppose that the teacher and student neurons have the same n number of input lines. Teacher here means that the table of a truth-function values is given as a reference for comparison or a (h,T) separating vector and threshold pair specifies the reference if it is separable. The first case is more probable since this adaptive algorithm is just a test of separability. The student state is s_0 initially(e.g. $s_0=0$ $\in R^n$ is suitable).

A training sequence is generated now iteratively as follows:

$$\begin{array}{ll} s_{k+1}=s_k & \text{if } t(u_k)=1 \text{ and } u_k s_k > T \\ s_{k+1}=s_k & \text{if } t(w_k)=0 \text{ and } w_k s_k \leq T \\ s_{k+1}=s_k+ru_k & \text{if } t(u_k)=1 \text{ and } u_k s_k \leq T \\ s_{k+1}=s_k-rw_k & \text{if } t(w_k)=0 \text{ and } w_k s_k > T \end{array} \qquad (7)$$

Here u_k and w_k are true and false input (i.e. p_k question) vectors; $u_k s_k$ and $w_k s_k$ are inner products of the questions and the actual state s_k.

If a threshold correction is not built into the algorithm, then choose T freely. If it was built, then investigate the t(0) reference value and choose T=+1 if t(0)=0 or T=-1 if t(0)=1.

8.THE REDUCED TRAINING SEQUENCE.

The (p_i,s_i) state-input pairs form the complete training sequences. However, if the pairs are omitted for which no real transition has occured (i.e when $s_{i-1}=s_i$) then a shorter sequence is created including solely the real corrections. In the following the subscripts of (p_i,s_i) will refer only to this sequence of non-mute iterations. If s_k denotes the state after the k-th correction, then

$$s_k = \left[\sum_{j=1}^{k} ru_j - \sum_{j=1}^{k-a} rw_j\right] + s_0 \qquad (8)$$

9.THE SEPARABILTY CONDITION.

This requirement means, that the teacher or reference-function, i.e.the $t \in B^n \rightarrow B^1$ truth-function is linearly separable. In this case t is representable as $t \sim (h,T)$, where h R^n and T is the threshold number: $T \in [M,m)$.

10.THE WAY OF CONVERGENCE PROOFS.

10.1.LOWER BOUND FOR THE LENGTH OF STATE VECTOR.AN OBSTACLE.

Since the separabilty condition includes lower estimations of $u_j h$ and $-w_j h$ scalar products, the following relations hold:

$$s_k h = r\left[\sum_{j=1}^{a} u_j h - \sum_{j=1}^{k-a} w_j h\right] > raT+r(a-k)T = rT(2a-k) \qquad (9)$$

An obstacle to the succesful continuation of the proof appears here. If the $s_k h$

absolute value can be estimated with invariant direction of inequality then the following step may come:

$$|s_k|^2 |h|^2 \geq |s_k h|^2 \geq r^2 T^2 (2a-k)^2 \tag{10}$$

where the Cauchy-Schwartz relation was applied. This is a usual, inherent part of the proofs of PCOTs (Minsky, Papert, 1969; Lewis and Coates, 1968).

The sign of $s_k h$ and also the sign of lower bounds might be negative and a priori very little can be said concerning the ralation of a and k-a. The threshold cannot always be taken zero. Three cases have to be distinguished: (1) $M \leq T < m < 0$; (2) $0 = M \leq T < m$; and (3) $0 < M \leq T < m$. In the proofs of similar theorems the case (2) occurs. However the turning to absolute values is blocked by the ambiguity of signs.

10.2.UPPER BOUND FOR THE LENGTH OF STATE VECTOR.

This is a problem-free part of convergence-proofs since the separability condition is not utilized. The estimation exploits the change of state vector in a single correcting step and by summing the relations obtained an upper bound for the square of the length of the student vector can be derived. At first:

$$|s_{k+1}|^2 = |s_k|^2 \pm 2rp_k s_k + r^2 |p_k|^2 \tag{11}$$

where again the sign of $p_j s_j$ depends on the type of error. The error conditions give upper bounds for $u_j s_j$ and for $-w_j s_j$. For the square of length of input $|p_j|^2 \leq n$ holds. Summing according to the steps of corrections $|s_k|^2 = \sum_{j=1}^{k} (|s_{j+1}|^2 - |s_j|^2)$ holds, and hence:

$$|s_k|^2 \leq 2raT + 2r(k-a)T + knr^2 \tag{12}$$

10.3.CONDITION OF THE FINITNESS OF REDUCED TRAINING SEQUENCE.

If both of the upper and lower bounds are applicable, then the finitness of the reduced training sequence follows.

11.EXAMPLES OF CYCLING.

CASE 1 - A 3-fan in neuron: $f_{11} \sim 00001011$. It is realizable by (2,1,-1) separating vector; $U = \{4,6,7\}$, $W = \{0,1,2,3,5\}$; $M=1$, $m=2$; $1 \leq T < 2$; Chow-parameters: (3;3,2,1). Choose: $s_0 = (0,0,0)$, initial state, $T=1$ threshold and $r=1$ correction factor. Let the series of inputs be the following sequence: (0,4,1,5,2,6,3,7) where these integers represent input vectors (000,001,010,011,etc..).

Fig. 1 shows the state transition matrix and also definitions of the created states. A sample pattern of the state transitions and questions is as follows:

00 40 1F 5F 2F 6F 3R 7R 0R 4R 1R 5R 2S 6S 3S 7S 0T 4T 1T 5T 2U 6S 3S 7S etc...

Here the sequence is formed by the consecutive $\langle p_k,s_k \rangle$ pairs and the states are as follows: O=(0,0,0),F=(1,0,0),R=(2,1,0),S=(1,1,-1),T=(2,2,0),U=(1,2,-1)

At the given conditions (that is at correct threshold, s_0=0,r=1 and natural question pattern) no convergence takes place and the (ABC) cycle emerges. Investigating the matrix of Fig 1 numerous other possibilities may be detected to create similar state-cycles.

Comparable matrix structures have been observed also when the following formal neurons were synthetized(all of them with 3 input lines): f_{11}, f_{13}, f_{35}, f_{49}, f_{69}, f_{81}, f_{174}, f_{186}, f_{206},f_{220}, f_{242}, f_{244}. These functions form an equivalence class generated by negation and permutation of variables.

Two other phenomena are demonstrable by the same case: depending on question strategy or initial state,cycling or convergence may occur.

CASE 2 - The same function generates different state-transition matrices depending on the threshold, assuming that other conditions are identical and both thresholds are correct (i.e. realizable). Compare the training of f_{186}, n=3 with T=-2 and T=-3. In the first case cycling may appear, in the last one the convergence is absolute.

CASE 3 - If the function is linearly separable, but the threshold is not adequate, then cycling may arise. Try with n=2 'and-gate'. This is realizable with h=(1,1) and T=1. However, with a trial of T=0 or $T<0$, no success is available and cycling arises. Similar 'threshold-inadequate' cases are not regarded as 'surprising' examples.

CASE 4 The significance of the r correction factor is demonstrable by the next synthesis cases: n=4,s_0=(0,-2,-4,0); the reference neuron is specified by ((-1,-1,-1,-1),4),M=-4 and m=-3. The input sequence starts with 15=(1,1,1,1)$\in B^4$ and continues with 0,1,2,3,4...14,15. This whole B^4 set is repeated cyclically. If r=4, then the following reduced training sequence was observed:

(0,-2,4,0)15(-4,-6,0,-4)1(-4,-6,0,0)4(-4,-2,0,0)8(0,-2,0,0)15(-4,-6,-4,-4)1(- 4,-6,-4 ,0)2(-4,-6,0,0)4(-4,-2,0,0)15...

The length of cycle is L=5: the input 4 evokes the same correction as before. The cycle condition (see Section 13) is satisfied since the two vectorial sums are equal: (0100)+(1000)+(0001)+(0010)=(1111), where L_1=4 and L_2=1.

A similar cycling occurs if s_0 =(0,0,0,0) and r=4. But decreasing the value of r the length of cycle also decreased. At $r>4/3$ L=3 was also observed. However if $1 < r \leq 4/3$ was satisfied then a quick convergence was achieved with the same initial states.

The examples prove that cycling may arise in the following inadequate choices even if the other conditions were formally satisfied:

(1) training of non-separable truth-function(case of Minsky and Papert, 1969)

(2) an attempt to separate with incorrect threshold

(3) separation of a separable function with adequate threshold but with bad initial state or

(4) with bad question pattern or

(5) by a too large value of correction factor

12.AN ESCAPE FROM THE CYCLING PITFALL BY RANDOM INPUT STRATEGY.

As it was shown, when the f_{11} neuron was taught, a cycling may arise either at lexicographic repetitive or regular sequences of inputs as a result of iterations. However, generating the inputs in specific random way, then an 'escape of cycling traps' toward the learned state (fixed point) was possible.

A definition of probabilistic learning automaton may be as follows:

1. An initial distribution of states must be chosen. E.g.: P(0)=1 for the 0 initial state and the other probabilities are zeroes.

2. The transition probability matrix is designed on the basis of the non-random case. Probabilities for any non-prohibited transitions are equal to $1/2^n$. If from a state, k different inputs leads to non-dummy transitions, then $(2^n-k)/2^n$ is the chance of a state to remain unchanged. This is a 'natural' but not the unique possibility of a stochastic matrix definition.

The next distribution now is computable: $D_{n+1}=D_nW$, where $D_i \in R^m$ and W is an mxm stochastic matrix, m is the number of created states.

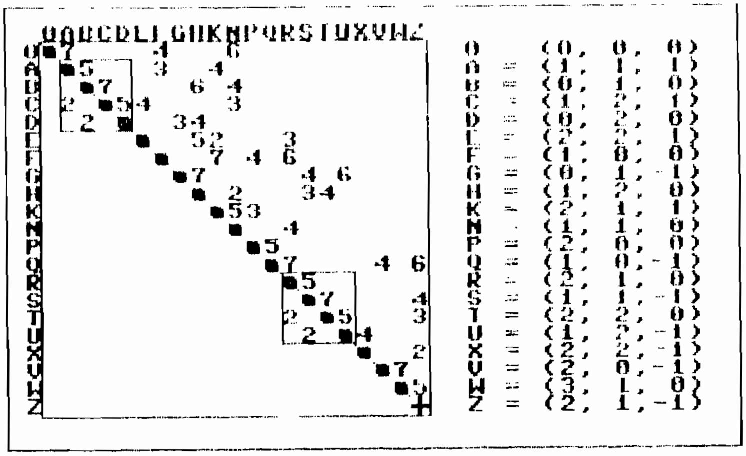
In Fig 2 the probabilities of six sets of states are plotted against the parameter of iteration (i.e. the 'time'). The probability of the state set which includes the unique fixed point will be near to one after a few iteration.

It is obvious that the convergence is dependent on the transition graph structure since it was designed on this basis. If cycling basins have escape pathways along which the probability of passage toward the fixed point set is non-zero, then this 'random convergence' to learned state is guaranteed.

A metaphora of such a process is e.g the fate of one liter of fluid in some vessels (=state sets) which are interconnected by tubes of which the conductances are proportional to the transition probabilities. The P(t) results of computation includes polynomial and exponential terms as a function of the t time-parameter since this process corresponds to a well amenable discrete homogenous linear process.

13.A CONDITION OF CYCLING.

Let s_0 be an initial state of the learning machine. If after L corrections this state reappears first, then L is the length of state cycle and the sum of corrections in it must be zero:


0 A B C D E F G H K M P Q R S T U X V W Z
0 A B C D E F G H K M P Q R S T U X V W Z

0		⟨0, 0, 0⟩
A	=	⟨1, 1, 1⟩
B	=	⟨0, 1, 0⟩
C	=	⟨1, 2, 1⟩
D	=	⟨0, 2, 0⟩
E	=	⟨2, 2, 1⟩
F	=	⟨1, 0, 0⟩
G	=	⟨0, 1, -1⟩
H	=	⟨1, 2, 0⟩
K	=	⟨2, 1, 1⟩
M	=	⟨1, 1, 0⟩
P	=	⟨2, 0, 0⟩
Q	=	⟨1, 0, -1⟩
R	=	⟨2, 1, 0⟩
S	=	⟨1, 1, -1⟩
T	=	⟨2, 2, 0⟩
U	=	⟨1, 2, -1⟩
X	=	⟨2, 2, -1⟩
V	=	⟨2, 0, -1⟩
W	=	⟨3, 1, 0⟩
Z	=	⟨2, 1, -1⟩

Figure 1 State transition matrix of a learning machine. The f_{11} (n=3) performance was taught. Rows represent s_k actual and columns next s_{k+1} states. Numbers denote input-codes under the effect of which transitions are realized. Empty places correspond to prohibited transitions. The diagonal spots represent mute transitions. Altogether 21 states and 46 non-dummy transitions are observable. The encircled two submatrices enclose cycles. The state symbols are defined as R^3 vectors.

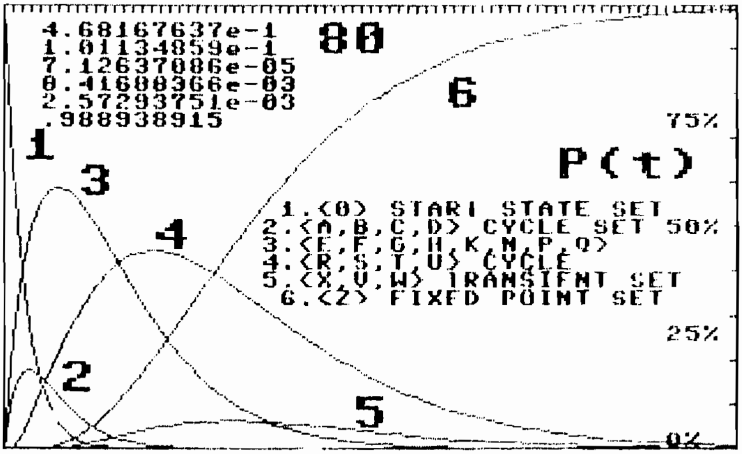


Figure 2 Abscissa: t number of iterations. Ordinate: P(t) probabilities of the defined six sets of states. The machine of Fig 1 was fed by random input sequence. At start P(0)=1. Each non-prohibited transition was equally probable. Chance to remain unchanged corresponds to the sum of mute transition probabilities. Observe that to stay in the two basins of this machine show transient peaks, while the probability of learning tends to 1.

$$s_0 + r \sum_{i=0}^{L} p_i = s_0 \quad \text{or} \quad \sum_{i=0}^{L} p_i = 0 \quad R^n \tag{13}$$

As a part of corrections consists of additions and others represent substractions of the corresponding inputs, it follows that:

$$\sum_{i=1}^{a} u_i = \sum_{i=1}^{L-a} w_i \quad \text{where} \quad u_i \in U, w_i \in W \tag{14}$$

Since the added inputs are true, the substracted ones are false vectors of the reference function, then a sum of true vectors must be equal to a sum of false vectors. However, there exists the theorem of Elgot (1961) and Chow (1961) called assummability theorem which states that linear sparability is equivalent with assummability (proof see in Muroga, 1971 p.175). This means that if $u_i \in U$ and $w_i \in W$, then the following relation cannot be satisfied:

$$\sum_{i=1}^{L} u_i = \sum_{i=1}^{L} w_i \tag{15}$$

Nevertheless, if the number of vectors in these two sums is not identical, then the relation may be true. For example (1,1,1) is false and (1,0,0) or (0,1,1) are true vectors of $f_{186} \sim$ 10111010 and the first one is the sum of the two latter ones.

THEOREM: For the inputs which evoke cycle of states in the learning automaton, $L_1 \neq L_2$ must hold in the following necessary relation:

$$\sum_{i=1}^{L_1} u_i = \sum_{i=1}^{L_2} w_i \tag{16}$$

Remark: The condition means a kind of summability of true (u) or false (w) vectors of threshold gates.

14. THE POSSIBILITY OF BOUNDED APERIODIC MOTIONS.

Aperiodic sequence of states is designable by suitable choice of input sequence. Take first a function of which the state transition graph includes cycling basin with alternative connected cycles. E.g. in Fig. 1 $C_1=(BCA)^*$ and $C_2=(BCD)^*$ are such ones. Aperiodic but regular input sequences may control walking from C_1 to C_2 and back while turning numbers are e.g. increasing. Such non-autonomous aperiodic motions in a learning automaton are possible.

15. DIFFICULTIES IN ADAPTIVE SYNTHESIS OF NETWORKS.

A negligible part only of the truth-functions is synthetizable as a single threshold gate but may be realized with suitable - at the worst case very large - networks. The adaptive (or other) synthesis algorithms now available do not solve the problem of minimalization of the network, since no a priori knowledge is available about the size of the corresponding networks.

16.CONCLUSIONS.

The essential message of this paper is as follows. The convergence proofs of adaptive single threshold gate synthesis in which it is claimed that only separability is necessary for the convergence (e.g. Minsky and Papert, 1969) are correct. Nevertheless, the separability concept applied in them is surely too narrow, because based exclusively on the cases realizable with zero threshold. For the remaining cases it is said that are non-separable and therefore cycling (Minsky and Papert, 1969).

However, more than half of the threshold gates are synthetizable only with definite negative or positive thresholds.As the examples of this lecture show, the absolute convergence cannot be extended to these cases. It is necassary to specify: (1) initial state, (2) threshold correction, (3) adequate values of correction factor ,(4) a suitable input or question-sequence generation, and (5) an adequate choice of (h,T) representation.

The most important open problems are related to the structure of transition graph or matrix: (1) Is it at least one fixed point or not in the transition graph if the function is separable; (2) Is the fixed point set reachable from any initial state by choosing a suitable question strategy and correction factor or not?

REFERENCES

1. Block,H.D.: The Perceptron: A Model for Brain Functioning.Rev.Mod.Phys. 34 (1962) 123-135.
2. Chow,C.K.: On the Characterization of Threshold Functions.Proc.of the Symp.on Switching Circuit Theory and Logical Oesign,AIEE. (1961) 34-38.
3. Oertouzos,M.L.: Threshold Logic: A Synthesis Approach. MIT Press, Cambridge, Mass., 1965.
4. Elgot,C.C.: Truth Functions Realizable by Single Threshold Organs.Proc.of the Symp. on Switching Circuit Theory and Logical Design, AIEE, (1961) 225-245.
5. Fukushima,K.: Cognitron: A Self-Organizing Multilayered Neuronal Network Model.NHK Technical Monograph, No30., (1981) 3-25.
6. Hu-Sze-Tseu: Threshold Logic. University of California Press, Berkeley, 1965.
7. Labos,E.: Periodic and non-periodic motions in different classes of formal neuronal networks and chaotic spike generators.In Cybernetics and System Research 2(R.Trappl.,Ed.) Elsevier, Amsterdam, (1984) 237-243.
8. Kohonen,T.: Associative Memory. Springer, Berlin, 1977.
9. Lewis,P.E.II. and Coates,C.L.: Threshold Logic. Wiley, New York, 1967.
10. Loe,K.F. and Goto,E.: DC Flux Parametron. World Scientific Ser. in Computer Science Vol.6., Singapore, 1986.
11. McCulloch,W.S. and Pitts,W.H.: A Logical Calculus of the Ideas Immanent in Nervous Activity. Bull.Math.Biophys. 5 (1943) 115-133.

12. Mendel,J.M. and Fu,K.S.: Adaptive, Learning and Pattern Recognition Systems.Academic Press, New York, 1970.

13. Minsky,M. and Papert,S.: Perceptrons.The MIT Press,Cambridge,Mass.,1969.

14. Muroga,S.: Threshold Logics and Its Application. Wiley-Interscience, New York,1971.

15. Milsson,N.J.: Learning Machines.McGraw-Hill,New York, 1965

16. Novikoff,A.: On Convergence Proofs for Perceptrons. Proc.of the Symp.on Mathematical Theory of Automata. Polytechnic Institute of Brooklyn 12 (1962) 615-621.

17. Rosenblatt,F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Rev.65 (1958) 386-408.

18. Rosenblatt,F. Principles of Neurodynamics. Spartan Books, Macmillan, New York, 1962.

19. Sheng,C.L.: Threshold Logic. Academic Press, New York., 1969.

20. Widrow,B.: A Statistical Theory of Adaptation. In Adaptive Control Systems(eds.: Caruthers,F.P. and Levenstein,H.). Pergamon Press, Oxford, 1963, pp 97-122.

21. Winder,R.O.: The Fundamentals of Threshold Logic. In Applied Automata Theory (J.Tou Ed.). Academic Press, New York, 1968, pp 236-318.

22. Winder,R.O.: Chow Parameters in Threshold Logic. Memorandum at RCA Princeton Laboratory (April 1969) 45 pages.

Adaptive Stabilization Without High-Gain

Bengt Mårtensson

Department of Automatic Control
Lund Institute of Technology
Box 118, S-221 00 Lund
Sweden

Abstract: During the last few years there has been a very intense discussion on the applicability of *adaptive control* and on 'standard assumptions' made in the traditional theory. Some years ago, the question of *what is really the relevant information needed for successful adaptive control* was starting to receive some attention. The present work belongs to this tradition.

A very brief introduction to the concept of adaptive control is first given. The prototype problem of stabilizing an unstable, unknown plant is studied. The main result is the complete characterization of necessary and sufficient a priori knowledge needed for adaptive stabilization, namely knowledge of the order of any stabilizing controller. The concept of switching function controller is introduced, and some properties stated. 'The Turing Machine of Universal Controllers' is then presented. As the title indicates, this adaptive controller possessed the greatest stabilizing power a smooth, non-linear controller can have. The preceding works in this field have all dealt with variations on the theme of high-gain stabilization. This paper deals only with adaptive stabilization algorithms not requiring high-gain-stabilizability. Finally, the problem of stabilization to a possibly non-zero reference value is solved.

1. Introduction

The discipline of *Control Theory* studies the problem of achieving "satisfactory performance" of a *plant*, i.e. a dynamical system to be controlled, by manipulating the input u in order to e.g. keep the output y close to 0, or to follow a reference signal r. The most general problem of control theory can in loose terms be described as the following: *Given a set $\mathcal{G}$ of plants, we are to find one controller K that achieves "satisfactory performance" (or optimal in some sense) to each one of the plants $G \in \mathcal{G}$.* Figure 1 illustrates the concept. The dependence of the input u of the output is exactly the concept of *feedback*.

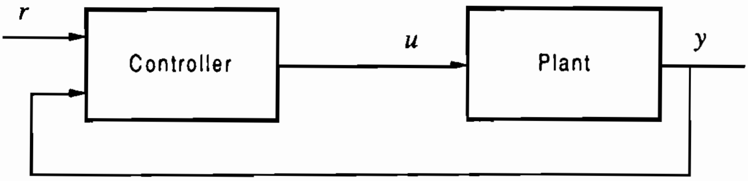


Figure 1. The Most General Control Configuration.

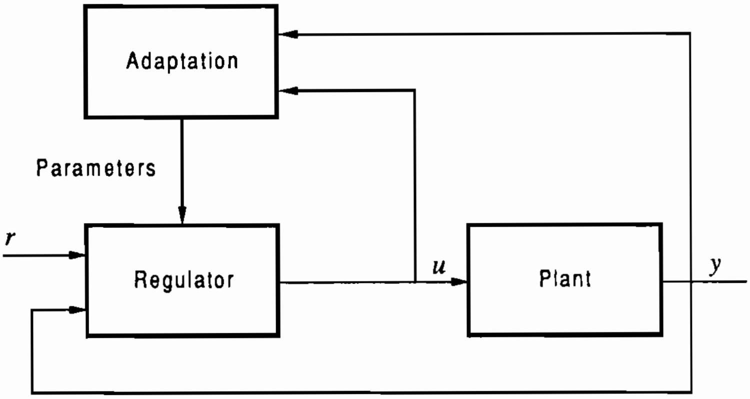


Figure 2. The General Adaptive Controller.

Adaptive Control is one—out of several other possible—approaches to solving this problem. It is an approach based on the concept of *learning*, i.e. the splitting of the 'true' state space of the non-linear system an adaptive controller constitutes in *parameters* and *states*. See Figure 2! The parameters reside in the "adaptation box", while the states reside in the "regulator box". The parameters are moving "slower" that the states, thereby motivating the values of the parameters as a state of knowledge on the dynamics of the plant.

Adaptive control is a vital subfield within control theory, with over 100 papers published every year. For an excellent overview of the field see [Åström].

In the end of the seventies and the beginning of the eighties, proof for convergence and stability of the commonly used adaptive schemes appeared. These proofs all required some variant of the following assumptions:

(i) A bound n^* on the order of the transfer function $g(s) = n(s)/d(s)$ is known.

(ii) The relative degree $r = \deg d(s) - \deg n(s)$ is known exactly.

(iii) The plant is minimum phase.

(iv) The sign of the 'instantaneous gain', i.e. the leading coefficient of $n(s)$*, is known.

This work is concerned with the fundamental limitations and possibilities of adaptive control, regardless of the particular algorithm used. In particular—are the four assumptions (i)-(iv) really necessary? To this end, what is believed to be the most fundamental problem is studied, namely the stabilization of an unstable plant. It can be argued that this is the "prototype problem", if we can do this there is hope for more achievements, and vice versa. It is also a very clean, quantitative problem.

We next give some more precise definitions for the sequel.

Definitions

Consider Figure 2! In general, with fixed values of the parameters, the dynamics in the states are assumed to be linear. Under this condition, we make the following definition.

* We assume that $d(s)$ is monic

Definition 1.1. Let the set of plants $\mathcal{G}$, its times $\mathcal{T}$, its input space $\mathcal{U}$, its output space $\mathcal{Y}$, and its space of reference signals $\mathcal{R}$ are given. Let l be a non-negative integer and $\mathcal{X}$ a vector bundle of rank l over the C^∞-manifold $\mathcal{M}$. We shall call the mapping

$$\mathcal{S} : \mathcal{Y} \times \mathcal{R} \times \mathcal{X} \longrightarrow \mathcal{U}$$

a *linear adaptive controller* with *state space* $\mathbb{R}^l$ and *parameter space* $\mathcal{M}$ if it is smooth in the sense of a control system, [Brockett], and for fixed $k \in \mathcal{M}$ the mapping $\mathcal{S}_k : \mathcal{Y} \times \mathcal{R} \times \mathbb{R}^l \longrightarrow \mathcal{U}$ is linear. That is, it can locally be written as

$$\begin{aligned}
\dot{z} &= F(t,k)z + G(t,k)y \qquad x \in \mathbb{R}^l \\
u &= H(t,k)z + K(t,k)y \\
\dot{k} &= f(y,r,t,z,k)
\end{aligned}$$

where F, G, H, K, and f are locally defined C^∞-functions. Here $x = (z^T, k^T)^T$ is a decomposition of the state of the controller corresponding to the local decomposition of $\mathcal{X}$ in $\mathbb{R}^l$ and $\mathcal{M}$. □

For a global, coordinate free description of a non-linear control system as a section of a certain pull-back bundle, see [Brockett].

With this definition, what makes a nonlinear controller into a linear adaptive controller is the (local) decomposition of the state space into a vector space times a manifold, together with linearity for fixed values of the parameters.

This definition covers the traditional approaches to adaptive control, namely model reference adaptive control and the self tuning regulator. Compare Figure 2!

Convergence of Adaptive Control

We will next make precise what we mean by convergence of a certain adaptive controller, controlling a certain plant. Only the stabilization problem, i.e. when $r \equiv 0$, will be considered. We restrict our attention to stabilization of strictly proper, time-invariant, linear plants described by finite dimensional differential equations, with vector spaces as their state space. That is, plants that can be written on state space form as

$$\begin{aligned}
\dot{x} &= Ax + Bu, \qquad x \in \mathbb{R}^n, \qquad u \in \mathbb{R}^m \\
y &= Cx, \qquad\quad\;\; y \in \mathbb{R}^p
\end{aligned} \qquad \text{(MIMO)}$$

Definition 1.2. We shall say that the linear adaptive controller $K \not\equiv 0$, controlling the plant G, whose state space is $\mathbb{R}^n$, converges, if, as $t \to \infty$, $\mathcal{M} \ni k$ converges to a finite value k_∞, while $\mathbb{R}^l \ni z \to 0$, and $\mathbb{R}^n \ni x \to 0$ as $t \to \infty$. □

Adaptive Control Problems

Finally, this is what shall be meant by an *adaptive control problem*.

Definition 1.3. We shall call the following an *adaptive control problem*: Let $\mathcal{G}$ be a set of plants. The adaptive control problem consists of finding a linear adaptive controller K, such that for any plant $G \in \mathcal{G}$, the controller K, controlling G, converges in the sense above. □

The 'size' of $\mathcal{G}$ can be considered as a measure of the uncertainty of the plant.

2. Necessary and Sufficient Conditions for Adaptive Stabilization

This section contains the complete characterization of the a priori knowledge needed to adaptively stabilize an unknown plant, namely the order of *any* fixed linear controller capable of stabilizing the plant. The necessity was proved in [Byrnes-Helmke-Morse], while the sufficiency was proved in [Mårtensson 1985]. A new proof of the sufficiency part is given, based on the results on switching functions presented in Section 5.

The Main Theorem

The following theorem is the most general result on adaptive stabilization.

THEOREM 2.1. *Let $\mathcal{G}$ be a set of plants of the type* (MIMO). *The necessary and sufficient a priori knowledge for adaptive stabilization is knowledge of an integer l such that for any plant $G \in \mathcal{G}$ there exists a fixed linear controller of order l stabilizing G.*

Proof of Necessity. See [Byrnes-Helmke-Morse]. ∎

The original proof of the sufficiency of this a priori information is the controller given in Section 6. The result can also be obtained by the method of switching functions introduced in Section 5.

We will devote the next sections to the development of some tools for proving this result.

3. A Viewpoint on Dynamic Feedback

In this section it is shown that, from a certain point of view, dynamic feedback can conceptually be replaced by static feedback.

The idea is very simple: the plant is augmented by a box of integrators, each with its own input and output. Static feedback is then applied to the augmented plant, i.e. the plant together with the integrators. The situation is depicted in Figure 3.

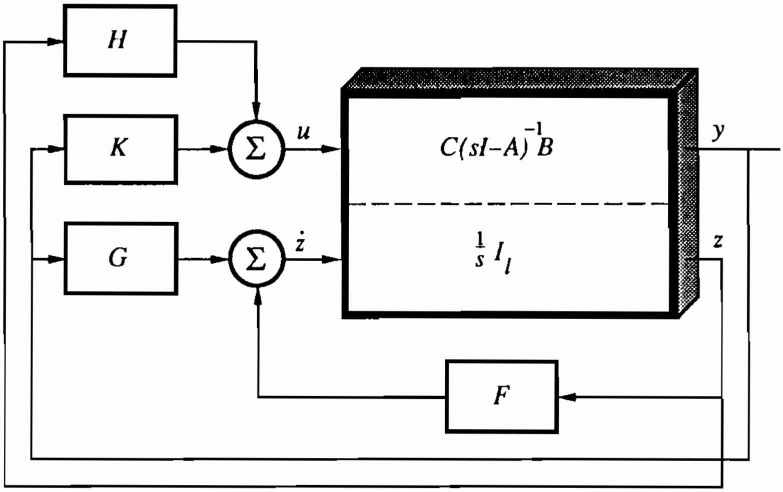


Figure 3. Dynamic feedback considered as static feedback.

More formally: Consider the following dynamic feedback problem: Given the plant

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad & x \in \mathbb{R}^n, \qquad & u \in \mathbb{R}^m \\ y &= Cx, & y \in \mathbb{R}^p & \end{aligned} \qquad \text{(MIMO)}$$

and the controller

$$\begin{aligned} \dot{z} &= Fz + Gy, \qquad z \in \mathbb{R}^l \\ u &= Hz + Ky \end{aligned}$$

It is easy to see that this is equivalent to the static feedback problem

$$\begin{aligned} \dot{\tilde{x}} &= \tilde{A}\tilde{x} + \tilde{B}\tilde{u} \\ \tilde{y} &= \tilde{C}\tilde{x} \\ \tilde{u} &= \tilde{K}\tilde{y} \end{aligned} \qquad \text{(MIMOA)}$$

where

$$\tilde{x} = \begin{pmatrix} x \\ z \end{pmatrix} \qquad \tilde{u} = \begin{pmatrix} u \\ \dot{z} \end{pmatrix} \qquad \tilde{y} = \begin{pmatrix} y \\ z \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \qquad \tilde{B} = \begin{pmatrix} B & 0 \\ 0 & I \end{pmatrix} \qquad \tilde{C} = \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix} \qquad \tilde{K} = \begin{pmatrix} K & H \\ G & F \end{pmatrix}$$

Remark 3.1. This observation might seem very powerful at least at first sight, but note the highly non-generic nature of $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$. This means e.g. that results on generic pole placement by static output feedback, see [Brockett-Byrnes], [Byrnes], do not translate at all. □

4. Estimation of the Norm of the State

In this section a lemma is proven, which gives an estimate of the norm of the state x of (MIMO), expressed in the L^2 norm of y and u. The lemma has a simple corollary, which implies that, under mild conditions, to show that an adaptive algorithm converges and stabilizes the plant, it is enough to show that the controller stays bounded. First we give the continuous time version.

LEMMA 4.1. *Assume that the linear system* (MIMO) *is observable. Then:*

(i) *For all $x(0)$, there are constants c_0 and c_1 such that*

$$\|x(t)\|^2 \leq c_0 + c_1 \left(\int_0^t \|y(\tau)\|^2 \, d\tau + \int_0^t \|u(\tau)\|^2 \, d\tau \right)$$

for all $u(\,.\,)$, and $t \geq 0$. Here c_0 does not depend on t or u; and c_1 does not depend on t, $u(\,.\,)$ or $x(0)$.

(ii) *For $T > 0$, c_1 can be taken so*

$$\|x(t)\|^2 \leq c_1 \left(\int_{t-T}^t \|y(\tau)\|^2 \, d\tau + \int_{t-T}^t \|u(\tau)\|^2 \, d\tau \right)$$

for all t, $u(\,.\,)$, and $x(t-T)$.

Remark 4.2. In (ii) we can consider c_1 as a function of T. This function can clearly be taken continuous and decreasing. □

Remark 4.3. Note that, for t bounded from below (i) follows trivially from (ii). Also note that the c_0-term is necessary if and only if we allow arbitrary small $t > 0$. □

Remark 4.4. It is not possible to improve the result by deleting the integral of u. A simple counter-example can be constructed by letting (MIMO) be an integrator, the initial state $x(0) = 0$, and the input $u(\tau) = \delta\left(\tau - (t - \varepsilon)\right)$, for some small $\varepsilon > 0$. Choose coordinates in the state space so that $x = y$. Then clearly $x(t) = 1$, and $\int y^2 d\tau = \varepsilon$, so by letting $\varepsilon \to 0$, we arrive at a contradiction. The lemma is true without the u-dependent term if and only if $G(s)$ has a proper left-inverse. □

Proof. In an obvious operator notation we have

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)}Bu(\tau)\,d\tau =: L_1^t x(0) + L_2^t u(\,.\,) \qquad (\smile)$$

$$y(\,.\,) = L_3 x(0) + L_4 u(\,.\,) \qquad (\frown)$$

where L_1^t, L_2^t, L_3, and L_4 are bounded linear operators between suitable Hilbert spaces. We first prove (ii). Let $T > 0$ be given. By using time invariance, it is enough to show (ii) for $t = T$. From observability, $(\frown)$ can be solved with respect to $x(0)$, i.e. $x(0)$ is the image of $y(\,.\,)$ and $u(\,.\,)$ under a continuous linear mapping. Inserted into $(\smile)$, this proves (ii).

By Remark 4.3, it only remains to show (i) for small t, say $t \leq 1$. For this, note that the operators $\mathcal{L}_1 = \{L_1^t : 0 \leq t \leq 1\}$ and $\mathcal{L}_2 = \{L_2^t : 0 \leq t \leq 1\}$ are uniformly bounded by, say, k_1 and k_2. From these observations, (i) follows (for $t \leq 1$) from $(\smile)$, since $\int_0^t \|u\|^2\,d\tau \leq \int_0^t \left(\|u\|^2 + \|y\|^2\right)\,d\tau$. The proof is finished. ■

A Useful Corollary

The lemma has the following immediate corollary, which will be used in the connection with adaptive stabilizers. We make the following definition:

Definition 4.5. A function $f : \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ will be called *L^2-compatible* if it is satisfies a Lipschitz-condition and there exists a constant $c > 0$ such that $f(y, u, k, t) \geq c\left(\|y\|^2 + \|u\|^2\right)$ for all k and all t. □

The name is motivated by the fact that for f being an L^2-compatible function, we can estimate the L^2-norm of (y, u) by the integral of f, as will be done in the proof of the following corollary.

COROLLARY 4.6. *Consider the plant* (MIMO), *and let* $u(\,.\,)$ *be a continuous time-function. Let k satisfy*

$$\dot{k} = f(y, u, k, t), \quad k(0) = k_0$$

where f is an L^2-compatible function. Then, if k converges to a finite limit k_∞ as $t \to \infty$, it holds that $\|x(t)\| \to 0$ as $t \to \infty$.

Proof. Clearly

$$\int_0^\infty \left(\|y\|^2 + \|u\|^2\right) dt \leq \frac{1}{c}\int_0^\infty f(y, u, k, t)\,dt = \frac{1}{c}\left(k_\infty - k_0\right) < \infty$$

Thus, for any $T > 0$, the right hand side of (ii) in Lemma 4.1 approaches zero when t approaches infinity. The corollary follows. ■

Remark 4.7. In previous 'universal' stabilizing algorithms, the step of showing that $x(t) \to 0$ as $t \to \infty$ has involved a minimum phase argument. This is not required here. □

5. Switching Function Controllers

In this section we will deal with the following problem: We want to adaptively stabilize an unknown plant G of type (MIMO), for which we know that G belongs to a set $\mathcal{G}$. Here $\mathcal{G}$ is a set of plants for which there exists a finite or countable set of controllers $\mathcal{K}$, such that for any $G \in \mathcal{G}$, there is at least one controller $K \in \mathcal{K}$ such that the control law $u = Ky$ will stabilize G.

A heuristically appealing algorithm for stabilizing the unknown plant G would be to try each one of the K's for ε units of time, until we find one that stabilizes the system. It is shown in [Mårtensson 1986] that this is possible if and only if we know a bound on the McMillan degree of the plants belonging to $\mathcal{G}$. Instead we try each one of the controllers for some time, according to some criterion, in a way that will hopefully converge, and thus will switch among the controllers only a finite number of times. A *switching function* is a criterion of this type.

The concept of switching function was first introduced in [Willems-Byrnes], where the set of plants $\mathcal{G}$ under consideration was single-input, single-output, minimum phase plants of relative degree one. In [Byrnes-Willems] this was generalized to multivariable plants satisfying analogous conditions.

In the remainder of this section, we introduce the pertinent concepts formally, and give a result on switching function based adaptive stabilization.

Definitions

Definition 5.1. Let $s(k)$ be a function of a real variable, and $\{\tau_i\}_{i=0}^{\infty}$ a sequence of increasing real numbers. For $r = 2, 3, \ldots, \aleph_0$, we shall say that $s(k)$ is a *switching function of rank* r with *associated switching points* $\{\tau_i\}$, if $s(k)$ is constant for $k \notin \{\tau_i\}$, and, for all $a \in \mathbb{R}$, $s(\{k \geq a\}) = \{1, \ldots, r\}$. Further, just as a notational convenience, we require a switching function to be right continuous. □

Remark 5.2. Note that it follows from the definition that infinity is the only limit point of the sequence $\{\tau_i\}$. □

By switching function controller we shall mean the following.

Definition 5.3. For $r = 2, 3, \ldots, \aleph_0$, let $\mathcal{K} = \{K_1, \ldots, K_r\}$ be a set of controllers, with card $\mathcal{K} = r$. Let f be a Lipschitz-continuous function and $s(k)$ a switching law of rank r. A controller of the type

$$\begin{aligned} u &= K_{s(k)}y \\ \dot{k} &= f(y, u, k, t) \end{aligned} \qquad \text{(SFC)}$$

will be called a *switching function controller*. □

Remark 5.4. Note that in general the control law $u = K_i y$ must be interpreted in an operator-theoretic way, not as a matrix multiplication. □

Remark 5.5. The way (SFC) is written requires all the controllers $K_1, \ldots, K_r$ to be simultaneously connected to the output of the plant, while the switching law chooses which controller's output to connect to the plants input, at least if the K_i's contain dynamics. For r large or infinite, this is clearly not a practical way of implementing a controller. However, if all the controllers have a (not necessarily minimal) realization on a state space of a certain dimension, then this difficulty can be circumvented by considering the augmented plant as in Section 3, and considering the controllers as static controllers. □

For further reference, we shall make clear what we mean by a set of controllers stabilizing a set of plants in some sense.

Definition 5.6. Let f an L^2-compatible function, $\mathcal{G}$ a set of plants of the type (MIMO), all of which having the same number of inputs and outputs, and $\mathcal{K}$ a set of controllers of compatible dimensions. For $k_0 \in \mathbb{R}$, let k be the unique solution to $\dot{k} = f(y, u, k, t)$, $k(0) = k_0$. We shall say that $\mathcal{K}$ *is stabilizing for* $\mathcal{G}$ *with respect to* f (or *is* f*-stabilizing for* $\mathcal{G}$) if the following holds: For any plant $G \in \mathcal{G}$ there is a controller $K \in \mathcal{K}$ and constants c, T such that the control law $u = Ky$ will stabilize G in the sense that

$$\int_{t_0}^{\infty} f(y, u, k, t)\, dt \leq c\|x(t_0)\|^2$$

for all $x(0) \in \mathbb{R}^n$ and for all $k_0 \in \mathbb{R}, t_0 \geq T$. □

Remark 5.7. In particular, the left hand side stays finite, so it follows from Corollary 4.6 that $x(t) \to 0$ as $t \to \infty$. It also follows that the solution to the differential equation is indeed globally defined. □

Remark 5.8. By considering singleton sets in the definition, it is clear what we shall mean by the statement *the controller* K *stabilizes the plant* G *with respect to* f. □

The Main Result on Switching Functions

With the machinery developed so far, we can now easily prove the following results on switching function controllers.

THEOREM 5.9. *Suppose that f is an L^2-compatible function, and that the set of controllers $\mathcal{K}$ is f-stabilizing for the set of plants $\mathcal{G}$. Then there is a sequence $\sigma = \{\tau_i\}$ such that for $s(k)$ any switching function of rank equal to* card $\mathcal{K}$, *with associated switching points $\{\tau_i\}$, the control law* (SFC) *will stabilize any plant $G \in \mathcal{G}$ in the sense that for all $x(0), k(0)$, it holds that $\|x(t)\| \to 0$ as $t \to \infty$, while k converges to a finite limit. Further, there is a 'universal' switching point sequence σ, independent of the individual set $\mathcal{G}$.*

Proof. The steps in the proof are the following: To say that the theorem is false is to say that for all switching sequences, there is a switching function with the stated properties such that stabilization does not take place. It will be shown that, if stabilization does not take place, the sequence $\{\tau_i\}$ has to satisfy a certain requirement, depending on $\mathcal{G}$, namely ($\mathcal{L}$) below. A sequence σ is given, with the property that for all allowed $\mathcal{G}$, the requirement is violated. We conclude that with this very sequence stabilization takes place, which will establish the theorem.

From Corollary 4.6, and since k is increasing, it follows that in order to show stabilization it is enough to show that k is bounded. By the definition of switching function, this is equivalent to the statement that s, considered as a function of time, only switches a finite number of times. So we assume that this is not the case, and investigate the implications of this assumption.

Consider an arbitrary, but fixed, $G \in \mathcal{G}$. Say that controller K_i is f-stabilizing for G, and that the controller K_i is used with start at time t_0. That is, $k(t_0) = \tau_j$, where $s(\tau_j) = i$. By the assumptions, this will happen for arbitrarily large k and t. Therefore, with T as in Definition 5.6, we shall make the assumption that $t_0 \geq T$.

The assumption that s switches an infinite number of times implies that we will reach the next switching point τ_{j+1} after a finite time. But this is exactly the statement that

$$\int_{t_0}^{\infty} f(y,u,k,t)\,dt \geq \tau_{j+1} - \tau_j \qquad (¿)$$

where the left hand side, by assumption finite, is evaluated as if the controller K_i was used forever. We will show that the sequence $\{\tau_i\}$ can be taken in a way so that (¿) cannot be satisfied for j sufficiently large, which will prove the theorem.

By definition of f being L^2-compatible, there is a c, so that the left hand side of (¿) can be estimated as

$$\int_{t_0}^{\infty} f(y,u,k,t)\,dt \leq c\|x(t_0)\|^2$$

Using the same argument as in the proof of Corollary 4.6, it follows from Lemma 4.1, part(i), that for all $x(0)$, there exist constants c_0 and c_1 such that

$$\|x(t)\|^2 \leq c_0 + c_1 k(t)$$

for all t. Substituting $t = t_0$, $k = \tau_j$, and combining the last two estimates, we see that a necessary condition for (¿) to be satisfied, is that

$$\tau_{j+1} - \tau_j \leq cc_0 + cc_1\tau_j \qquad (£)$$

But there are sequences $\{\tau_i\}$ such that, for any c, c_0, c_1, the statement (£) will be false for all sufficiently large j. This is the case e.g. for the sequence defined by

$$\begin{aligned} \tau_{j+1} &= \tau_j^2, \qquad i = 2,3,\ldots \\ \tau_1 &= 2 \end{aligned}$$

Therefore, with a switching sequence like this chosen, the assumption of s to switch infinitely many times leads to a contradiction. Since G was arbitrary, the proof is complete. ∎

Proof of Sufficiency in Theorem 2.1

The proof is a fairly straightforward application of Theorem 5.9. Consider a controller in the spirit of Section 3, namely as a constant $M \times P$-matrix, where $M := m + l$, and $P := p + l$. The set of controllers $\mathcal{K}$ is taken to be all such with rational coefficients, i.e. $\mathcal{K} := \mathbb{Q}^{M\times P}$. Let f be defined as $f(y,u,k,t) = \|y\|^2 + \|u\|^2$. This is an L^2-compatible function. A stabilizing controller places the closed loop poles in the open left half plane. The poles depend continuously of the parameters in the controller. Since $\mathcal{K}$ is dense in the space of all controllers of order l, i.e. $\mathbb{R}^{M\times P}$, $\mathcal{K}$ is thus f-stabilizing for $\mathcal{G}$. Theorem 5.9 establishes the existence of a switching function such that the corresponding switching function controller (SFC) stabilizes any plant in $\mathcal{G}$. This completes the proof. ∎

Remark 5.10. By some additional effort, an explicit algorithm based on the ideas in the proof can be constructed. □

In [Mårtensson 1986], it is shown that the controller can also be taken to be continuous by 'smoothing-out' the discontinuities. Another approach is presented in the next section.

6. "The Turing Machine" of Universal Stabilizers

In this section we will consider the problem of adaptively stabilizing the plant (MIMO), given only the a priori information that an integer l is known, such that there exists a fixed linear time-invariant controller of order l that will stabilize the system. An explicit algorithm for this will be given. This will be given only very briefly, without proof. A more detailed discussion, including a discrete time version, is given in [Mårtensson 1986]. The proof is also given in [Mårtensson 1985].

As shown in Section 3, it suffices to consider adaptive control based on static feedback. A (fixed) controller is then nothing but a matrix in $\mathbb{R}^{M\times P}$, where M and P denotes the number of inputs and outputs to the augmented plant (MIMOA). For the sequel, we assume that this augmentation has been done, and therefore we only consider static feedback. Since a (fixed) controller achieving internal stability to the closed loop system places all the eigenvalues in the open left-half plane, (or the open unit disc) and these depend continuously on the parameters of the controller, there is an open set in parameter space yielding a stable system. Equip $\mathbb{R}^{M\times P}$ with the norm

$$\|A\|^2 = \sum_{i,j}(A)_{ij}^2$$

Thus we identify $\mathbb{R}^{M\times P}$, as a normed space, with $\mathbb{R}^{MP}$, equipped with the Euclidean norm. For the rest of this section, we let $\|\,.\,\|$ denote the this vector norm, or the corresponding induced matrix norm. Partition $\mathbb{R}^{M\times P} = \mathbb{R}^+ \times S^{MP-1}$ in a natural way, namely by dividing out the norm of every non-zero matrix. S^{MP-1} is now the unit sphere in a normed space of controllers. Let the controller be

$$\tilde{u} = g(h(k))N(h(k))\tilde{y} \tag{1}$$

$$\dot{k} = \|\tilde{y}\|^2 + \|\tilde{u}\|^2 \tag{2}$$

where

$$N(h) \text{ is `almost periodic' and dense on } S^{MP-1} \tag{3}$$

while h and g are continuous, scalar functions satisfying

$$h(k) \nearrow \infty, \qquad k \to \infty \tag{4}$$

$$\text{There exists an } a \text{ such that } \left|\frac{dg}{dh}\right| < a \tag{5}$$

$$g\big(\{\alpha\nu + (\beta,\gamma)\}_{\nu=n}^{\infty}\big) = \mathbb{R}^+ \quad \text{for } n \in \mathbb{Z}, \quad \alpha \neq 0, \quad \gamma > \beta \tag{6}$$

$$kg\big(h(k)\big)\frac{dh}{dk} \to 0, \qquad k \to \infty \tag{7}$$

THEOREM 6.1. *Consider the minimal plant* (MIMO). *Assume that l is chosen so that there exists a fixed linear stabilizing controller, and that the augmentation to (MIMOA) has beed done. The controller (1) – (2), subject to (3) – (7), will then stabilize the system in the sense that*

$$\big(x(t), z(t), k(t)\big) \to (0, 0, k_\infty) \qquad \text{as } t \to \infty$$

where $k_\infty < \infty$.

One set of functions satisfying (4) – (7) is

$$h(k) = \sqrt{\log k}, \qquad k \geq 1$$

$$g(h) = \sqrt{h}\left(\sin\sqrt{h} + 1\right)$$

The construction of the function $N(h)$ is a standard exercise in calculus on manifolds. One such is given explicitly in the references cited above.

7. Setpoint Stabilization

In this section it will be shown how to introduce integrators in the loop, thereby being able to track a constant reference signal with error approaching zero asymptotically. The problem is as follows: Let $\mathcal{G}$ be a set of plants as before, and $r \in \mathbb{R}^p$ a given constant (a reference value). We want to find a controller K such that for all $G \in \mathcal{G}$ it holds that

$$\begin{aligned} x &\to \hat{x} \qquad (= \text{constant}) \\ y &\to r \\ z &\to \hat{z} \\ k &\to k_\infty \end{aligned}$$

as $t \to \infty$.

Tracking with Zero Error Asymptotically

Every engineer knows that you cannot track a constant reference signal with zero error asymptotically without having integrators in the loop*. The analogous statement of course applies to multivariable plants. Conversely, with integrators in every loop, the asymptotic tracking error is zero, provided the closed loop system is stable. This shall mean that every fixed linear combination of rows or columns of the matrix $G(s)$ has a pole at the origin.

The construction for adaptively stabilizing a plant, with a constant reference signal $r(t) \equiv r_0$ is very simple: We just put the diagonal 'precompensator' $\bar{K} = s^{-1}I_m$ in front of the plant. For the sequel, consider the problem of adaptively stabilizing the 'plant' $\tilde{G}(s) := G(s)\bar{K}(s)$ instead. This is depicted in Figure 4.

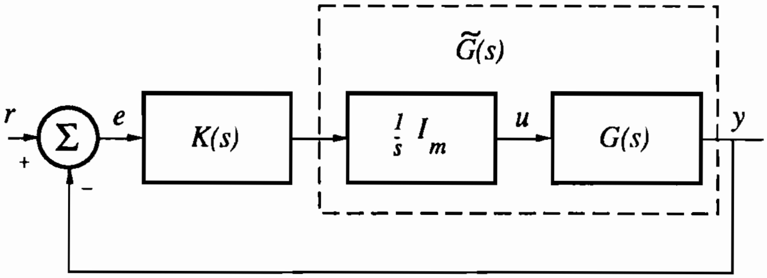


Figure 4. Setpoint Stabilization by Introducing Integrators.

More precisely, we have the following result.

* Quick and *dirty* proof: $y(\infty) = r(\infty) \iff g(0)/(1 + g(0)) = 1 \iff g(0) = \infty$ ∎

THEOREM 7.1. *Assume that the controller K stabilizes the plant G in the usual sense. Let $r \in \mathbb{R}^p$ be given. Suppose that there exists a unique $\hat{x}$ such that*

$$\begin{aligned} 0 &= A\hat{x} \\ r &= C\hat{x} \end{aligned}$$

Let K operate on $-e := y - r$ instead of y. Then, as $t \to \infty$ it holds that $y \to r$, $x \to \hat{x}$, and $k \to k_\infty$.

Remark 7.2. The uniqueness follows automatically from observability. □

Proof. We can write

$$\begin{aligned} \frac{d}{dt}(x - \hat{x}) &= A(x - \hat{x}) + Bu \\ y - r &= C(x - \hat{x}) \end{aligned}$$

So, assuming we have a proof of a theorem saying that the assumptions are satisfyed, we only have to substitute all occurances of x by $x - \hat{x}$, and all occurances of y by $y - r$ in order to construct a proof of the above theorem for the case in question. So Theorem 7.1 is really a meta-theorem on adaptive stabilization. ■

The most natural use of Theorem 7.1 is in the form of the following corollary:

COROLLARY 7.3. *Assume K stabilizes $G(s) = \frac{1}{s}G_1(s)$, where $\det G \neq 0$. Then with error feedback K will also do set-point stabilization for any $r \in \mathbb{R}^p$.*

Extensions and Comments

Everyone with experience of practical control engineering knows that plants of high relative degree are very hard to control manually, but often fairly simple to control with simple controllers, such as standard PID-controllers. Something similar is true about adaptive control. We need some extra dynamics in our controllers, that is all. By preceding the plant by integrators as in the construction above, the minimal order of a stabilizing controller might increase. A classical control engineer would say that we do this at the expense of a decrease of the phase by 90°, and thus need some extra phase advancing to stabilize the plant.

The same argument may be used to introduce multiple integrators in the loop, thus being able to track ramps of higher order.

References

ÅSTRÖM, K. J. (1983): "Theory and Applications of Adaptive Control—A Survey", *Automatica* **19** no. 5, 471–486.

BROCKETT, R. W. (1983): "Asymptotic Stability and Feedback Stabilization", in BROCKETT, R. W., R. S. MILLMAN, and H. J. SUSSMANN, (Eds.): *Differential Geometric Control Theory*, Birkhäuser, Boston.

BROCKETT, R. W. and C. I. BYRNES (1981): "Multivariable Nyquist Criteria, Root Loci, and Pole Placement: A Geometric Viewpoint", *IEEE Transactions on Automatic Control* **AC-26** no. 1, 271–284.

BYRNES, C. I. (1983): "Control Theory, Inverse Spectral Problems, and Real Algebraic Geometry", in BROCKETT, R. W., R. S. MILLMAN, and H. J. SUSSMANN, (Eds.): *Differential Geometric Control Theory*, Birkhäuser, Boston.

BYRNES, C. I., U. HELMKE, and A. S. MORSE (1985): "Necessary Conditions In Adaptive Control", in BYRNES, C. I. and A. LINDQUIST (Eds.): *Modelling, Identification and Robust Control, Proc. 7th International Symposium on the Mathematical Theory of Network and Systems*, North Holland, Amsterdam.

BYRNES, C. I. and J. C. WILLEMS (1984): "Adaptive Stabilization of Multivariable Linear Systems", *Proceedings of the 23rd IEEE Conference on Decision & Control*, Las Vegas, NV, pp. 1574–1577.

MÅRTENSSON, B. (1985): "The Order of Any Stabilizing Regulator is Sufficient A Priori Information for Adaptive Stabilization", *Systems & Control Letters* **6** no. 2, 87–91.

——— (1986): "Adaptive Stabilization", Report CODEN: LUTFD2/(TFRT-1028)/1-122/(1986), Dept. of Automatic Control, Lund Institute of Technology, Lund, Sweden. Ph.D. Thesis.

WILLEMS, J. C. and C. I. BYRNES (1984): "Global Adaptive Stabilization in the Absence of Information on the Sign of the High Frequency Gain", *Proc. INRIA Conf. on Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences* 62, 49–57, Springer-Verlag, Berlin.

Topological properties of observability for a system of parabolic type

S. Miyamoto

International Institute for Applied Systems Analysis, Laxenburg, Austria

ABSTRACT

The purpose of the present paper is to demonstrate topological properties of observable regions in a distributed parameter system. A parabolic partial differential equation with constant coefficients is considered. According to Sakawa's definition, observability is defined to be the possibility of the unique determination of the initial value by point measurements, or by spatially averaged measurements. Furthermore, n-mode observability is defined to be the possibility of the unique determination of the coefficients corresponding to the first n eigenvalues, based on the expansion of the solution by eigenfunctions. Then it is proved that n-mode observability is generic, that is, open and dense, whereas observability is shown to be dense in the whole space of measurements. In case of point measurements, it is shown that observability is valid almost everywhere with respect to the Lebesque measure. Moreover genericity of n-mode controllability and the related properties of controllability will be shown for the dual systems with controls.

1. Introduction

The problem of observability in distributed parameter systems has a different aspect from that in lumped parameter systems, because the former includes the specification of the spatial distribution of measurements, which we need not take into account for ordinary differential equations. For example, in distributed systems we have some local information of the state variable such as the point measurement which should be extended to the whole spatial domain. Therefore some efforts have been devoted to the unique determination of the state from local measurements.

Goodson and Klein [1] considered the problem of uniqueness with respect to point observation. Moreover they proposed the definition of n-mode observability, which means the coefficients that correspond to first n eigenvalues in the eigenfunction expansion of the initial state is uniquely determined. Furthermore, Sakawa [4] considered a broader class of parabolic systems and gave the conditions of observability with respect to point measurement and spatially averaged measurement.

In view of their results, the measurement space can be divided into two regions, one where observability holds and the other where some portions of the state is "unobservable". Here a problem of topological properties of the observable region arises. For example, in case of lumped parameter systems, observability has been proved to be generic, that is, open and dense in the whole domain of definition (cf. Wonham [6]).

We consider here this problem with respect to a class of parabolic differential equations and examine whether observability and n-mode observability are generic, dense, or not in the space of measurement.

2. Preliminary consideration

This section depends mainly on Sakawa [4]. Let D be an open bounded region in n-dimensional Euclidean space $\mathbf{R}^n$ (n >0) with a smooth boundary ∂D. Then we consider the following system:

$$\frac{\partial u}{\partial t}(t,x) = Au(t,x) \qquad (t,x)\in(0,T)\times D \tag{1}$$

$$A = \Delta - a_0 = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i} - a_0 \tag{2}$$

$$c_1 u(t,\xi)+(1-c_1)\frac{\partial u}{\partial \nu}(t,\xi) = 0 \qquad (t,\xi)\in(0,T)\times\partial D \tag{3}$$

$$0 \leq c_1 \leq 1$$

where a_0 is a real constant or an analytic function, c_1 is a real constant, and ν is the exterior normal to the boundary ∂D.

We assume the initial condition to be

$$u(0,x) = u_0(x) \qquad x\in D \tag{4}$$

Let the eigenvalues and the corresponding eigenfunctions be

$$\{\lambda_i, \varphi_{ij}; j = 1,2,\ldots,m_i, i = 1,2,\ldots\} \tag{5}$$

and assume that the multiplicity of the eigenvalues is finite:

$$\sup_i m_i = m < +\infty \tag{6}$$

Then the solution of the system and the initial condition are represented as:

$$u(t,x) = \sum_{i=1}^{\infty} \exp(-\lambda_i t) \sum_{j=1}^{m_i} u_{ij} \varphi_{ij}(x) \tag{7}$$

$$u_0(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} u_{ij} \varphi_{ij}(x) \tag{8}$$

respectively.

Then we consider the following observations described by Sakawa [4]:

$$(type1) \quad y_k(t) = \int_D w_k(x) u(t,x) dx \quad 1 \le k \le N \tag{9}$$

$$(type2) \quad y_k(t) = u(t,x_k) \quad 1 \le k \le N \tag{10}$$

On the other hand we describe the definition of observability and n-mode observability; the latter was studied by Goodson and Klein [1].

Definition 1

The system described by (1), (2), and (3) is said to be observable (resp. observable in time T) if the initial state $u_0(x)$ can be uniquely determined from the observation $Y(t) = (y_1(t), y_2(t), \ldots, y_N(t))$, $0 < t < \infty$. (resp. $0 < t < T$) .

Definition 2

The system described by (1), (2), and (3) is said to be n-mode observable (resp. n-mode observable in T) if u_{ij}, $j = 1,2,\ldots,m_i$, $i = 1,2,\ldots,n$ in (6) and (7) can be uniquely determined from the observation $Y(t) = (y_1(t), \ldots, y_N(t))$, $0 < t < \infty$ (resp. $0 < t < T$).

Remark 1

As is shown in Sakawa [4], observability in T (resp. n-mode observability in T) means observability (resp. n-mode observability) in view of analytic property of the solutions. Therefore, we do not distinguish the two notions below.

Then we have the following two propositions:

Proposition 1. (Sakawa [4])

For type 1 (resp. type 2) measurement, assume $w_{ij}^k = (w_k, \varphi_{ij})$ (resp. $w_{ij}^k = \varphi_{ij}(x_k)$), where (,) denotes the inner product of L^2-space. Then in order that the system (1), (2), and (3) is observable, it is necessary and sufficient that

$$rank\ W_i \doteq m_i \quad \text{for } all\ i \geq 1 ,$$

for

$$W_i = \begin{bmatrix} w_{i1}^1 & w_{i2}^1 & \cdots & w_{im_i}^1 \\ w_{i1}^2 & w_{i2}^2 & \cdots & w_{im_i}^2 \\ \vdots & \vdots & \cdots & \vdots \\ w_{i1}^N & w_{i2}^N & \cdots & w_{im_i}^N \end{bmatrix} \tag{11}$$

where we assume

$$\sum_{i=2}^{\infty} \frac{1}{(\lambda_i - \lambda_1)^2} < +\infty , \quad \sup_{i,j} |\varphi_{ij}(x)| < +\infty \tag{12}$$

in case of type 2 measurement.

Proposition 2

Under the hypotheses in Proposition 1, in order that the system (1), (2), and (3) is n-mode observable for $n > 0$, it is necessary and sufficient that

$$rank\ W_i = m_i \quad \text{for } all\ 1 \leq i \leq n .$$

The proof of Proposition 2 is a slight modification of that in Proposition 1 [4] and is omitted.

3. n-mode observability

Roughly speaking, the word "genericity" expresses that a property is valid at "almost all" points of the domain of definition. This notion is widely used in the theory of dynamical systems [2].

Definition 3

A property P defined on a topological space S is said to be generic if P is valid on an open and dense set in S.

Hence, in order that P is generic, it is sufficient to show that the subset where P does not hold is closed and nowhere dense (it has no interior points).

First, we suppose $N > m = \sup m_i$. otherwise the system can not be made observable. Then the following lemma is obvious.

Lemma 1

Assume that $\det \bar{W}_i(j)$, $j = 1,2,\ldots,\begin{bmatrix} N \\ m_i \end{bmatrix}$ be a m_i-th order minor determinant of W_i and let

$$D_i = {\sum_j}' \mid \det \bar{W}_i(j) \mid$$

where $\mid \; \mid$ denotes absolute value and the sum ${\sum_j}'$ is taken for all the m_i-th order minor determinants. Then, in order that

$$rank \; W_i = m_i$$

it is necessary and sufficient that

$$D_i \neq 0 .$$

Remark 2

In the below, D_i in the above lemma is sometimes represented as $D_i(w)$, since it is a function of the measurement w.

Lemma 2

Let

$$w^l(x) = (w_1^l(x), w_2^l(x), \ldots, w_N^l(x)) \in L^2(D)^N .$$

If the sequence $\{w^l, \; l=1,2, \cdots\}$ converges to w in $L^2(D)^N$, then the corresponding $\{D_i(w^l)\}$ converges to $D_i(w)$.

(Proof) Since $\{\varphi_{ij}\}$ is complete, we can expand w_k^l and w_k as:

$$w_k^l = \sum_{i,j} d_{kij}^l \varphi_{ij} , \quad w_k = \sum_{i,j} d_{kij} \varphi_{ij} .$$

From the assumption we have

$$d_{kij}^l \to d_{kij} \quad \text{for } all \; k, i, j .$$

Since D_i is a finite sum of the absolute values of the finite polynomials of d_{kij}, it is clear that

$$D_i(w^l) \to D_i(w) .$$

Lemma 3

The set $\{w \mid D_i(w) = 0, w \in C(D)\}$ has no interior points, where $C(D)$ is the class of continuous functions on D.

(Proof) For simplicity, we write an arbitrary one of $\overline{W}_k(j)$'s as V. Letting $l = m_k$, we assume that

$$V = \begin{vmatrix} (w_1, \varphi_{k1}) & (w_1, \varphi_{k2}) & \cdots\cdots & (w_1, \varphi_{kl}) \\ (w_2, \varphi_{k1}) & (w_2, \varphi_{k2}) & \cdots\cdots & (w_2, \varphi_{kl}) \\ \cdot\cdot & \cdot\cdot & \cdots\cdots & \cdot\cdot \\ (w_l, \varphi_{k1}) & (w_l, \varphi_{k2}) & \cdots\cdots & (w_l, \varphi_{kl}) \end{vmatrix} \tag{13}$$

without loss of generality. Considering V to be a function of $w_1(x)$ and expanding it with respect to the first row, we have the following:

$$\det V = \sum_{j=1}^{l} f_j \int w_1(x)\varphi_{kj}(x)dx = 0$$

where f_j's are functions of $w_2, \ldots\ldots, w_l, \varphi_{k1}, \ldots\ldots, \varphi_{kl}$. Let

$$\varphi(x) = \sum_{j=1}^{l} f_j \ \varphi_{kj}(x)$$

then it follows that

$$\det V = \int w_1(x)\varphi(x) = 0 .$$

If we assume that $\varphi(x)$ is not identically zero, then we can show that in an arbitrary neighborhood of w_1 in $C(D)$, there exists a function $\bar{w}$ such that

$$\int_D \bar{w}(x)\varphi(x)dx \neq 0 .$$

If we assume w_1 not to be identically zero on D, then there exists an open sphere B such that

$$w_1(x)\varphi(x) > 0 \ , \ w_1(x) > 0 \ , \ x \in B$$

or

$$w_1(x)\varphi(x) > 0 \ , \ w_1(x) < 0 \ , \ x \in B$$

For any $\varepsilon > 0$, there is an infinitely differentiable function ζ satisfying

$$support\ (\zeta) \subset B \ , \quad 0 \leq \inf_{x \in D} \zeta(x),\ \sup_{x \in D} \zeta(x) \leq \varepsilon .$$

Then we take $\bar{w}$ to be

$$\bar{w}(x) = w_1(x)+\zeta(x)\ ,\ \ x\in D\ \ \text{if } w_1(x')>0\ ,\ x'\in B$$

or

$$\bar{w}(x) = w_1(x) - \zeta(x)\ ,\ \ x\in D\ \ \text{if } w_1(x')<0\ ,\ \ x'\in B\ .$$

Hence

$$\sup_{x\in D} |\bar{w}(x) - w_1(x)| \le \varepsilon$$

and

$$\int_D \bar{w}(x)\varphi(x)dx > 0\ .$$

If $\varphi(x)\equiv 0$, it means that $f_i = 0$, $1\le i\le l$, since eigenfunctions are independent. The function f_i is a minor determinant of $(l-1)$st order. Therefore we can continue the same argument as above for w_2.

Lemma 4

The set

$$\{w\ |D_l(w) = 0\ ,\ w\in L^2(D)^N\}$$

has no interior points.

(Proof) Let

$$w_1 = \sum_{i,j} d_{ij}\varphi_{ij}$$

and expanding (12) with respect to the first row, we have the following form:

$$\sum_{j=1}^{l} f_j d_{kj} = 0 \tag{14}$$

where f_j's are functions of $w_2,\ldots\ldots,w_N,\ \varphi_{k1},\ldots\ldots,\varphi_{kl}$. If some of f_j's are nonzero, then it is clear that the above equation does not contain an open sphere of L^2:

$$\{\bar{d}_{kj}\ |\ \sum_{j=1}^{\infty}\sum_{k=1}^{m_i}(\bar{d}_{kj}-d_{kj})<\varepsilon\}\ .$$

If $f_1=f_2=\cdots=f_l=0$, we can continue the same argument for w_2.

Lemma 5

The set

$$\{w \mid D_i(w) = 0 \ , \ w \in L^2(D)^N\}$$

is closed.

(Proof) From Lemma 2, we have $D_i(w^l) \to D_i(w)$ as $w^l \to w$. Hence if $D_i(w^l) = 0$, $l = 1,2,\ldots\ldots$, then $D_i(w) = 0$.

Corollary 1.

The set

$$\{w \mid D_i(w) = 0 \ , \ w \in C(D)^N\}$$

is closed.

(Proof) Since D is bounded, the relation that $w^l \to w$ in C means $w^l \to w$ in L^2. The result follows immediately from the previous lemma.

From these lemmas we have the following:

Theorem 1

In case of type 1 meansurement, n-mode observability is generic on $\{w \mid w \in L^2(D)^N\}$ or on $\{w \mid w \in C(D)^N\}$ for any $n > 0$.

(Proof) From Lemma 4 and Lemma 5, the set

$$\{w \mid D_i(w) = 0 \ , \ w \in L^2(D)^N\} \ , \quad 1 \leq i \leq n$$

is nowhere dense.

Since the set where n-mode observability does not hold is represented as

$$\bigcup_{i=1}^{n} \{w \mid D_i(w) = 0 \ , \ w \in L^2(D)^N\} \ ,$$

which is obviously nowhere dense. Therefore the genericity in $L^2(D)^N$ is proved. The same argument as above proves the genericity in $C(D)^N$.

Lemma 6

Assume that for $x = (x_1, x_2, \ldots\ldots, x_N)$, $x \in D^N$,

$$det \ V(x_1, x_2, \ldots\ldots, x_N) = 0 \ ,$$

where V is an arbitrary one of $\overline{W}_k(j)$'s:

$$V = \begin{bmatrix} \varphi_{k1}(x_1) & \varphi_{k2}(x_1) & \ldots\ldots & \varphi_{kl}(x_1) \\ \varphi_{k1}(x_2) & \varphi_{k2}(x_2) & \ldots\ldots & \varphi_{kl}(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_{k1}(x_l) & \varphi_{k2}(x_l) & \ldots\ldots & \varphi_{kl}(x_l) \end{bmatrix}$$

Then for any $\varepsilon > 0$, there exists $x' \in D^N$ such that $\|x' - x\| < \varepsilon$ and

$$det\ V(x_l', x_2', \ldots\ldots, x_N') \neq 0$$

(Proof) If we assume the contrary, it follows that for some x and ε_0,

$$det\ V(x_1', x_2', \ldots\ldots, x_N') = 0$$

for all x' such that $\|x' - x\| \leq \varepsilon_0$.

Expanding (16) with respect to $\varphi_{kl}(x_1')$, we have

$$\sum_{i=1}^{l} f_i \varphi_{ki}(x_1') = 0\ , \quad \|x_1' - x_1\| \leq \varepsilon_0$$

Since φ_{ki}'s are analytic [5], the above equation may be extended to D, that is,

$$\sum_{i=1}^{l} f_i \varphi_{ki}(x_1') = 0 \quad x_1' \in D$$

Since eigenfunctions are independent, we obtain $f_1 = f_2 = \ldots\ldots = f_l = 0$. Since f_i's are minor determinants of $(l-1)$st order, we can continue the same argument for x_2, x_3, and so on. Finally we have $\varphi_{kl}(x_j) \equiv 0$ on D for some j, which is a contradiction.

Lemma 7

The set

$$\{x \mid D_l(x) = 0\ ,\ x \in D^N\}$$

is closed, where the topology of D^N is defined by the Euclidean distance of $(\mathbf{R}^n)^N = \mathbf{R}^{nN}$.

(proof) The proof is immediate because $D_l(x)$ is a continuous function of x.

Then we have the following theorem.

Theorem 2

In case of type 2 measurement, n-mode observability is generic in D^N for any $n > 0$.

This theorem can be proved in the same way as in Theorem 1 and we omit the detail.

4. Properties of observable regions

Although n-mode observability is proved to be generic for type 1 and type 2 observations, the same property does not necessarily hold for observability. Therefore we examine whether observable subset is dense in the whole space or not. For this, we need the following definition of a set of the first category.

Definition 3 (cf. Mizohata [3].)

Let E be a metric space. G ($\subset E$) is said to be a set of the first category if and only if it is the union of a countable family of nowhere dense sets, where a nowhere dense set means that its closure has no interior points.

Lemma 8

In case of type 1 measurement, the set K_L (resp. K_C) where observability does not hold in $L^2(D)^N$ (resp. $C(D)^N$) is of the first category.

(proof) Let K_L^n be the set where n-mode observability does not hold. Then

$$K_L = \bigcup_{n=0}^{\infty} K_L^n .$$

(See Theorem 1 in Sakawa [4] or Proposition 2 in this paper.) As is shown in Theorem 1, K_L^n is nowhere dense, hence K_L is of the first category. The same argument is valid for K_C.

Lemma 9

In case of type 2 measurement, the set $K \subset D^N$ where observability does not hold is of the first category.

The proof is the same as that of Lemma 8 and is omitted.

Theorem 3

In case of type 1 measurement, the set in $L^2(D)^N$ or in $C(D)^N$ where observability holds is dense in the respective space.

(proof) The procedure of the proof depends on Mizohata [3].

Let E be a metric space and K denote a set where observability does not hold and is written as

$$K = \bigcup_{n=0}^{\infty} K_n$$

where K_n is nowhere dense (Lemmas 8 and 9).

Assume that the complement $E-K$ is not dense, then there exists a closed sphere $B_0 = \{x \mid \|x-x_0\|_E \le r\}$ such that $B_0 \subset K$ and $B_0 \cap K_0 = \phi$, since K_0 is nowhere dense. Next, we can take another closed sphere $B_1 = \{x \mid \|x-x_1\|_E \le r_1\}$ such that $B_1 \subset B_0$, $B_1 \cap K_1 = \phi$, $r_1 < (1/2)r_0$, since K_1 is nowhere dense. In this way, there exists a sequence of closed sphere

$$B_0 \supset B_1 \supset B_2 \supset \ldots..$$

such that

(a) the diameter r_n converges to zero

(b) $B_n \cap K_n = \phi$.

Then as is shown in Mizohata [3], there exists a $\bar{x}(\in E)$ satisfying $\bar{x} \in \bigcap_{n=0}^{\infty} B_n \subset K$. On the other hand, $\bar{x} \notin K_n$ for any K_n. Hence $\bar{x} \notin K$, which is a contradiction.

In the same manner we can prove:

Theorem 4

In case of type 2 measurement, the set of points in D^N where observability holds is dense in D^N.

Furthermore, in case of type 2 measurement we have the following result by applying standard measure theory.

Theorem 5

In case of type 2 measurement, observability is valid almost everywhere with respect to the Lebesque measure on $\mathbf{R}^{nN}$. In other words the set where observability is not valid is measure zero.

(Proof) It is known that in case of the Lebesque measure the measure of the set K_n is equal to zero. (See [7]) Since K is expressed as a countable union of K_n, the measure of K is also equal to zero.

5. Controllability and n-mode controllability of the dual systems

As dual systems of the system with observation type 1 and type 2, we consider the following.

(type 1')

$$\frac{\partial u}{\partial t} = Au + \sum_{i=1}^{N} w_i(x) z_i(t), \quad (t,x) \in (0,T) \times D$$

$$c_1 u(t,\xi)+(1-c_1)\frac{\partial u}{\partial v}(t,\xi) = 0\ ,\ \ (t,\xi)\in(0,T)\times\partial D$$

$$u(0,x) = u_0(x)$$

where $u_0(x)$ is a *known* initial value, $w_i(x)\in L^2(D)$ or $w_i(x)\in C(D)$, and $Z(t) = (z_1(t),\ldots,z_N(t))\in L^2(0,T)^N$ represents the control.

(type 2')

$$\frac{\partial u}{\partial t} = Au + \sum_{i=1}^{N} \delta(x-x_i)z_i(t)\ ,\ \ (t,x)\in(0,T)\times D$$

$$c_1 u(t,\xi)+(1-c_1)\frac{\partial u}{\partial v}(t,\xi) = 0\ ,\ \ (t,\xi)\in(0,T)\times\partial D$$

$$u(0,x) = u_0(x)$$

where $u_0(x)$ is known and $z(t) = (z_1(t),\ldots,z_N(t))\in L^2(0,T)^N$ is the control.

Below we write the state $u(t,x)$ with control $z\in L^2(0,T)^N$ as $u(t,x;z)$.

Definition 4

The type 1' system or the type 2' system is said to be controllable (in a weak sense) at $t=T$ if the set $\{u(T,x;z)\,|\,z\in L^2(0,T)^N\}$ is dense in $L^2(D)$.

In the following definition it should be noted that the state $u(T,x)$ can be expanded in terms of eigenfunctions:

$$u(T,x) = \sum_{i=1}^{\infty}\sum_{j=1}^{m_i} \xi_{ij}\varphi_{ij}(x)\ .$$

Definition 5

The type 1' system or type 2' system is said to be n-mode controllable if for arbitrary given real numbers η_{ij}, $i=1,2,\ldots,n$; $j=1,2,\ldots,m_i$; it is possible to find a control $\bar{z}$ such that in the expansion of $u(T,x;\bar{z})$:

$$u(T,x;\bar{z}) = \sum_{i=1}^{\infty}\sum_{j=1}^{m_i} \bar{\xi}_{ij}\varphi_{ij}(x)\ ,$$

the relation

$$\eta_{ij} = \bar{\xi}_{ij}\ ,\ \ i=1,2,\ldots,n\ ;\ \ j=1,2,\ldots,m_i$$

is valid.

Then we have

Theorem 6

If we assume $w_{ij}^k = (w_k, \varphi_{ij})$ for type 1' system and $w_{ij}^k = \varphi_{ij}(x_k)$ for type 2' system, then the necessary and sufficient condition for controllability is

$$rank\ W_i = m_i,\ for\ i = 1,2,...$$

where W_i is given by (11).

In the same way, the necessary and sufficient condition for n-mode controllability ($n > 0$) is

$$rank\ W_i = m_i,\ i = 1,2,\ldots,n\ .$$

(Proof) We give the proof for controllability of type 2' system. Define a linear operator $T: L^2(D) \to L^2(0,T)^N$ as $Tu_0 = (y_1(t),\ldots,y_N(t))$, where $y_1(t),\ldots,y_N(t)$ are defined by the original system with type 2 observation. Let us consider a system

$$-\frac{\partial p}{\partial t} = Ap + \sum_{i=1}^{N} \delta(x - x_i) z_i(t),\ in\ (0,T)\times D \tag{15}$$

$$z_i(t) \in L^2(0,T),\ i = 1,2,\ldots,N$$

$$p(T) = 0$$

$$c_1 p + (1 - c_1)\frac{\partial p}{\partial v} = 0\ on\ (0,T)\times D\ .$$

Then it is easy to see that

$$(y,z)_{L^2(0,T)^N} = (u_0, p(0))_{L^2(D)}\ .$$

Therefore

$$T^* z = p(0)$$

Note that the system (15) is equivalent to type 2' system concerning the controllability and the n-mode controllability. In view of the relation $(Ker\ T)^{\perp} = \overline{I_m T^*}$, we have the above condition for the controllability.

As for the n-mode controllability, it is sufficient to consider a linear operator $T\hat{I}: S \to L^2(0,T)^N$ where S is a finite dimensional subspace of $L^2(D)$ and $\hat{I}$ is an imbedding map of S into $L^2(D)$. Note that $\hat{I}^*$ is a projection onto a finite dimensional subspace.

Corollary 2

The n-mode controllability is generic and the controllability is dense for the type 1' system with $w_i(x) \in L^2(D)$ or $w_i(x) \in C(D)$, $i = 1, 2, \ldots, N$.

Corollary 3

The n-mode controllability is generic and the controllability holds almost everywhere on D^N with respect to the Lebesque measure for type 2' system.

6. Conclusions

The objective in this paper is to introduce a degree of easiness in constructing measurements or controllors for parabolic systems. If the observability is dense, we can find sufficiently many points everywhere for measurement. If the observabiliy holds almost everywhere, and if we select randomly the points for measurement, then the probability that we have "unobservable" points is zero. In case of genericity the construction of measurements is still easier.

The above properties are closely related to the zeros of the eigenfunctions φ_{ij}. If $\varphi_{ij} \equiv 0$ for some i, j on a subset $D' \subset D$, then it is easy to see that the observability and the n-mode observability for sufficiently large n are not valid on D'. Therefore generalization of the properties considered here to a broader class of parabolic systems needs examination of zeros of the eigenfunctions.

7. References

[1] R.E. Goodson and R.E. Klein. A Definition and Some Results for Distributed System Observability. IEEE Trans. Auto. Cont., 15, 2, (1970) pp. 165-174.

[2] M.W. Hirsh and S. Smale. Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, New York, (1974).

[3] S. Mizohata. The Theory of Partial Differential Equations. Cambridge Universty Press (1973).

[4] Y. Sakawa. Observability and Related Problems for Partial Differential Equations of Parabolic Type. SIAM J. Cont., 13, 1, (1975) pp. 14-27.

[5] H. Tanabe. On Differentiability and Analiticity of Solutions of Weighted Elliptic Boundary Value Problems. Osaka Math. J., 2, (1965) pp. 163-190.

[6] W.M. Wonham. Linear Multivariable Control: a Geometric Approach, Second Edition. Springer-Verlag, Berlin (1979).

[7] S. Ito. Introduction to the Theory of the Lebesque Integral, Shoka-bo, Tokyo (1963), in Japanese.

SIMPLE ALGORITHMS FOR ADAPTIVE STABILIZATION[†]

A. S. Morse
Department of Electrical Engineering
Yale University
New Haven, Ct. 06520/USA

Introduction

Within the past few years there has been a resurgence of interest in the development of adaptive stabilizers for processes modelled by finite dimensional linear systems. This renewed interest is due in part to a paper by R. D. Nussbaum [1] which proves constructively, for a one-dimensional linear system, that one of the classical process model assumptions of adaptive control is unnecessary. Subsequent work by others [2-7] shows that these assumptions can be relaxed very much further while [8] addresses the necessity of the assumptions themselves.

The purpose of this paper is to describe several different algorithms for adaptive stabilization. Some are new, while others have been discussed previously in [2-4]. In contrast with the general adaptive stabilizers of [6,7],each algorithm considered here is very simple in structure and easy to analyze. All are "minimal compensator based", {cf. [9]} and consequently of the high-gain feedback type. None uses a "probing signal" or an "augmented error" {cf. [3]} and one is applicable to process models of relative degree 3 or less.

In the discussion which follows use is made of several concepts and constructions which differ sharply from those of classical adaptive control {e.g. [10]}. In §1 a Nussbaum Gain is employed as a component of an adaptive stabilizer for a one-dimensional system - and closed-loop stability is proved using a nonclassical "indicator function" [1]. In §2, a nonclassical parameterization is used to prove that the algorithm of §1 also stabilizes relative degree one minimum phase systems [2]. A parametrically dependent indicator function is used in §3 to prove that the algorithm of §1 also stabilizes relative degree two minimum phase systems with "positive damping". Finally in §4 it is shown that any minimum phase system of relative degree two or less can be stabilized by a one-dimensional adaptive stabilizer which is nonlinearly dependent on a single tunable parameter.

1. One-Dimensional Systems

Let us begin by considering the problem of adaptively stabilizing the one-dimensional linear system

$$\dot{y} = ay + gu \tag{1}$$

with unknown but constant parameters a and g, assuming $g \neq 0$. Our objective is to

[†] The work of the author was supported in part by the U. S. Air Force Office of Scientific Research under Grant No. 84-0242.

construct an m-dimensional, nonlinear dynamical system of the form

$$\left.\begin{aligned} \dot{x} &= f(x,y) \\ u &= h(x,y) \end{aligned}\right\} \qquad (2)$$

which stabilizes (1) in the sense that, for each initial state (y_0,x_0), the solution $(y(t),x(t))$ to the closed-loop dynamical system (1), (2) exists and is bounded on $[0,\infty)$ and $y(t) \to 0$ as $t \to \infty$. Here $f\colon \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ and $h\colon \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ are "smooth" functions not depending on a or g.

If σ_g = sign(g) is known, stabilization can be achieved with the classical adaptive controller

$$u = -\sigma_g ky \qquad (3a)$$

$$\dot{k} = y^2 \qquad (3b)$$

It is easy to prove that the resulting closed-loop system

$$\dot{y} = (a-|g|k)y \qquad (4a)$$

$$\dot{k} = y^2 \qquad (4b)$$

is stable. For this, first choose a constant k_0 so that $a - |g|k_0 < 0$. Next, evaluate the derivative of the "indicator function"

$$V = y^2/2 + |g|(k-k_0)^2/2 \qquad (5)$$

along solutions to (4); i.e. $\dot{V} = (a-|g|k_0)y^2$. Since $\dot{V} \leq 0$, V is monotone nonincreasing. Clearly $0 \leq V(t) \leq V(0)$ so y and k are in L^∞, the space of bounded functions on $(0,\infty)$. Next observe that (4a) implies $\dot{y} \in L^\infty$, whereas (4b) implies $y \in L^2$, the space of square integrable functions on $(0,\infty)$; it follows that $y(t) \to 0$ as $t \to \infty$. Thus controller (3) stabilizes (1).

The preceding analysis is classical. An indicator function (5),(actually a bonafide Lyapunov function in this particular case) which is quadratic in y and "parameter error" $k - k_0$,is used to prove stability. There is another way to prove stability which has the virtue of being applicable in a variety of more general situations. This alternative method uses an indicator function which is quadratic in just y; i.e.

$$V = y^2/2 \qquad (6)$$

In this case evaluation of $\dot{V}$ along solutions to (4) gives $\dot{V} = (a-|g|k)y^2$; hence from (4b),

$$\dot{V} = (a-|g|k)\dot{k}$$

This equation can be integrated to yield

$$V(t) = ak(t) - |g|k^2(t)/2 + C \qquad (7)$$

where C is a constant. Examination of (7) reveals that $k \in L^\infty$ - for if this were not so, then for $|k|$ sufficiently large, V would become negative which by (6) is impossible. Clearly $V \varepsilon \mathcal{L}^\infty$, so from (6) $y \varepsilon \mathcal{L}^\infty$ as well. With $(y,k) \varepsilon \mathcal{L}^\infty$ now established, (4) can be used just as in the classical analysis discussed earlier to prove that $y \to 0$. We shall use this non-classical method of proof again in a moment.

Consider again the adaptive stabilization of (1), but now with σ_g unknown. To deal with this situation, we replace control law (3a) with

$$u = N(k)ky \tag{3a'}$$

where $N(\cdot)$ is a Nussbaum Gain; i.e., any integrable function satisfying

$$\left.\begin{aligned} \sup_{x>0} \frac{1}{x}\int_0^x N(\mu)\mu d\mu &= \infty \\ \inf_{x>0} \frac{1}{x}\int_0^x N(\mu)\mu d\mu &= -\infty, \end{aligned}\right\} \tag{8}$$

e.g., $N(\mu) = \mu \cos(\mu)$. To prove that the resulting closed-loop system

$$\dot{y} = ay + gN(k)ky \tag{9a}$$

$$\dot{k} = y^2 \tag{9b}$$

is stable, we proceed just as before by evaluating the rate of change of the indicator function $V = y^2/2$ along solutions to (9). Thus $\dot{V} = (a + gN(k)k)y^2$; hence from (9b), $\dot{V} = (a+gN(k)k)\dot{k}$. Therefore by integrating

$$V(t) = ak(t) + g\int_0^{k(t)} N(\mu)\mu d\mu + C \tag{10}$$

The definition of $N(\cdot)$ in (8) clearly implies that for some number $k^* \geq k(0)$,

$$ak^* + g\int_0^{k^*} N(\mu)\mu d\mu + C < 0$$

Since by definition $V \geq 0$, $k(t)$ cannot attain this value. It follows that $k(0) \leq k(t) < k^*$ or that $k \varepsilon \mathcal{L}^\infty$. The definition of V together with (10) thus imply that $y \varepsilon \mathcal{L}^\infty$ as well. With $(y,k) \varepsilon \mathcal{L}^\infty$, (9) can now be used, just as in the classical proof discussed earlier, to show that $y \to 0$.

This proves that controller (3a'), (3b) adaptively stabilizes (1). The concept of a Nussbaum Gain and the nonclassical stability analysis we've just used, are based on ideas introduced by Nussbaum in [1].

2. Relative Degree One Systems

We now consider the problem of adaptively stabilizing a process with scalar input u and scalar output y, which can be modelled by a linear system Σ with transfer function

$$g\,\frac{\alpha(s)}{\beta(s)}$$

where g is a nonzero constant (the high frequency gain), $\alpha(s)$ and $\beta(s)$ are monic and coprime polynomials, and $\alpha(s)$ is stable (i.e., Σ is minimum phase). A useful state space realization of Σ is provided by the following Lemma:

Lemma 1: *Write* γ *and* ρ *for the unique quotient and remainder of* β *divided by* α; i.e. $\beta = \alpha\gamma + \rho$, degree (ρ) < degree (α). *Then* Σ *admits a state-space model of the form*

$$y = c_1x_1$$

$$\dot{x}_1 = A_1x_1 + b_1(gu + L(y))$$

$$L(y) = c_2x_2$$

$$\dot{x}_2 = A_2x_2 + b_2y$$

where $\Sigma_1 = (c_1,A_1,b_1)$ *and* $\Sigma_2 = (c_2,A_2,b_2)$ *are canonical realizations of* $1/\gamma$ *and* $-\rho/\alpha$ *respectively*.

For a simple proof of this lemma, see [4].

Σ_1 and Σ_2 are called respectively, the *quotient* and *remainder* subsystems of Σ. As a consequence of the minimum phase assumption, Σ_2 is necessarily stable. In addition, note that the dimension of Σ_1 equals $n^* =$ degree (β) - degree (α), the *relative degree* of Σ.

Assume $n^* = 1$. In view of Lemma 1, Σ can be described in the state space by the equations

$$\dot{y} = ay + gu + L(y) \tag{11a}$$

$$L(y) = c_2x_2 \tag{11b}$$

$$\dot{x}_2 = A_2x_2 + b_2y \tag{11c}$$

where A_2 is a stability matrix. We wish to prove that control (3a'), (3b), previously shown to stabilize (1), also stabilizes (11). In this case, the closed-loop system is described by

$$\dot{y} = ay + gN(k)ky + L(y) \tag{12a}$$

$$\dot{k} = y^2 \tag{12b}$$

together with (11b) and (11c).

To prove stability, we shall proceed along exactly the same lines as before. The first step is to evaluate the rate of change of the indicator function $V = y^2/2$ along solutions to (11), (12). Thus

$$\dot{V} = ay^2 + gN(k)ky^2 + yL(y)$$

As before, we can substitute $\dot{k}$ for y^2 and integrate. The result is

$$V(t) = ak(t) + g\int_0^{k(t)} N(\mu)\mu d\mu + C + \int_0^t y(\tau)L_\tau(y)d\tau \tag{13}$$

Except for the term $\int yL_\tau(y)d\tau$, this expression for V is the same as (10). To deal with this term, one additional technical result is needed [2].

Lemma 2: *Let* $\dot{x} = Ax + B\zeta$, $\omega = Cx + D\zeta$ *be a stable linear system. There exist positive constants* C_1 *and* C_2, *depending only on* (A,B,C,D) *such that for each initial state* $x(0) = x_0$ *and each piecewise-continuous input* $\zeta(\cdot)$,

$$\int_0^t \omega(\tau)\zeta(\tau)d\tau \le C_1 \|x_0\|^2 + C_2 \int_0^t \|\zeta(\tau)\|^2 d\tau$$

In view of Lemma 2, we see that for some numbers $\tilde{C}_1$ and C_2 not depending on t,

$$\int_0^t y(\tau)L_\tau(y)d\tau \le \tilde{C}_1 + C_2 \int_0^t y^2(\tau)d\tau$$

Replacing y^2 by $\dot{k}$ and substituting into (13), there results the inequality

$$V(t) \le \bar{a}k(t) + g \int_0^{k(t)} N(\mu)\mu d\mu + \bar{C}$$

where $\bar{a} = a + C_2$ and $\bar{C} = C + \tilde{C}_1 - C_2k(0)$. Observe that this expression for V is of exactly the same form as the expression for V in (10). Hence the same reasoning as before can be used to conclude that $(y,k) \in L^\infty$ and that $y(t) \to 0$, as $t \to \infty$; and with A_2 a stability matrix, it follows from (11c) that $x_2(t) \to 0$ as well.

To summarize, we have shown that controller (3a'), (3b) used to stabilize (1), also stabilizes any relative degree one minimum phase system Σ in the sense that the state (y,x_2,k) of the closed-loop system (11), (12) is bounded on $(0,\infty)$ and Σ's state $(y,x_2) \to 0$ as $t \to \infty$. The method of proof is basically the same as that used in §1. The new ideas needed for the generalization to relative degree one systems - the parameterization of Σ in (11) and the inequality of Lemma 2 - are due to Willems and Byrnes [2].

3. Relative Degree Two Systems with Positive Damping

It is well-known from root-locus considerations that minimum phase relative degree one systems can always be stabilized (in a nonadaptive context) with high-gain control laws of the form u = ky provided gain k is of the appropriate sign and sufficiently large in magnitude. Root locus arguments can also be used to identify those relative degree two, minimum phase systems which can be similarly stabilized. In particular, if $\gamma(s) = s^2 + as + b$ is the denominator of the transfer function of the quotient system of Σ {cf. §2}, then Σ can be stabilized with a high-gain feedback u = ky provided Σ's "damping coefficient" $a > 0$. It is natural to expect that controls such as (3a), (3b) or (3a'), (3b) can adaptively stabilize such systems. At present we are unable to prove that (3a'), (3b) stabilizes; however, for the case when σ_g is known, controller (3a), (3b) can in fact adaptively stabilize. Our objective is to show that this is so. Our ideas here have been heavily influenced by M. Corless who, in an informal communication, proved that controller (3) stabilizes any system with transfer function $g/(s^2+as+b)$ provided $a > 0$.

To proceed, assume Σ is minimum phcse, of relative degree two, with damping $a > 0$. Since Σ's quotient subsystem has transfer function $1/(s^2+as+b)$, by Lemma 1 Σ can be modelled in the state space by the equations

$$\dot{y} = -\frac{a}{2}y + x_1 \qquad (14a)$$

$$\dot{x}_1 = -\frac{a}{2}x_1 + gu + L(y) \qquad (14b)$$

$$L(y) = C_2x_2 + (a/2 - b)y \tag{14c}$$

$$\dot{x}_2 = A_2x_2 + b_2y \tag{14d}$$

where A_2 is a stability matrix. Application of control (3) results in the closed-loop system (14a), (14c), (14d) together with

$$\dot{x}_1 = -\frac{a}{2}x_1 - |g|ky + L(y) \tag{15a}$$

$$\dot{k} = y^2 \tag{15b}$$

Note from (15b) that either $k \in L^\infty$ or k grows positively without bound. Our first objective is to show that the latter is impossible. For this, assume the contrary. Therefore, for some time t_0, $k(t) > 0$ for $t \geq t_0$. For such values of t, computation of the rate of change of the indicator function

$$V = |g|y^2/2 + x_1^2/2k \tag{16}$$

along solutions to (14)-(15), yields

$$\dot{V} = -\frac{|g|a}{2}y^2 - \omega x_1^2 - \frac{a}{2}\frac{x_1^2}{k} + \frac{x_1}{k}L(y) \tag{17}$$

where $\omega = \dot{k}/2k^2 \geq 0$. Let d be any positive number, and note that $\frac{x_1}{k}L(y) \leq \frac{d}{2}\frac{x_1^2}{k^2} + \frac{1}{2d}L^2(y)$. Application of this to (17) and then integrating yields

$$V \leq -\int_{t_0}^{t}[(\frac{|g|a}{2}y^2(\tau) - \frac{1}{2d}L^2_\tau(y) + (\omega(\tau) + \frac{a}{2k(\tau)} - \frac{d}{2k^2(\tau)})x_1^2(\tau)]d\tau + C \tag{18}$$

At this point we need the following technical result [10]:

Lemma 3: Let $\dot{x} = Ax + B\zeta$, $w = Cx + D\zeta$ be a stable linear system. There exist positive constants C_1 and C_2 depending only on (A,B,C,D) such that for each initial state $x(t_0) = x_0$ and each piecewise-continuous input $\zeta(\cdot)$

$$\int_{t_0}^{t} w^2(\tau)d\tau \leq C_1\|x_0\|^2 + C_2\int_{t_0}^{t}\|\zeta(\tau)\|^2d\tau$$

In view of Lemma 3, there are constants $\tilde{C}_1$ and C_2 not depending on t such that

$$\int_{t_0}^{t} L^2_\tau(y)d\tau \leq \tilde{C}_1 + C_2\int_{t_0}^{t} y^2(\tau)d\tau$$

Hence with $d = 2C_2/|g|a$, the preceding can be used together with (18) and (15b) to obtain

$$V(t) \leq -\frac{|g|a}{4}k(t) - \int_{t_0}^{t}(\omega(\tau) + \frac{a}{2k(\tau)} - \frac{d}{2k^2(\tau)})x_1^2(\tau)d\tau + \bar{C} \tag{19}$$

where $\bar{C} = C + \tilde{C}_1/2d + \frac{|g|a}{4}k(0)$. Examination of (19) clearly reveals that if k(t) were to grow positively without bound, then V(t) would eventually become negative which is impossible. Therefore $k \in L^\infty$.

Having achieved our first objective, we now must show that $k \in L^\infty$ implies that $(y,x_1,x_2) \to 0$ as $t \to \infty$. For this we make use of the following special result.

Lemma 4: Let (C(t),A(t)) be continuous matrices with $\lim_{t\to\infty}(C(t),A(t)) = (\bar{C},\bar{A})$ existing and observable. If for some initial state x_0, and some L^2-input b(t), to the linear system $\dot{x} = A(t)x + b(t)$, the output $y(t) = C(t)x(t)$ satisfies $y \in \mathcal{L}^2$, then $x(t) \to 0$ as $t \to \infty$.

Proof: Since $(\bar{C},\bar{A})$ is observable there exists a matrix K such that $\bar{A} + K\bar{C}$ is stable. Then $\dot{x} = (\bar{A} + K\bar{C})x + (\tilde{A} + K\tilde{C})x - Ky + d$ where $\tilde{A} = A - \bar{A}$ and $\tilde{C} = C - \bar{C}$. Since $-Ky + d \in \mathcal{L}^2$ and $\tilde{A}(t) + K\tilde{C}(t) \to 0$ as $t \to \infty$, it follows from Theorem 2, Chapter 2 of [11], that the lemma is true.

Lemma 4 can be used as follows. First observe that (14a) and (15a) can be written as $y = Cx$, $\dot{x} = A_k x + d$ where $x = \begin{pmatrix} y \\ x_1 \end{pmatrix}$ $C = [1,0]$, $A_k = \begin{bmatrix} -a/2 & 1 \\ -|g|k & -a/2 \end{bmatrix}$ and $d = \begin{bmatrix} 0 \\ L(y) \end{bmatrix}$. Since $k \in \mathcal{L}^\infty$, (15b) implies that k approaches a finite limit $\bar{k}$ and that $y \in \mathcal{L}^2$. Clearly $d \in \mathcal{L}^2$ and $(C,A_{\bar{k}})$ is observable. Therefore by Lemma 4, $(y,x_1) \to 0$; since A_2 is stable it follows from (14d) that $x_2 \to 0$ as well.

The preceding analysis proves that controller (3) can adaptively stabilize any relative degree two minimum phase system with positive damping, provided σ_g is known. The novel feature of the analysis is the use of an indicator function (16) which depends on a control parameter k.

4. Relative Degree Two Systems

We now turn to the problem of developing adaptive stabilizers for the class of all relative degree two, minimum phase systems. Since there are systems in this class which cannot be stabilized (in a nonadaptive context) with the simple high-gain feedback law $u = ky$, something more general than the controllersof §1 will have to be used if adaptive stability is to be achieved for every possible system in the class. One possible control structure with this potential is described by the equations

$$\left.\begin{aligned} u &= -k_2\theta - k_1k_2y \\ \dot{\theta} + \lambda\theta &= u \end{aligned}\right\} \qquad (20)$$

where λ is a positive constant. For if this controller is applied to a relative degree two, minimum phase system Σ with transfer function $g\alpha/\beta$, then for sufficiently large values of parameter constants k_1 and k_2 stability will result. This can easily be proved by examining the closed-loop system characteristic polynomial

$$\pi(s) = (s+\lambda)\beta(s) + k_2(\beta(s) + k_1 g\alpha(s)(s+\lambda))$$

Since $\alpha(s)(s+\lambda)/\beta(s)$ is a minimum phase, relative degree one transfer function, for $k_1 g$ sufficiently large, $\beta(s) + k_1 g\alpha(s)(s+\lambda)$ will be stable. With k_1 fixed at such a value, $(\beta(s) + k_1 g\alpha(s))/(s+\lambda)\beta(s)$ is also a minimum phase, relative degree one transfer function so for k_2 sufficiently large $\pi(s)$ will be a stable polynomial.

An adaptive version of (20) has been shown to be capable of stabilizing any minimum phase system with relative degree not exceeding two [4]. The tuning formulas for this controller are

$$k_2 = -N((k_\theta^2+k_y^2)^{\frac{1}{2}})k_\theta$$

$$k_1k_2 = -N((k_\theta^2+k_y^2)^{\frac{1}{2}})k_y$$

$$k_\theta = \theta y + z_\theta$$

$$k_y = \tfrac{1}{2}\, y^2 + z_y$$

$$\dot{z}_\theta = (\lambda+\lambda_1)\theta y - uy$$

$$\dot{z}_y = \lambda_1 y^2$$

where λ_1 is a positive constant, and $N(\cdot)$ is a Nussbaum Gain.

In the sequel we consider an alternative controller, depending on only one parameter k. In particular we assume σ_g is known, set $k_1 = \sigma_g k$, $k_2 = k$ and adjust k according to the rule $\dot{k} = y^2$. The resulting controller {see also [12]} is thus described by the equations

$$\begin{aligned} u &= -k\theta - \sigma_g k^2 y \\ \dot{\theta} + \lambda\theta &= u \qquad\qquad (21) \\ \dot{k} &= y^2 \end{aligned}$$

Our objective is to show that this controller can stabilize any minimum phase system with relative degree not exceeding two. For this, assume that for $i = 1,2$, Σ_i is a minimum phase system of relative degree i. By Lemma 1, Σ_1 admits a state space model of the form

$$\dot{y} = -\lambda y + gu + L(y)$$

$$L(y) = (\lambda-a)y + c_2x_2$$

$$\dot{x}_2 = A_2x_2 + b_2y$$

where A_2 is a stable matrix and $1/(s+a)$ is the transfer function of Σ_1's quotient system. A direct calculation shows that

$$y = g\theta + \bar{L}(y)$$

where

$$\bar{L}(y) = z$$

$$\dot{z} = -\lambda z + L(y)$$

provided $z(0) = y(0) - g\theta(0)$. Thus using state (y,x_2,k,z) rather than (y,x_2,k,θ), the closed-loop system which results when controller (21) is applied to Σ_1 is

$$\dot{y} = -\lambda y - |g|k^2y - ky + L(y) + k\bar{L}(y) \qquad (22a)$$

$$\dot{k} = y^2 \qquad (22b)$$

$$L(y) = (\lambda-a)y + c_2x_2 \qquad (22c)$$

$$\dot{x}_2 = A_2x_2 + b_2y \qquad (22d)$$

$$\bar{L}(y) = z \tag{22e}$$

$$\dot{z} = -\lambda z + L(y) \tag{22f}$$

A similar system of equations can be derived for the case when the controlled system is Σ_2. By Lemma 1, Σ_2 admits a state space model of the form

$$\dot{y} = -\lambda y + x_1$$

$$\dot{x}_1 = (\lambda-a)x_1 + gu + L(y)$$

$$L(y) = (\lambda(a-\lambda)-b)y + c_2x_2$$

$$\dot{x}_2 = A_2x_2 + b_2y$$

where $1/(s^2+as+b)$ is the transfer function of Σ_2's quotient system and A_2 is a stability matrix. By direct calculation, it is easy to verify that

$$x_1 = g\theta + \bar{L}(y)$$

where

$$\bar{L}(y) = (2\lambda-a)y + z$$

$$\dot{z} = -\lambda z + L(y)$$

provided $z(0) = x_1(0) - g\theta(0) - (2\lambda-a)y(0)$. Thus using state (y,x_1,x_2,k,z) rather than (y,x_1,x_2,k,θ), the closed-loop system which results when controller (21) is applied to Σ_2 is

$$\dot{y} = -\lambda y + x_1 \tag{23a}$$

$$\dot{x}_1 = (\lambda-a-k)x_1 - k^2|g|y + L(y) + k\bar{L}(y) \tag{23b}$$

$$\dot{k} = y^2 \tag{23c}$$

$$L(y) = (\lambda(a-\lambda)-b)y + c_2x_2 \tag{23d}$$

$$\dot{x}_2 = A_2x_2 + b_2y \tag{23e}$$

$$\bar{L}(y) = (2\lambda-a)y + z \tag{23f}$$

$$\dot{z} = -\lambda z + L(y) \tag{23g}$$

Our objective now is to show that systems (22) and (23) are each stable, in the sense that $k \in L^\infty$ and all other state variables go to zero as $t \to \infty$. Note that for either system either $k \in L^\infty$ or k grows without bound. Suppose the former is true. Then $y \in L^2$ and k approaches a finite limit $\bar{k}$. Thus $L(y)$ and $\bar{L}(y)$ are in L^2 and Lemma 4 can be used to show that y and x_1 go to zero. From this it then follows that (x_2,z) goes to zero as well. To prove stability, it is therefore enough to show that $k \in L^\infty$.

Consider first (22). Evaluation of the rate of change of the indicator function $V = y^2/2$ gives

$$\dot{V} = -(\lambda + |g|k^2 + k)y^2 + yL(y) + yk\bar{L}(y)$$

$$\leq -(\lambda + \frac{|g|}{2}k^2 + k)y^2 + yL(y) + \frac{1}{2|g|}\bar{L}^2(y)$$

Thus by integrating, and then using Lemmas 2 and 3,

$$V(t) \leq -\int_0^t (\lambda - C_1 + \frac{|g|}{2}k^2(\tau) + k(\tau))y^2(\tau)d\tau + C_2$$

where C_1 and C_2 are positive constants. Replacing y^2 by $\dot{k}$ and carrying out the integration there results

$$V(t) \leq -(\lambda - C_1)k(t) - \frac{|g|}{4}k^3(t) - \frac{k^2(t)}{2} + C_3$$

Quite clearly, if k were to grow without bound, then V would become negative which is impossible. Therefore $k \in L^\infty$, which proves that system (22) is stable.

We now turn to system (23). Assume k(t) grows without bound. Then for some time $t_0 \geq 0$, $k(t) > 0$ for $t \geq t_0$. For $t \geq t_0$, the rate of change of the indicator function $V = |g|y^2/2 + x_1^2/2k^2$ along solutions to (23) is

$$\dot{V} = -\lambda|g|y^2 - (\omega + \frac{k+a-\lambda}{k^2})x_1^2 + \frac{x_1L(y)}{k^2} + \frac{x_1\bar{L}(y)}{k}$$

$$\leq -\lambda|g|y^2 - (\omega + \frac{1}{k} + \frac{(a-\lambda+ 1/2\bar{d})}{k^2} + \frac{(1/2d)}{k^4})x_1^2 + \frac{d}{2}L^2(y) + \frac{\bar{d}}{2}\bar{L}^2(y)$$

where $\omega = \dot{k}/k^3 \geq 0$ and d and $\bar{d}$ are any positive constants. By integrating this inequality, using Lemma 3 for $L^2(y)$ and $\bar{L}^2(y)$, then appropriately selecting d and $\bar{d}$, and replacing y^2 by $\dot{k}$, we obtain

$$V(t) \leq -\frac{\lambda|g|}{2}k(t) - \int_{t_0}^t (\omega(\tau) + \frac{1}{k(\tau)} + \frac{(a-\lambda- 1/2\bar{d})}{k^2(\tau)} - \frac{(1/2d)}{k^4(\tau)})x_1^2(\tau)d\tau + C$$

where C is a constant. Examination of this expression shows that if k were to grow without bound, then V would become negative which is impossible. Therefore $k \in L^\infty$ and system (23) is stable.

Concluding Remarks

In this paper we have analyzed three simple algorithms for adaptive stabilization. The most general of these-the one-dimensional system $u = -k\theta - \sigma_g k^2 y$, $\dot{\theta} + \lambda\theta = u$ with a single parameter k adjusted by the rule $\dot{k} = y^2$ - is capable of stabilizing any minimum phase system of relative-degree two or less, provided σ_g is known. For the case when σ_g is unknown it is likely that adaptive stabilization can be achieved using the above controller with σ_g replaced by a Nussbaum Gain N(k) - but this remains to be seen.

It is natural to expect the ideas in this paper to generalize to systems of relative degree greater than two. It can be shown that the one-parameter, two-dimensional controller $u = -k^3\theta_1 - k^2\theta_2 - \sigma_g k^4 y$, $\dot{\theta}_1 + \lambda_1\theta_1 = \theta_2$, $\dot{\theta}_2 + \lambda_2\theta_2 = u$, with λ_1 and λ_2 positive constants, together with adjustment rule $\dot{k} = y^2$, stabilizes any relative degree three or less minimum phase system with σ_g known. A proof of this will appear elsewhere.

What's especially interesting about these results is that they strongly suggest that at least three concepts - namely error models, error augmentation and positive realness - are not as crucial to adaptive stabilization theory as they were once thought to be. It seems that this paper raises more questions than it answers and that there is a great deal of work to be done.

References

[1] R. D. Nussbaum, "Some Remarks on a Conjecture in Parameter Adaptive Control," Systems & Control Letters, 3, November 1983, pp. 243-246.

[2] J. C. Willems and C. I. Byrnes, "Global Adaptive Stabilization in the Absence of Information as the Sign of the High Frequency Gain," Proc. Sixth International Conference on Analysis and Optimization of Systems, Springer Lecture Notes in Control and Information Sciences, 62, June 1984, pp. 49-57.

[3] D. R. Mudgett and A. S. Morse, "Adaptive Stabilization of Linear Systems with Unknown High-Frequency Gains," IEEE Transactions on Automatic Control, AC-30, No. 6, June 1985, pp. 549-554.

[4] A. S. Morse, "A Three-Dimensional Universal Controller for the Adaptive Stabilization of Any Strictly Proper Minimum-Phase System with Relative Degree Not Exceeding Two," IEEE Transactions on Automatic Control, AC-30, No. 12, December 1985, pp. 1188-1191.

[5] A. S. Morse, "A Model Reference Controller for the Adaptive Stabilization of Any Strictly Proper, Minimum Phase, Linear System with Relative Degree Not Exceeding Two," Proc. 1985 MTNS Conf., Stockholm, Sweden, June 1985.

[6] B. Martensson, "The Order of any Stabilizing Regulator is Sufficient Apriori Information for Adaptive Stabilization," Systems and Control Letters, 1985.

[7] M. Fu and B. R. Barmish, "Adaptive Stabilization of Linear Systems Via Switching Control," IEEE Trans. Auto. Control, AC-31, December 1986.

[8] C. I. Byrnes, O. Helmke and A. S. Morse, "Necessary Conditions for Adaptive Control," Proc. 1985 MTNS Conf., Stockholm, June 1985.

[9] A. S. Morse, "New Directions in Parameter Control," Proc. 23rd Conf. on Decision and Control, Las Vegas, Nv., December 1984, pp. 1566-1568.

[10] A. S. Morse, "Global Stability of Parameter-Adaptive Control Systems," IEEE Trans. Automat. Contr., AC-25, June 1980, pp. 433-439.

[11] R. Bellman, Stability Theory of Differential Equations, Dover Publications, 1953.

[12] C. I. Byrnes and A. Isidori, "Asymptotic Expansions, Root-Loci and the Global Stability of Nonlinear Feedback Systems," Proc. of the Conf. on the Algebraic and Geometric Methods in Non-Linear Control Theory, C.N.R.S., Paris, June 1985.

General structure of hierarchy control and management with both of objective and subjective approach

Toichiro Nakagawa

Consulting director of Chichibu Cement Co., Ltd.,

President of System Sogo Kaihatu Co., Ltd. Tokyo, Japan.

1. Introduction

Control problems of the production process generally signifies the field of stabilizing control of process variables in most cases. However, optimum production level and setting level of the corresponding process variables have not been regulated at all, and have been selected empirically on the basies of static heat and mass balance without colored noise in the process, and the basic operation is generally subjective rather than objective. The behavior of production is dynamic as far as it is in operation field, and it is improper to obtain optimum production level only from static specifications of equipments. In other words, static control is only passive management after event is occured, and does not solve the problem of how and what to manage.

In this report, a hierarchy system of stabilizing control and control management including determination of optimum production level, pursuit control to this level, is described, and the system of how and what to do is presented briefly with examples in practical field. Further more, the cooperation with both objective and subjective (Artificial Intelligence) approach is also referred. Fig-1 shows the system structure and Fig-2 shows diagram of the relation-ships with each field.

2. Production Management and Control System

Control system in a broad sense including production management has generally hierarchy structure. Conventional PID control, which is widely used, and optimum regulator problem by modern control theory are controls of the local stabilization So far, set points of each variable are given tentatively case by case. Selection of each set point is made in a trial-and-error manner by considering the original unit cost, quality, ease of operation etc.

Therefore, even if stabilization control is realized, the effect of the stabilization control from the aspect of production control, or the reduction of costs or guarantee of high quality, has been indefinite. Even though various optimization techniques have been developed in the field of operations research, they have not been applied to the actual optimization of industrial process. This is considered to be originated in the fact that it has been difficult to construct practical production levels of process models without dynamical consideration of process.

If optimization model of production level is obtained, as refered later, a set of set points is determined by the optimization technique, and stabilizing control operates to stabilize around this optimum set points. Moreover, if a set of optimum set points changes because of the shift of process state, pursuit control operates to shift the process forward to each of the new optimum set points. Fig.1 shows the flow-diagram of analysis and control procedure, and the rough draft of generalized system structure is shown in Fig.2 and presents the relations among the fields.

The systems are composed of surface-level models and deep models of reasoning using of a multi-level approach. The surface knowledge is described by the production rule type and the deep model is mathematical and implemented as a complex software tools, such as process and human simulator.

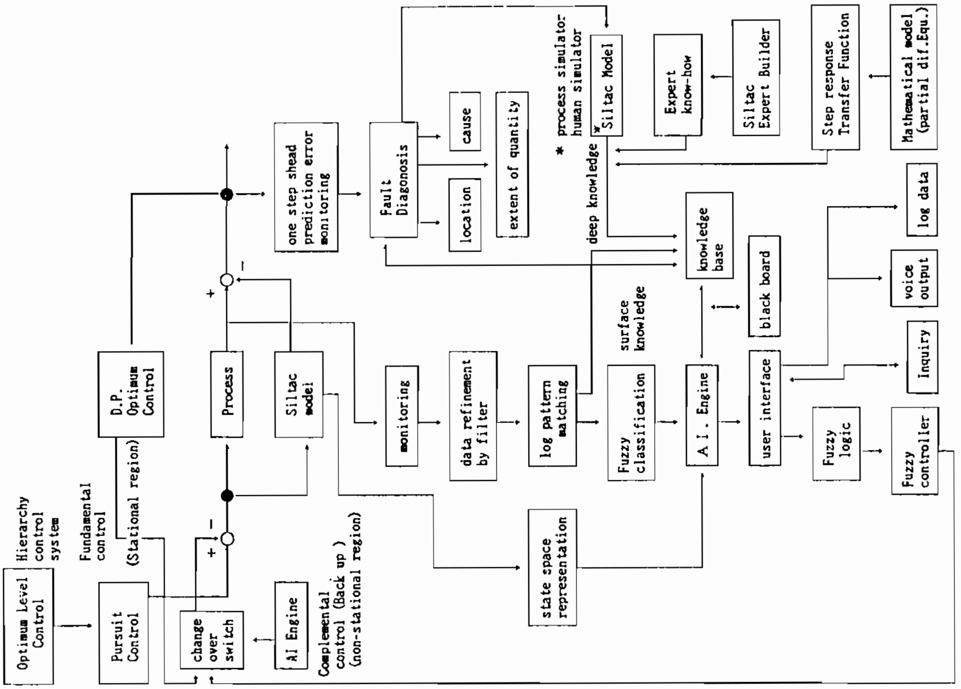


Fig. 2 diagram of the relation-ships with each field

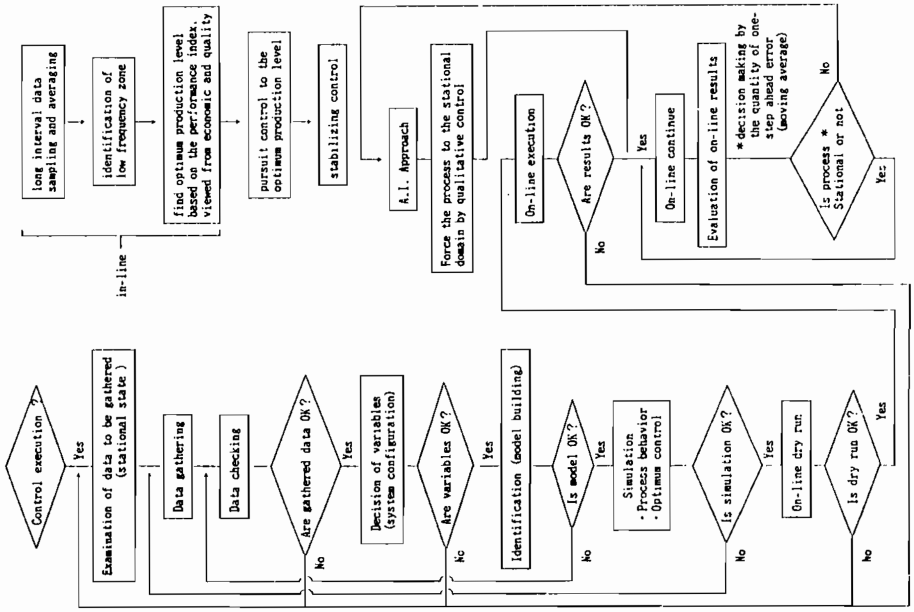

long interval data sampling and averaging
identification of low frequency zone
find optimum production level based on the performance index, viewed from economic and quality
pursuit control to the optimum production level
stabilizing control
in-line
A.I. Approach
Force the process to the stational domain by qualitative control
On-line execution
Are results OK?
On-line continue
Evaluation of on-line results
* decision making by the quantity of one-step ahead error (moving average)
Is process * Stational or not
Control execution ?
Examination of data to be gathered (stational state)
Data gathering
Data checking
Are gathered data OK?
Decision of variables (system configuration)
Are variables OK?
Identification (model building)
Is model OK?
Simulation
· Process behavior
· Optimum control
Is simulation OK?
On-line dry run
Is dry run OK?
Yes
No

Fig. 1 system structure

3. Stabilizing control of the kiln process

Since a very good exposition of cement manufacturing process is already available, only a brief description of the kiln process, which is depicted in Fig.3, will be given.

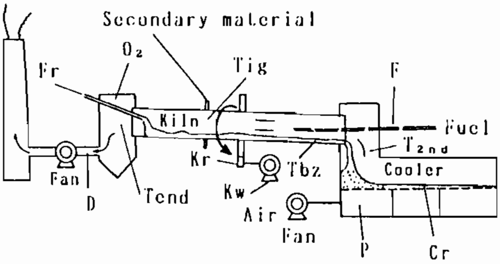


Fig. 3 Rotary kiln and clinker cooler system

The raw materials such as lime stone, clay and pyrite cinder which are ground in mill are fed into the higher end of the kiln and the raw material is moved downwards by the rotation of the kiln. During the travel through the kiln the raw material is first dried, calcined and further heated to reaction temperature to form the clinker after several phases of physical-chemical reactions. The clinker is then quenched and cooled in the clinker cooler. The necessary heat for the reactions within the kiln is supplied by burning fuel at the lower end of the kiln. Rotary kiln with suspension preheater (so called SF Kiln) is dry system and thermal effectivity of SF kiln is strongly improved, comparing conventional long kiln.

Table 1 shows parametes of auto regressive model through identification [1][2][3][8][9][10][11] and optimum control gain fo wet rotary kiln process.

Table 2 shows one example of the model parameters fordry SF rotary kiln process. As is shown in Table 2, these auto regressive equations represent the process behavior model and operator model at the same time. Then operator can realize and check his own operation by himself.

This behavior with feed back of operation's action is reproduced and can be easily simulated on the CRT and printer. Then we can play the simulation-game from the operation board just like the atmosphere of the real process. Of course, optimum

```
                    ***  KILN 1 A,B AND G CHECK  ***

     LL=  4  NR=  3  HT=  6

              A(M,I,J)       FILE NO. = 1001                          B(M,I,J)       FILE NO. = 1011
M=  1          W                P                T                         F               K                D                C
  W    0.824270E 00 -0.219830E-02 -0.457220E-01       W    0.103830E-01  0.457660E-01  0.129690E 00 -0.261820E-01
  P   -0.998810E-01  0.140530E 01 -0.242530E 00       P    0.147260E 01  0.763540E 00 -0.241080E 00 -0.100960E 01
  T    0.105500E-01  0.185090E-02  0.135740E 01       T   -0.394510E-01  0.229410E-01 -0.155580E 00 -0.516630E-02
M=  2          W                P                T                         F               K                O                C
  W   -0.433340E 00 -0.858820E-02 -0.162980E 00       W   -0.370840E 00  0.229980E 00 -0.323160E 00  0.240270E-02
  P   -0.515820F-01 -0.575290E 00 -0.725310E-01       P    0.378230E 00 -0.245750E 00  0.483760E 00  0.336160E 00
  T   -0.122130E-01 -0.345670E-02 -0.526130E 00       T   -0.231770E-01 -0.889460E-02 -0.198790E-01  0.202110E-02
M=  3          W                P                T                         F               K                D                C
  W    0.423630E 00 -0.443310E-02  0.313110E 00       W    0.653920E 00  0.453540E 00  0.284270E 00 -0.222170E-01
  P    0.237880E 00  0.701000E 00  0.605770E 00       P   -0.574380E 00  0.338870E 00 -0.710670E 00 -0.338300E 00
  T   -0.100150E-02  0.300360E-02  0.354310E 00       T    0.627190E 00 -0.804130E-01  0.231690E 00 -0.227000E-01
M=  4          W                P                T                         F               K                O                C
  W    0.336750F-01 -0.715230E-02 -0.818370E-01       W    0.739720E 00 -0.336440E 00 -0.167680E 00  0.320980E-01
  P   -0.130120E 00 -0.792320E-01 -0.667090E 00       P    0.286950E 01 -0.335950E 00  0.883980E 00  0.794680E 00
  T    0.597860E-02 -0.147340E-01 -0.176290E 00       T   -0.372930E-01 -0.418400E-02  0.253680E 00  0.360690E-01
M=  5          W                P                T                         F               K                O                C
  W    0.201250E 00  0.900960E-02  0.224610E-01       W   -0.835210E-01  0.578330E-01  0.236570E 00 -0.202940E-01
  P    0.320890E 00  0.347740E-01  0.201540E 00       P   -0.202570E 01  0.977870E-01 -0.656170E 00  0.120550E 00
  T   -0.127880E-02  0.384270E-02  0.407060E-01       T   -0.309850E 00  0.603510E-01 -0.262020E 00  0.203310E-02
M=  6          W                P                T                         F               K                D                C
  W   -0.250920E 00 -0.664200E-02  0.123520E 00       W   -0.233310E 00 -0.399790E 00 -0.268640E 00 -0.317720E-02
  P   -0.230050E 00 -0.368110E-01  0.134990E 00       P    0.391360E 00 -0.200570E 00 -0.168790E 00 -0.317180E-01
  T   -0.516770E-02  0.107820E-01 -0.688900E-01       T   -0.638510E-02 -0.286860E-01 -0.142470E-01 -0.141710E-01


              G(M,I,J)       FILE NO. = 1021
M=  1          W                P                T              M=  4       W                P                T
  F   -0.797568E-02  0.847277E-03 -0.127989E-03       F   -0.336322E-01  0.283896E-03 -0.152680E-01
  K   -0.478817E-03  0.662570E-05  0.106823E-02       K   -0.139737E-02 -0.847000E-05  0.593260E-03
  D    0.100330E-01 -0.177352E-01 -0.151177E 00       D    0.265783E-01 -0.245954E-01 -0.238065E 00
  C   -0.605565E-01  0.344783E 00 -0.156880E 00       C   -0.685306E-01  0.786531E-01 -0.483212E-01
M=  2          W                P                T              M=  5       W                P                T
  F   -0.233072F-02  0.670030E-03 -0.323503E-01       F   -0.148937E-01  0.970841E-04 -0.128413E-01
  K   -0.568011E-03 -0.317936E-04  0.857275E-03       K   -0.205168E-03  0.433759E-05  0.467837E-03
  D    0.801776E-01 -0.199741E-01 -0.156238E 00       D    0.103675E-01 -0.190314E-01 -0.220442E 00
  C    0.262656E-02  0.310515E 00 -0.109690E 00       C   -0.112261E 00 -0.792663E-02 -0.758610E-01
M=  3          W                P                T              M=  6       W                P                T
  F   -0.310763E-02  0.482460E-03 -0.135378E-01       F   -0.348355E-02  0.108082E-05 -0.921545E-02
  K   -0.124022E-02 -0.114314E-04  0.663584E-03       K    0.159235E-03 -0.293978E-03  0.398037E-03
  D    0.131806E 00 -0.257308E-01 -0.195887E 00       D    0.439358E-01 -0.507333E-02 -0.136157E 00
  C   -0.793786E-02  0.188004E 00 -0.563069E-01       C   -0.273685E-01 -0.160361E-01 -0.600213E-01

 // END
```

Table 1

```
             === AR-MODEL FITTING EXPRESSION ===              TH5201

     [ SYS IN ]           _______________         [ SYS OUT ]
     Y1: N205--S  -->|      - ID -       |--> C4 O2-S  :X1
     Y2: V252-S   -->|     CT8 -B20      |--> NOx-SL   :X2
     Y3: N200-S   -->|                   |--> T205-S   :X3
                     | - OPTIMUM LAG -   |--> T201-S   :X4
                     |        1          |--> V200-S   :X5
                     |___________________|

   [ PROCESS MODEL ]            [ PROCESS MODEL ]            [ PROCESS MODEL ]
X1(n) = .6755E   0X1(n-1) X2(n) = .1430E   2X1(n-1)  X3(n) = .2745E   1X1(n-1)
      + .7998E  -3X2(n-1)       + .8279E   0X2(n-1)        + -.4888E  -1X2(n-1)
      + -.1095E -2X3(n-1)       + -.3692E  0X3(n-1)        + .5235E   0X3(n-1)
      + .8441E  -3X4(n-1)       + -.5340E  0X4(n-1)        + -.3143E  0X4(n-1)
      + .5547E  -3X5(n-1)       + -.7348E -1X5(n-1)        + .5246E  -2X5(n-1)
      + .3502E  -3Y1(n-1)       + -.9790E -1Y1(n-1)        + -.1619E  0Y1(n-1)
      + .6991E  -2Y2(n-1)       + .1433E   2Y2(n-1)        + .2468E   1Y2(n-1)
      + -.7135E -2Y3(n-1)       + .8052E   0Y3(n-1)        + .5145E   0Y3(n-1)
```

```
             === AR-MODEL FITTING EXPRESSION ===              TH5201

     [ SYS IN ]           _______________         [ SYS OUT ]
     Y1: N205--S  -->|      - ID -       |--> C4 O2-S  :X1
     Y2: V252-S   -->|     CT8 -B20      |--> NOx-SL   :X2
     Y3: N200-S   -->|                   |--> T205-S   :X3
                     | - OPTIMUM LAG -   |--> T201-S   :X4
                     |        1          |--> V200-S   :X5
                     |___________________|

   [ PROCESS MODEL ]            [ PROCESS MODEL ]
X4(n) = .2234E   1X1(n-1) X5(n) =-.1443E   2X1(n-1)
      + .8663E  -2X2(n-1)       + .7377E  -1X2(n-1)
      + -.7539E -1X3(n-1)       + -.1551E  0X3(n-1)
      + .4698E   0X4(n-1)       + -.4250E  0X4(n-1)
      + -.3932E -1X5(n-1)       + .6750E   0X5(n-1)
      + .3597E  -1Y1(n-1)       + -.1074E  0Y1(n-1)
      + -.1985E  0Y2(n-1)       + -.5416E  1Y2(n-1)
      + .2472E   0Y3(n-1)       + -.4229E  0Y3(n-1)
```

```
             === AR-MODEL FITTING EXPRESSION ===              TH5201

     [ SYS IN ]           _______________         [ SYS OUT ]
     Y1: N205--S  -->|      - ID -       |--> C4 O2-S  :X1
     Y2: V252-S   -->|     CT8 -B20      |--> NOx-SL   :X2
     Y3: N200-S   -->|                   |--> T205-S   :X3
                     | - OPTIMUM LAG -   |--> T201-S   :X4
                     |        1          |--> V200-S   :X5
                     |___________________|

   [ OPERATOR MODEL ]           [ OPERATOR MODEL ]           [ OPERATOR MODEL ]
Y1(n) =-.1421E   2X1(n-1) Y2(n) =-.1228E   0X1(n-1)  Y3(n) =-.3028E   1X1(n-1)
      + .1008E  -1X2(n-1)       + .2290E  -3X2(n-1)        + .1199E  -1X2(n-1)
      + -.4318E -1X3(n-1)       + .2529E  -3X3(n-1)        + -.1050E -1X3(n-1)
      + -.1060E  0X4(n-1)       + -.3254E -2X4(n-1)        + -.2769E -1X4(n-1)
      + -.1484E -1X5(n-1)       + -.3046E -3X5(n-1)        + .0567E  -2X5(n-1)
      + .6325E   0Y1(n-1)       + -.1403E -2Y1(n-1)        + .7696E  -2Y1(n-1)
      + -.2025E  1Y2(n-1)       + .7255E   0Y2(n-1)        + -.1746E  1Y2(n-1)
      + -.1069E -1Y3(n-1)       + -.8788E -2Y3(n-1)        + .5570E   0Y3(n-1)
```

Table 2

control through the design by dynamic programming can be easily executed and displayed. Furthermore, optimum production level and pursuit control are also realizable for hierarchy control systems.

These functions and contents are referred by paper (11) (12) (14) .

Fig.4 is a charts of the result of one step ahead prediction, for Tbz (burning zone temp) , KW (power of driving motor) and kiln speed (rev./hour) , and one ahead prediction error chart is also presented. Fig.5 shows simulation of mud-ring failure. As is shown in Fig.5, mud-ring failures are figured out and finding of this event occurrence is significantly valuable for thereafter operation against probable emergent situation.

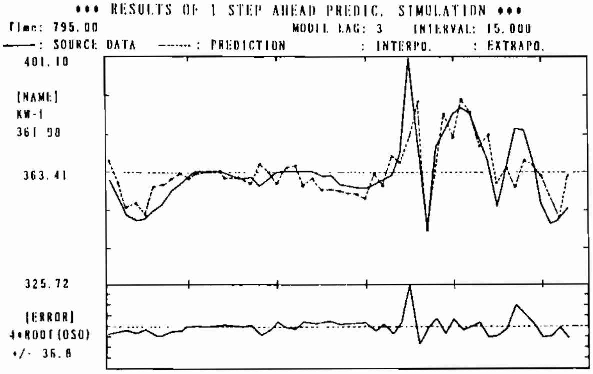


Fig. 4

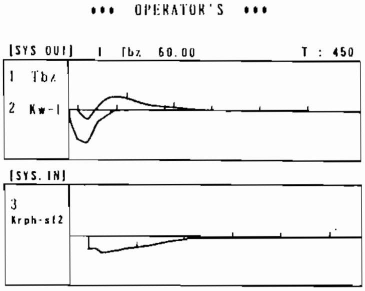


Fig. 5

4. Application of analytical artificial intelligence

The control mode must be selected under the decision whether the process is stational or not.

Classification between stational state and non-stational state, and detection of these transient state are necessary.

Fig.6 shows moving average of one-step ahead prediction error of the running data under the computer control of cement rotary kiln (3) .

When this moving average of one ahead prediction error happens to shift up-ward or down-ward, we can realize something happens in the process.

For example, when fuel quality such as kilo calories is changed gradually, we can not usually find this change of quality in real time base, because we can only monitor the quantity of flow.

Against the probable causes as mention above, corresponding operation are executed, After these procedure, process shifts back again to stational state from non-stational state.

When it is decided that the process is fallen in non-stational state, control mode is changed over from objective control system, so called ※1 siltac system, to subjective system, so called expert system.

Detection of cause by inference of backward resoning, and control decision by forward reasoning are performed by ※2 ESPA (Trading name of Expert Shell) [5][6][16][17]

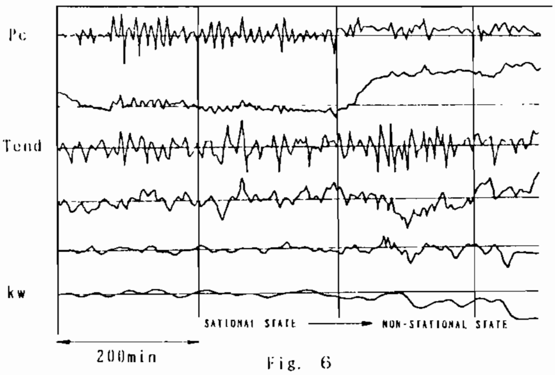


Fig. 6

5. Spectral observation of hidden information for process insight

It is natural that variation of KW power of the kiln driving motor is depended on the load. And the state of coating covered on the inside of kiln is irregular and rough and the travelling speed of raw material passing through the kiln has strong correlation with these uneven shape of coating. Then heat transfer efficiency is strongly depended on this [7] .

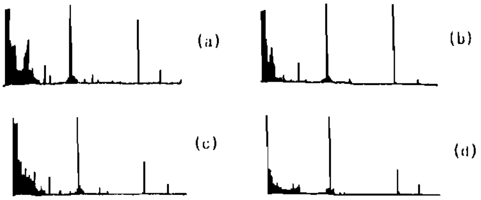
Therefore, spectral analysis of power within the range around one revolution, is likely considered to be valuable for the insight into kiln process situation.

When process is under normal and good conditions, the spectral value of specific frequency is rather high, and to the contrary, when bad condition, then spectral value is lower.

Examples of these results are shown in Fig. 7.

Fig. 8 shows time series spectral values of draft in combustion chamber. These intensity of spectra seem to be valuable information about state of combustion. That is, when combustion condition is good, spectral value is rather high and under bad condition, spectral value is lower at the specified frequency.

These examples present the existence of the information behind the process. These measures with intelligence are thought to be useful for expert system, even though these cause and effect relation is qualitative in stead of quantitative.



(a) normal (b) mud-ring failure (c) after cleaning of tower inside wall by water gun (d) bad condition

Fig. 7

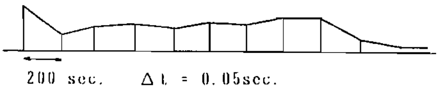


Fig. 8

6. Optimum production level (optimum set point)

The process balances at the value (state) of controlled variables responding to the set points of manipulated variables. As set points of process variables have ever been adjusted by experience of operators or staffs based on static heat balance and mass balance, it doesn't always satisfy the requirement from production management to produce high quality products with a minimum cost. I think the set points based on experience doesn't always guarantee the requirement of them. The characteristic of low frequency near zero is occured by natural causes including external causes of process, and considered to be autonomous changing of process by itself. And then it provides a characteristic of low frequency of energy cost and quality.

Autoregressive model (AR model) of discrete type is built through the identification of behavior. Output variables of the process (controlled variables) are expressed as dimensional vector x (n) , and input variables to the process (manipulated variables) as l dimensional vector $y(n)$, then the model is expressed as the following equation :

$$x(n)-\overline{x}=\sum_{m=1}^{M} Am\{x(n-m)-\overline{x}\}+\sum_{m=1}^{M} Bm\{y(n-m)-\overline{y}\}+\Sigma\varepsilon(n) \qquad (1)$$

Where $\overline{x}$: mean value of actual.data $x(n)$

$\overline{y}$: mean value of actual data $y(n)$

Am : $r\times r$ dimensional coefficient matrix to each m where $m=1, \cdots\cdots\cdots M$

Bm : $r\times l$ dimensional coefficient matrix to each m where $m=1, \cdots\cdots\cdots M$

$\Sigma\varepsilon(n)$: r dimensional white doise vector of 0 average.

When the sampling intervalΔt is smaller than the time constant of the process, equation (1) represents process model for stabilization control, and expresses dynamic characteristics of relatively high frequency zone. On the other hand, whenΔt is large (Δt > time constant) , a model only with extracted low frequency zone is obtained by taking an average of running data during Δt's, because frequency characteristics is filtered out.

As the problem of production level is started from the problem of determining the set points, the model is expected to be expressed by equation (1) only with low frequency characteristics. Set points are originally determined by theoretical physical or chemical models, but in the practical control cases they mostly manage unobservable variables which cannot be treated by theoretical models. Therefore, models of equation (1) obtained from actual data of daily operation are needed as more practical models.

The above-mentioned method has been applied to the problem of optimum production level for a rotary kiln process. [13] [15] Optimum production is performed under a set of set points of each variable to manufacture high quality products with the least fuel original unit (L/ton) within the given constraints. The model is given in Fig.9 and Table 3. Variables of constructing the model, which are considered to be significant as production level, are selected.

Five variables, $x_1 \cdots x_5$, are chosen as process output (controlled variables) , and other five variables, $y_1 \cdots y_5$ are chosen as process input (manipulated variables) . Total of these ten variables construct the model. Data used to identify the process are shown in Fig.10. Time series data are average values during eight hours operation. Sampling intarvals is 8 hours, and data length N is 201. Fig.11 and Fig.12 present the relative contribution of power spectra density between noise and power of variable at frequency f, and will conveniently be used for graphical representation like these.

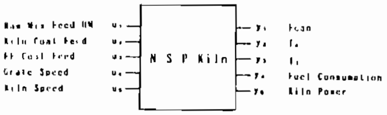


Fig. 9 Optimum Production Model

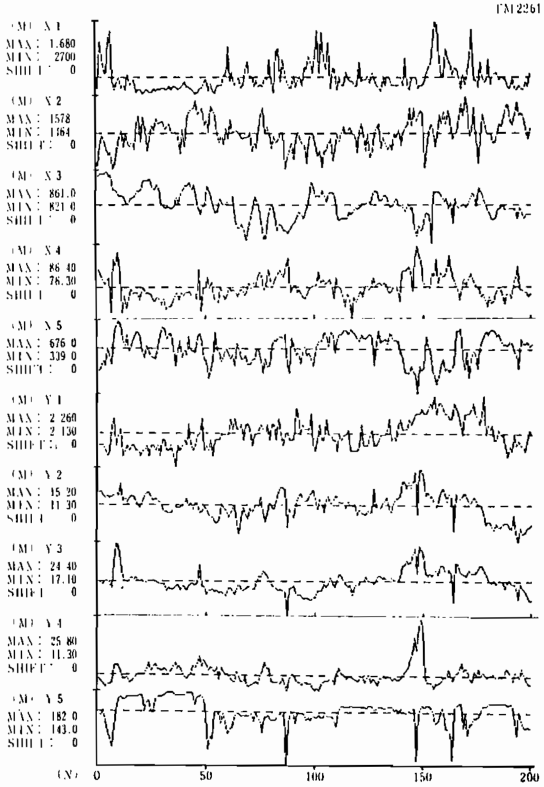


Fig. 10 Data of Optimum Production Level

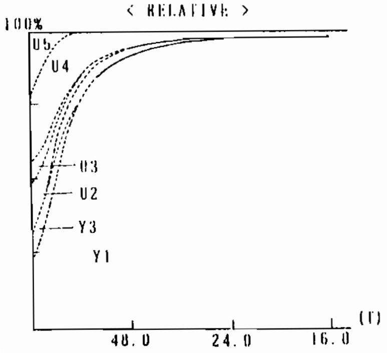


Fig. 11

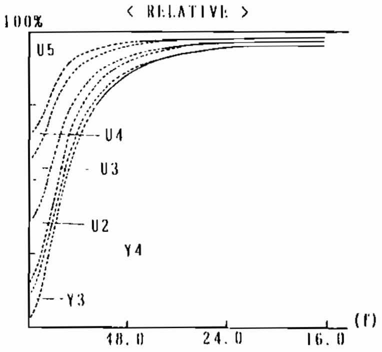


Fig. 12 Power Contribution

As shown in graph, for example, about 75% of quality (x) are explained to be effected by variables other than x at low frequency. Also, almost 100% of fuel original unit (x) is caused by other variables at low frequency region.

We have selected variables that are considered to be relative to fuel cost and quality, composed multi-dimensional autoregressive model shown in Fig. 9 and Table 3, and by using the average time series data of every eight hours, identified them by AIC criterion. The resultant model that we call Set point Model was expressed as following equation ;

$$y_i(n) = \sum_{i=1}^{M} A_i y(n-1) + \sum_{i=1}^{M} B_i u(n-1) + e_i(n) \qquad (2)$$

Output variables of the process and input variables to the process are expressed as five dimensional vector y(n) and u(n) . In order to introduce the characteristics of steady state, the final value theorem is applied to equation (2) and if steady state value of process is expressed as Ys and Us, the following equation is obtained.

$$Ys-\overline{y} = [I-\sum_{i=1}^{M} Ai] \; (\sum_{i=1}^{M} Bi) \; (Us-\overline{u}) \, .$$

$$kp = -[I-\sum_{i=1}^{M} Ai] \; (\sum_{i=1}^{M} Bi) \, , \text{ kp : stational gain of process.}$$

$$Ys-\overline{y} = -kp \; (Us-\overline{u})$$

Then stational model is presented as follows.

$$Ys+kpUs=Es \; (\overline{y}, \; \overline{u}) \qquad (3)$$

Optimum production problem is to minimize following objective function (4) under the equation (3) and constraints (5) .

$$J = \sum_{i=1}^{M} \alpha iYsi \; + \sum_{j=1}^{M} \beta jUsj \qquad (4)$$

$$Ly \leqq Ys \leqq Uy$$

$$Lu \leqq Us \leqq Uu \qquad (5)$$

Lastly principal component analysis has been carried out to determine parameters of an object function as given in Table 4. The first principal component P is found to be ralated to quality fuel original unit and combustibility from the size of the absolute value. Therefore, using the first principal component P as an object function, J is expressed as the following equation for eigenvectors of P

$$J = 0.26y + 0.04345y - 0.06994y + 0.49704y - 0.3835y$$
$$+ 0.403u + 0.3484u + 0.34282u + 0.19557u - 0.30251u$$

Then we obtain Ys and Us as the solution which make equation J minimum. Resultant optimum production level affect the fuel original unit, and it is estimated that about 1.47 L/ton is decreased. This means 1.8 % saving of total fuel consumption comparing with conventional product level as shown in Table 5. This naturally assumes preconditions that manipulated variables which do not be not adopted in the model are set almost perfect.

```
              A(I),B(I)

    I =      1

MATRIX        10     X    10

              1             2             3             4             5             6             7             8             9            10

  1      .26830E+00  -.76692E-03  -.17172E-02   .56084E-02  -.52127E-03   .71925E+00   .25235E-01   .25787E-01  -.38625E-01  -.21406E-02
  2     -.28041E+00   .50012E+00   .40670E+00   .21468E+01  -.37569E-01   .13210E+03  -.67107E+01  -.63745E+00   .10814E+01   .80251E+00
  3      .15953E+01  -.37401E-01   .80640E+00   .19111E+00  -.19944E-01  -.33742E+01  -.79327E+00  -.70720E+00   .12255E+00   .21618E+00
  4      .29574E+00  -.19661E-02  -.29006E-01   .21625E+00  -.45630E-02   .79398E+00   .37094E+00   .39914E+00  -.11302E+00  -.52672E-01
  5      .77376E+01  -.40726E+00   .77337E+00  -.31969E-01   .68406E+00  -.28730E+03  -.88471E+01   .11348E+01   .34717E+01  -.12039E+01
  6      .54336E-02  -.15394E-04  -.47647E-03  -.10626E-02  -.40663E-04   .47293E+00   .10867E-02   .75163E-02  -.91082E-03  -.76220E-03
  7     -.16102E+00  -.49402E-02   .10046E-02  -.28896E-01  -.57415E-03   .19537E+01   .68347E+00   .57940E-01   .19724E-02  -.13742E-01
  8     -.56071E+00  -.53317E-02   .15426E-01  -.78764E-01  -.13495E-02   .62021E+01  -.47013E-01   .84046E+00   .50768E-02  -.36565E-01
  9     -.32084E+00   .19413E-02   .15786E-01  -.76866E-01  -.11802E-02   .18490E+01  -.11547E-01   .20693E+00   .72901E+00  -.38562E-01
 10     -.34768E+01  -.47691E-01   .93821E-01  -.68255E-01   .45424E-02   .25477E+02  -.18600E+01   .22048E+00   .31020E+00   .58969E+00
```

Table 3 Parameters of Process Model

	P_1 2.76697	P_2 2.22318	P_3 1.42275	P_4 1.04725	P_5 .75515	P_6 .55163	P_7 .40225	P_8 .32902
X 1	.25000	− .42546	.13217	− .22239	− .43931	− .04577	.52599	.42414
X 2	.04345	.41921	− .32371	.24720	− .61255	− .01005	.23116	− .41499
X 3	− .06994	− .09522	.61729	.44260	− .38301	− .25571	− .34203	− .02544
X 4	.49704	− .05462	.07764	.02719	.33548	− .33904	.32432	− .44662
X 5	− .33254	.09255	.21378	− .36353	.02311	− .32557	.03611	− .27651
X 6	.40347	− .05327	− .18233	− .47756	− .33916	.18663	− .56080	− .11576
X 7	.24542	.14293	.47407	.06798	.14590	.52428	.05368	− .16277
X 8	.34252	.38534	.27363	− .24549	− .05190	− .37637	− .07906	.10318
X 9	.19557	.54485	− .12491	.11138	.13335	− .16490	− .04431	.56123
X10	− .30251	.39220	.20696	− .25328	− .11637	.31758	.27919	.06245

(P: PRINCIPAL COMPONENT)

Table 4

	New SV	Ave.	DEV.
Fcao : Y_1	0.3636	0.5984	-0.2348
T4 : Y_2	1513.9	1521.8	-7.9
T1 : Y_3	840.42	842.52	-2.10
Fuel Consumption : Y_4	79.110	80.586	-1.476
Kiln Power : Y_5	617.63	550.586	67.29
Raw Mix Feed BM : U_1	2.1651	2.1919	-0.0268
Kiln Coal Feed : U_2	12.565	13.256	-0.691
PF Coal Feed : U_3	19.473	20.594	-1.121
Grate Speed : U_4	16.548	14.576	1.972
Kiln Speed : U_5	170.28	172.02	-1.74

New SV = New Set Value
Ave. = Average
DEV. = New SV - Ave.

Table-5 Results of Optimum Production Level

7. Pursuit Control to Optimum Production Level

From the discussion mentioned above, an optimum production level can maintain a decrease in costs and high quality. However, in order to continually realize optimum production, it is necessary to use pursuit control with many variables for the present state of process to transfar smoothly to new production levels.

In Fig.13, simulation of pursuit control of set point exchange, using process model (Table 3) is shown. The upper half of Fig.13 indicates control when step-like disturbance is added to the variable x and the lower half is simulation of pursuit control to set point exchange of optimum production level.

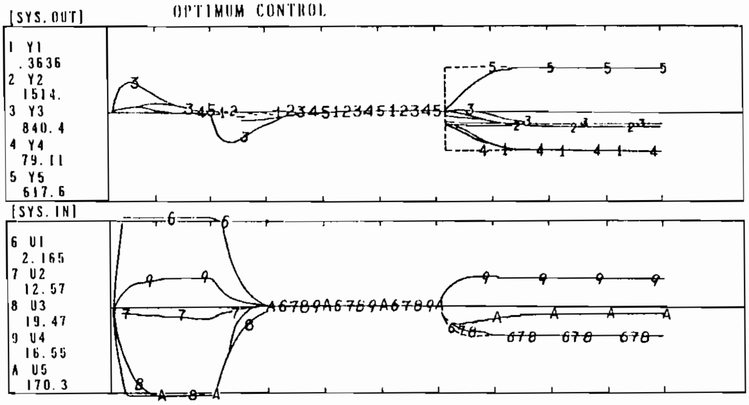


Fig. 13

8. Conclusion

We have reported a consistent approach with examples from design of optimum production level to pursuit control, and to their constant stabilizing control. The constant stabilizing control is devided into two regions, that is, stational state and non-stational state.

When big disturbances happen to occur, which are the inherent character and probable significant event, then process condition shifts to ill-defined situation from well defined situation. Under such a circumstance, temporaly back-up function based on artificial intelligence approach enforces the process to shift back again from the unstational state to stational state. A arrangement of these as a whole is so called hierarchy system.

In this paper, the validity of the cooperation with both objective and subjective approach, and building a coordinating and advisory system are also referred. Practical applications to the cement making process by means of Siltac tool and problem solving agent inference "ESPA" engine are presented.
The hybrid control system, with hierarchy structures which share with each function how and what to do, as mentioned meaning, is preferable for robustness of control and expansion of control capability. We don't expect expert system to completely replace conventional control or humans operation. Expert systems may be assistant as advisory systems around the objective control systems.

And further more, the knowledge acquisition through both mathematical and symbolical style is evaluated and complementary corporation of both is expected for further progress.

These systems with supporting circumstance which are called SOIDECS[※3] are to be demonstrated in IIASA conference which will be held in Kyoto Japan, Aug., 1986.

※1. Self-Instructive, Learning and Tutorial system for statistical Analysis and Control of dynamic systems.

※2. Expert Shell Partnership Agent.

※3. This epoch-making system named SOIDECS (SSK's Objective and Intelligent Model Design Environment and Control System) realizes a novel concept through combination of the AI techniques with the multi-input and output model identifying and control techniques on the basis of a series of SSK's system products.

The SOIDECS system basically consists of the following elements.

1) SILTAC : Self-Instructive, Learning and Tutorial system for statistical Analysis and Control of dynamic systems
This system for statistical analysis and control of a dynamic system implemented on a personal computer performs prediction and control of complicated processes.

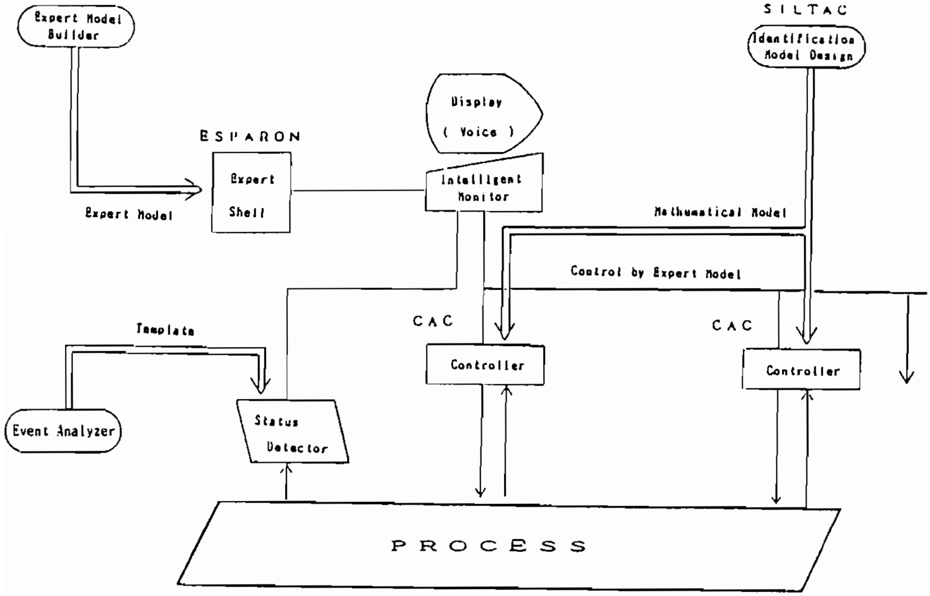
2) CAC : Compact Advanced Controller
This compact-size controller manipulates the control models of the multi-input and output system identified by SILTAC.

3) ESPARON : Expert Shell for Partnership Agent with Rule Organized Network
This expert shell implemented on a personal computer to construct a knowledge base system is capable of constructing cooperative type models and systems for FA networks.

4) Intelligent Monitor : Process monitor with expert system and intelligent detector

5) Display Human Interface : Display facility with animation and voice synthesis

Fig.14 shows the connection of functions.



Remarks : In this demonstration, a simulator with a plant model of SILTAC in place of an actual process is connected.

references.

1. H.Akaike "On the use of a linear model for the identification of feedback systems " Annals of the Institute of statistical mathematics. Vol.20, No.3, 1968

2. T.Otomo, T.Nakagawa, H.Akaike "Statistical Approach to computer control of cement rotary kilns" Automatica, Vol.8,pp35−48 Pergamon Press 1972

3. H.Akaike, T.Nakagawa "Statistical Analysis and control of dynamic systems " Science Book Publishing Co., Ltd. Tokyo Japan.

4. T.Nakagawa, Y.Yagihara "Identification of optimum production level and control " Lecture note in Company by English.

5. H.Ogawa "Representation of knowledge, based on Frame Work" Information Processing Society of Japan 26.12.1985

6. H.Ogawa "Hybrid Problem Solving Agent" Information Processing Society of Japan Jan. 1986

7. T.Nakagawa "Study on the control of cement rotary kiln" Phd. Dissertation, University of Tokyo (1964)

8. K.Y.Wong, K.M.Wing, E.J.Allbriton "Computer control of the clarksvill cement plant by state space design method " IEEE. cement industry Technical Conf. U.S.A. May 1968

9. J.I.Tou "Optimum Design of Digital Control Systems " Academic Press, New York, 1963

10. K.J.Astrom "Introduction to Stochastic Control Theory " Academic Press

11. Siltac operation manual, S.S.K. Co., Ltd. Tokyo Japan

12. K.Arai "Energy Economizing Measures in clinker burning process at Chichibu Cement Co." Proc. of 20th international cement seminar, Chicago, U.S.A. Dec. 1984 P.47－P.51

13. T.Nakagawa, Y.Yagihara "The approach to the design of optimum Production Level and pursuit control in industrial process " Jour. of SICE (Japan) vol.24, No.11, 1985

14. Manual of talk-master Y.N.S. Co., Japan

15. S.Hagimura, Y.Yagihara et al. "The hierarchy control with stability and Production Level control of cement NSP kilns" Submitted in IFAC conf. Tokyo, Japan, Aug., 1986.

16. Technical note on ESPA, S.S.K.Co.,Ltd., Tokyo, Japan.

17. T.Nakagawa, H.Ogawa "The Identification and control, partially added with the Artificial Intelligence approach " 4th IFAC／IFIP symp. Graz. Austria, 1986

ADAPTIVE CONTROL AND GROWTH PROCESSES

Peschel, M.,Academy of Sciences, GDR

Mende,W., Academy of Sciences, GDR

Breitenecker, F.,Technical University of Vienna

Kopacek, P.,Joh.Kepler University, Linz

Introduction

In this paper we are going to study the interrelationship between growth phenomena and control.

On one hand growth in complex systems like individual species or populations with interaction between different kinds of species is mostly based on socalled cooperative structures leading to clusters built up from some of the species or subsystems and competing with other but similar clusters.

This complex interaction structure can usually be considered as a certain network composed of interacting modules.

In this network we meet frequently feedback loops leading to internal controlmechanisms aiming to a global equilibrium.

This global equilibrium can be a static one equilibrium point or a dynamic one exposing oscillating regimes-limit cycles.

Thus in nature we find control phenomena and even adaptive control. Therefore it seems quite reasonable, to learn control principles from nature.

We use this source for concepts of new nonlinear basic controller as a substitute of the well-known classical PID-controller.

For this purpose we studied growth transitions in different applied fields, we established the socalled EVOLON-concept for a simple but reliable description of a growth-step, and we will now use this EVOLON-concept for the design of new basic controllers.

On the other hand in practice, especially in biotechnology we are often confronted with the necessity to apply an additional control from outside besides the internal autonomous control of the bio-ecosystem.

For the design of a good external control strategy we need a reliable model-description of the bio-ecosystem.

In the recent years we developped the socalled Lotka-Volterra approach for applied Systems Analysis which declares the Lotka-Volterra equations

$$F\, x_i = x_i\left(\sum G_{ij} x_j + \sum H_{is} y_s\right) \quad \text{with } F = d\, \ln / dt$$

as a relevant concept for modelling bio-ecosystems.

This approach together with a lot of concrete studies is published in /1/.

If we use the Volterra equations as model descriptions for real systems the problem arises, how to control such a system.

This could be done using control influences y_s as a linear superposition on the autonomous driving force on the right-hand sides of the model differential equations.

But there exist some different control concepts for applying an external control onto the system.

In the second part of this paper we inform about our control strategies for the control of growth in connection with the proposal of an adaptive controller.

Lotka-Volterra equations possess a very interesting expansion property which in some sense make them more attractive than a Taylor expansion of the nonlinear Dynamics of a system.

In the third part of this paper we communicate some informations about this idea.

Concepts for controllers and control strategies in connection with growth processes

Design of nonlinear basic controllers with the EVOLON-concept

We first give a short description for the EVOLON as a model for a growth step.

Every evolutionary process of a system, of an individuum, of a new technology, a population etc consists of a staircase(increasing or decreasing)of s-formed growth-steps.

Behind this form of such a growthstep usually an elementary evolution process is hidden, which we call an EVOLON, if some characteristic properties will be present.

Such an evolution step expresses a rich manifold of different interaction mechanisms within the system considered and between this system and its environment.

For the process called EVOLON we have to distinguish between an extensive phase at the very beginning and an intensive phase at the end of the process, when we can observe a saturation on the next higher level, on a new steady state.

In its extensive phase the system builds up a cooperative structure with the aim, to create himself optimal conditions for a quick growth. The consequence is a rapid consumption of the available resources and their transformation into ever increasing growth-rates.

But in the intensive phase cooperation for the purpose of increasing growth-rates makes no longer sense, more and more the system is encountered with competition phenomena caused by arising competitors and maybe a growing scarcity of resources.

Therefore the preestablished cooperative structure begins to loose its strong links, the system exposes a tendency to decomposition into parts accompanied by the trial to find a new cooperative structure by recombination of the parts which better fits into the new condition of development.

The system tries to open a door into the future by starting the next s-formed evolution step.

From basic considerations to implement this system philosophy in a corresponding formal model description and from our experiences with data-analysis for time-series in different fields we could draw the conclusion, that such an elementary growth process for a onedimensional growth indicator x in a highly aggregated robust system can be reliably described by the model of the socalled hyperlogistic differential equation

$$dx/dt = K\, x^{k} \left(B^{w} - x^{w}\right)^{l}$$

In this model the introduced parameters have the following meaning

- K is the driving force amplification factor.
- k is a measure of the complexity of the cooperative structure in the background of the growth phenomena.
- l is a measure for the complexity of the growth-damping interaction of the system with its environment.
- w is a velocity parameter of the approach of the growth indicator x against the next saturation level B.
- B is the steady state approached in the considered growth-step.

With this 5-parametric family of growth-curves a rather rich manifold of s-formed transitions can be described with a flexible adaption capacity to real growth phenomena.

Most of frequently used growth models in ecology, economy, agriculture are special cases of this family.

Also the broadly used logistic growth model with $k = l = w = 1$ is contained in the hyperlogistic growth model.

In some sense the hyperlogistic growth model is at the same time the most simple Lotka-Volterra system.

By introduction of two additional state-variables x_1 and x_2 according to the socalled Structure Design Principle /1/, the hyperlogistic model can be equivalently transformed into the following Lotk-Volterra system

$$dx/dt = x \,.\, K \,.\, x_1$$

$$dx_1/dt = x_1 \cdot K\left((k-1)\, x_1 - w \,.\, l \,.\, x_2\right)$$

$$dx_2/dt = x_2 \cdot K\left((k+w-1)\, .x_1 - w.(l-1).\, x_2\right)$$

This is an elementary autocatalytic predator-prey system with the prey x_1 and the predator x_2 on which a nonlinear observer- an exponential integrator- evaluates the growth of the prey x_1.

If we now try to make use of the EVOLON concept for the design of new basic controllers we meet for this purpose two different possibilities, namely, we can use immediately use the hyperlogistic differential equation, or we can apply this unfolded form after transformation of the hyperlogistic differential equation into the corresponding Lotka-Volterra equations.

In the following we will describe in general terms the possibilities we recognized for using the EVOLON as design concept for new controllers.

. EVOLON- behaviour in a nonlinear follower system

Here the control system S shall be a linear follower system, proportional with a transfer factor equal to 1.

B, the saturation value of an EVOLON usually a relatively slowly changing time-function $B = B(t)$, shall be the leading variable being providing at the input of the control system S.

After measuring the actual value x(t) of the control variable x the controller R has to produce such a correction at the input of the control system S, that we get a good following process of the curcuit with the control feedback.
Our essential orientation for this concept shall be that we demand a certain transition process of the closed curcuit which shall more or less exactly realize a wanted EVOLON.

. Modified EVOLON-concept for the control curcuit as a whole
In analogy to the case just discussed we demand that the whole behaviour of the closed control curcuit reproduces the behaviour of a prescribed EVOLON. In comparison with the first case we here introduce a demand-value x_S of the control variable x.
Therefore now the controller has to act on the base of a comparison between the actual value of x and the demand value x_S.
This difference gives rise to greater differences in the concrete implementation of the EVOLON, now we must admit sign changes of the power-functions within the EVOLONdescription. To avoid difficulties with these change of signs we must introduce special modifications of the power functions definitions (even and uneven power functions) for arbitrary real exponents.

. Application of the EVOLON immediately within the controller
In comparison with the two concepts just considered we know concentrate our attention no longer immediately on the control variable x, but on the controller output variable xI. Now xI for us will be the relevant growth indicator which now should follow the EVOLON model. As in the case 2 above discussed we must also here allow for change of signs in the arguments of the corresponding power-functions. In this case we can in general not expect that the controller R will fulfill one EVOLON-like state transition, if this is the case for the control variable x. Usually the

controller R has now to realize a whole evolution staircase, maybe going upstairs and downstairs during one EVOLON transition of the control- variable x.

Interesting in this aspect is also the idea to consider x and xI as species in a 2-nd order Lotka-Volterra-system or in a Lotka-Volterra network composed of two coupled 2-nd order Lotka-Volterra systems respective two coupled EVOLONs.

- Overall behaviour of the control curcuit shall be an EVOLON in its unfolded form as a Lotka-Volterra system

In this case we repeat the first approach, but we use for this purpose the EVOLON in its unfolded Lotka-Volterra form. This can be simply done only for the first case where sign changes of the arguments of the power- functions are not possible or can be avoided. This concept also works, but it can be remarkably qualified, if we now make explicit use of the 3 state description equations we have at hands. We introduced for example another control correction immediately into the equation for the state-variable x_1, the prey, proportional to the control deviation $x_S - x$, which gives a much better control process in comparison with the case without this additional correction.

Design of an adaptive multivariable controller for growth processes

Properties of classical relais switching controllers

The adaptive microcomputer controller proposed here is based on the socalled classical relais switching controller in combination with a simple basic controller(We can use for this purpose the wellknown PID- controller or a EVOLON basic controller just discussed in this paper).

In our considerations we rely much on the dissertations /2/,/3/, /4/. In these references the dissertation /2/ had a certain pioneer role.

At first we give some informations about onedimensional relais-switching curcuits working under the following conditions

- The forward plant, the control system S, is an arbitrary linear system with constant parameters, in most cases only a linear chain composed of PT_1-modules.
- In the references mostly the nonlinear autooscillations occurring in the closed loop were studied, therefore an application of demanded values x_S usually were omitted.
- In the feedback channel there is a symmetric relais plant, which in the normed case, without restriction of generality is switching between the two levels 1 and -1.
- With a prescribed switching- period T_a samples will be taken from the output of the relais and corresponding to the sign of the relais input signal either positive or negative pulses will be generated.
- The pulse taken with the period T_a from the relais output is then added to the actual statevalue Int of an integrator after being modified by the factor G_y, the transfer factor of the integrator.
- The integrator Int produces immediately the correcting input

signal of the control system S.

This feedback curcuit generates at the switcher output interesting and sometimes rather complex autooscillations. These combination oscillations consist of impulse tracks, socalled halfperiods, of consecutive pulses of the same sign.

Important for the following design of an adaptive controller are the following properties of the considered control curcuits.

. Practically the autooscillation which establishes in the closed control curcuit does not depend on the initial values of the state variables of the linear control system. Only in the transition process ending with the occurrence of a complex autooscillation some traces of the initial values of the states of the control system can be observed.

The established autooscillation at the end has forgotten completely its origin and only depends on the parameters of the control system S, the switching period T_a and the integrator initial value y_O.

. The integrator initial value in normalized form y_O / G_y will be not forgotten in the process of the establishing of the combination oscillation, in contrary this is a very important parameter for the nonlinear autooscillations. In whole connected intervals of the normalized integrator initial value and dependent ȯn the values of time-constants of the control systems or corresponding ratios with the switching period T_a we meet the same autooscillation.

In this aspect the ratio y_O / G_y is extremely important for the dynamics of the closed loop. But at the same time we recognize a farreaching symmetry of this influence.

There exists a fundamental interval, namely

$$-0.5 \leqslant y_0 / G_y < 0.5$$

Outside of this fundamental interval the whole picture of existing autooscillations repeats periodically with the periodicity of integers.

. Of great significance for the complexity of the autooscillations is the ratio

$$\varkappa = T / T_a$$

Here T is the biggest time-constant of the linear control system S. Here the following empirical finding is of importance.
The larger the value of $\varkappa$ is, the longer will be the length of the halfperiods of the arising autooscillation.
For the concept of the adaptive controller, we shall propose, this property is very important.

. In superficial consideration one might expect that in the closed loop curcuit only such combination oscillations can be stablized for which no longlasting deviation $x_S - x$ can exist. But this argument from the linear control theory does not hold here.
In the nonlinear case a nonzero mean value of the pulse tracks can easily be compensated by a corresponding integrator constant.

. Autooscillations in switching relais control curcuits are relatively robust from some different reasons.
There are whole existence areas constructed on the parameters T_i / T_a and y_0 / G_y, in which, despite of parameter variations within the limits of such an area, the same combination oscillation exists.
This phenomenon does not depend on the initial values of the

control system S as was already mentioned above. There this combination oscillation should be stable against some fluctuations of these control variables, that means the same autooscillation reestablishes after a certain transition regime following the disturbance has ended.

This promises that controllers based on the switching relais control principle should be rather robust against disturbances in the state space as well in the parameter space of the control system S.

Heuristical aspects for the design of an adaptive controller

The basic idea for an adaptive microcomputer controller makes use of the empirical finding, that with an increasing ratio:

$$\varkappa \approx T / T_a$$

the halfperiod duration of a complex combination oscillation in a closed switching controller curcuit will also increase at least in tendency.

Naturally with growing halfperiod duration also the control deviation

$$x_S - x$$

should increase and we should expect larger deviations from the demanded value of the control variable x.

Maybe D the basic computation tact of the microcomputer controller. Then we generally put

$$T_a \approx D. A$$

where A shall be a natural number.

For the generation of adaptive effects we follow the idea :

If we increase A, then with a certain tendency and accounting also the dynamic influence from the control system S the halfperiod duration of the combination oscillations will decrease because of the decrease of $\varkappa = T / T_a$. The result will be also a diminuation of the control deviation

$$x_S - x$$

Therefore we should l for concrete mechanisms on the base of which we can reasonably adaptively change the value of A.
At the same time we plan adaptively to change the integrator transfer factor G_y.
By a lot of simulation experiments we could show the following properties of the proposed adaptive controller.
We used for an adaptive multivariable controller a combination of a switching relais controller with adaptive change of A and G_y with a classical PID- controller separately in every feedback channel of a multivariable control curcuit without introducing a corresponding decoupling network.
Instead of decoupling we applied in different ways a reference control and a diagonal control - control of selfreproduction rates.
By these means we could observe satisfying control transitions for multivariable linear systems and Lotka-Volterra systems for the task to design a good follower controller.
We could use these control approaches also for the problem of parameter identification in linear- multivariable systems and in Lotka-Volterra systems.
We could with good success identify diagonal elements of these types of systems.
We met some difficulties in applying this control concept on systems with chaotic attractors (Schulmeister-Shelkov model).

Control layers for the control of nonlinear systems via transformation into a Lotka-Volterra description.

Frequently the Taylor expansion is used as an approximation of nonlinear dynamic systems, for example in the analysis of autooscillations in mechanical systems.

This approximation goes after the following steps.

We consider a nonlinear dynamical system with the following equations

$$dx_i / dt = f_i (x_1, x_2, \ldots, x_n) \qquad i\ 1,2,\ldots,n$$

We first try with a linear approximation based on a Taylor series expansion in the neighbourhood of a certain reference point, which has the form

$$dx_i / dt \approx \sum a_i x_i$$

A corresponding control concept would in this case refer to this linear multivariable description.

If we are forced to apply a better approximation of the given nonlinear system by a model, we would, following-the Taylor series expansion line, propose to take into account now the quadratic components, that means to use now the better approximation

$$dx_i / dt \approx \sum a_i x_i + \sum a_{ij} x_i x_k$$

This is obviously a more complicated model, for which we have now to design a good multivariable controller.

On the next step we would account for the products of 3 state-variables and so on.

It is likely to interpret this process of model refinement in the following way.

We interpret the linear model as a competition of the x_i within a common medium, a substrate.

Taking into account the quadratic terms we interpret this as a two-particle rendevous in connection with a competition within a common substrate. The third order model will then be characterized as an additional accounting of three particles interactions and so on.

We are afraid that following this way the control complexity ever increases.

This will be not the case, if we use the Lotka-Volterra approach for such a systems analysis.

Here we use as a first approximation the Lotka-Volterra approach.

$$dx_i / dt = x_i \left(\Sigma G_{ij} x_j\right)$$

interpreting this from the very beginning as a two-particle interaction without a rather neutral competition for a common medium.

From the very beginning we rely here on a two particle interaction. On the next step of model approximation we should take into account three particle interaction using now additionally the third orcomponents

$$dx_i / dt = x_i \left(\Sigma G_{ij} x_j + \Sigma G_{ijr} x_j x_r\right)$$

This obviously is in comparison with the Lotka Volterra description a model with higher order nonlinearities.

It is characteristic that nature always repeats the use of given means on lower and higher levels. This feature of nature can be reproduced by the Lotka-Volterra approach which offers something like a linear hull operation on different levels.

If we introduce for the products $x_r x_j$ new state- variables

$$x_{jr} = x_j \, x_r$$

we at once get also for these state- variables a Lotka- Volterra description, because of

$$F\,x_{jr} = F\,x_j + F\,x_r \qquad \text{with} \quad F = d\ln / dt$$

The same can be done if we take into account higher and higher orders of interaction.

Therefore, if we are able to design suitable control strategies for systems in a Lotka- Volterra description, we can propose similar controls for the models on different hierarchical levels arising together with a model refinement above described.

Therefore it seems to us so important to design adaptive controllers for the control of Lotka- Volterra systems as feasable growth models for rather complex highly aggregated nonlinear systems.

References

/1/ Peschel, M., Mende,W. : The Predator-Prey Model: Do we live in a Volterra World. Akademie-Verlag,Berlin and Springer-Verlag, Wien, 1986

/2/ Franke, K.: Beitrag zur Ermittlung stationaerer Eigenschwingungen einer Klasse von Abtastrelaisregelkreisen. Diss.A, Technische Hochschule Karl-Marx-Stadt,1968

/3/ Regel, P.: Untersuchungen zu Abtastrelaisregelkreisen. Diss.A, TH Karl-Marx-Stadt, 1971

/4/ Haase, H.,Helbig, H.: Untersuchung einer speziellen Klasse von Abtastrelaisregelkreisen und deren Ausnutzung zur Kennwertermittlung. Diss.A, TH Karl-Marx-Stadt, 1973

ON THE IDENTIFIABILITY OF FACTOR ANALYSIS MODELS

Giorgio Picci [+]
Electrical and Computer Engineering
Arizona State University, Tempe, AZ 85287

ABSTRACT:

We consider Factor Analysis models representing two blocks of variables and discuss the problem of "identifiability" or what we rather prefer to name <u>model selection</u>. For general Factor Analysis (or equivalently, for Error in Variables) models this problem is apparently still unsolved although raised and discussed in the literature since more than fifty years ago.

In this paper we show that there is a continuum of representations which connect together two extreme representations of the pure regression type. This continuum of models can be parametrized in terms of a projection matrix describing the part of the modelled vector which is represented <u>exactly</u> (i.e. with no random modelling error) by the Factor Analysis model. Any procedure of model selection is just a procedure for choosing the "exact part" of the modelled variables. Any choice results in certain modelling errors for the remaining variables, whose variance can be computed explicitly.

KEY WORDS: Factor Analysis, Error In Variables, Stochastic Realization, Splitting Subspaces, Identification.

[+] On leave from Istituto di Elettrotecnica ed Elettronica, Università di Padova, 35100 PADOVA, Italy

1 INTRODUCTION

Given two Hilbert spaces Y_1, Y_2 of zero mean second order real random variables with inner product $\langle \zeta, \eta \rangle = E(\zeta\ \eta)$ we say that a third subspace X is splitting for Y_1, Y_2 if

$$\langle \eta_1 - E^X\eta_1,\ \eta_2 - E^X\eta_2 \rangle = 0 \qquad (1.1)$$

for all random variables $\eta_1 \varepsilon Y_1$, $\eta_2 \varepsilon Y_2$. Here E^X denotes orthogonal projection (conditional expectation in the Gaussian case) onto the subspace X. A splitting subspace X makes Y_1 and Y_2 conditionally uncorrelated (independent in the Gaussian case) given X. Notation: $Y_1 \perp Y_2 \mid X$.

A splitting subspace X is said to be minimal if there are no proper subspaces $X' \subset X$ still satisfying (1.1). A thorough analysis of this concept is presented in [7],[8],[3]. One reason for its usefulness is the fundamental role played in various stochastic modelling problems. The following Proposition gives perhaps the simplest instance of relation between splitting and the construction of models for random phenomena. The proof follows immediately from the definition (1.1).

PROPOSITION 1.1

Let Y_1, Y_2 be the subspaces generated by (the scalar components of) two zero mean random vectors y_1, y_2 of dimensions m_1 and m_2. Let the random vector $x = [x_1, \ldots, x_n]'$ be a basis for a splitting subspace X for Y_1, Y_2. Then y_1, y_2 admit the representation

$$y_1 = H_1x + w_1$$
$$y_2 = H_2x + w_2 \qquad (1.2)$$

where H_i, $i = 1,2$ are $m_i$xn deterministic matrices and the random vectors w_1, x, w_2 are mutually uncorrelated i.e.

$$w_1 \perp x \perp w_2 \qquad (1.3).$$

Viceversa, let y_1, y_2 be generated by the scheme (1.2) with w_1, x, w_2 uncorrelated as in (1.3) Then X = span {x} is splitting for Y_1, Y_2.

Models of the type (1.2) are called Factor Analysis (F. A.) models for the random vector $y = [y_1', y_2']'$, ([6], [11]). The vector x, which we shall always take to be a basis, i.e. with a positive definite variance matrix Exx', is sometimes called the factor and X = span {x} the factor space of the model. Two F.A. models of the type (1.2) for which the factors, say x and $\bar{x}$, span the same splitting subspace will be called equivalent. Observe that from the orthogonality condition (1.3) it follows $w_i = y_i - E^X y_i$, i = 1,2 and hence two F.A. models are equivalent if they have the same "noise" vectors, $w_i = \bar{w}_i$ i = 1,2 and $\bar{H}_i = H_iT$, i = 1,2 for some nonsingular n x n matrix T. There is a one to one correspondence between splitting subspaces X for Y_1, Y_2 and equivalence classes (defined module choice of the factor) of F.A. models.

Proposition 1.1 generalizes in a straightforward way to the case of more than two blocks, when the data are N random vectors y_1, ..., y_N of dimensions m_k, $k = 1, \ldots, N$. In this case conditional orthogonality of $Y_1, \ldots, Y_N$ given X is defined by the condition

$$\langle \eta_i - E^X\eta_i, \eta_j - E^X\eta_j \rangle = 0 \tag{1.4}$$

for all $\eta_i \in Y_i$, $\eta_j \in Y_j$ and all $i \neq j$. By the same argument leading to Proposition 1.1 we could, more generally, state that *every equivalence class (defined modulus basis change* $\bar{x} = Tx$, T n x n *nonsingular)* *of Factor Analysis models,*

$$\begin{aligned} y_1 &= H_1\, x + w_1 \\ \cdots\ &\ \ \cdots \qquad \cdots \\ y_N &= H_N\, x + w_N \end{aligned} \tag{1.5}$$

where $w_1 \perp \cdots \perp w_N \perp x$, *is uniquely attached to a splitting subspace* X *for* $Y_1, \ldots, Y_N$.

In this paper however we shall only consider the case N=2.

F.A. models are potentially very useful devices in Multivariate Statistical Analysis and in Econometrics. Their structure is however very poorly understood. One difficulty with these models is their intrinsic *lack of uniqueness*. Even if we restrict to the class of *minimal models* by requiring X = span {x} to be a minimal splitting subspace, there are in general infinitely many (equivalence classes of) F.A. models describing the same data. This is due to the fact that there are in general infinitely many minimal splitting subspaces for given Y_1, Y_2. Indeed, the two predictor spaces

$$X_1 = E^{Y_1} Y_2, \qquad X_2 = E^{Y_2} Y_1 \tag{1.6}$$

are both minimal splitting. Other minimal splitting subspaces can be constructed by suitably combining X_1 and X_2 ([3], [10]). Note that $X_1 \subset Y_1$, $X_2 \subset Y_2$ and, unless some unlikely degeneracy occurs, X_1 and X_2 are very different objects. Now, if a F.A. model is to be used for identification of real data a preliminary question to solve is which minimal model (actually which equivalence class) should be chosen to fit the given data. Note that the choice of the model (i.e. of the minimal splitting subspace) has to be done a priori since all F.A. models generate the same data y_1, y_2, in particular the same covariance matrix (the same probability distribution in the Gaussian case) and are therefore indistinguishable by looking at sample values of y_1, y_2. The nonuniqueness manifests itself with the presence of "too many" parameters to estimate and has sometimes been called lack of identifiability in the literature [6],[11]. In our opinion this terminology is misleading. In effect identifiability is a concept related to coordinatization i.e. choice of a particular coordinate system to describe a model in a one to one way and is a condition that can always be achieved (at least locally)[5]. It has nothing to do with the (probabilistic) problem of selecting that particular model out of a model class.

The elucidation of some basic properties of F.A. models corresponding to different minimal X's and the suggestion of a possible criterion for the choice in the model class will be the main theme of this note. Due to reasons of space we shall only present and illustrate the main results without supplying proofs. A more complete version of the theory will be found in the forthcoming article [12].

In this paper we shall restrict our discussion to (models corresponding to) minimal splitting subspaces X contained in the data space $Y: = Y_1 \vee Y_2$. There are very good reasons to do so if our F.A. models are to be used for identification of real data. In identification all what is available are sample values of the random vectors y_1, y_2 and this means in particular that we will not be able to distinguish, on the basis of our observations, among factor vectors x having the same conditional expectation given the data y_1, y_2. To have a chance of reconstructing x unambiguously from the data we shall then have to use models in which x is a function of y_1 y_2.

2. A PARAMETRIZATION OF MINIMAL SPLITTING SUBSPACES

Let $Q \in R^{n \times n}$ be a positive definite symmetric matriz. A Q-orthogonal projector Π, is an idempotent nxn real matrix satisfying

$$\Pi \, Q = Q \, \Pi' \tag{2.1}$$

(the prime denotes transposition) or, equivalently,

$$\Pi \, Q \, (I - \Pi)' = 0 \tag{2.2}$$

This notion is simply that of an orthogonal projector in R^n with respect to the inner product $\langle x,y \rangle = x' \, Q^{-1} y$.

Let y_1, y_2 be m_1 and, respectively, m_2-dimensional zero mean random vectors with a nonsingular joint covariance matrix Λ. We shall write Λ in block-partitioned form as

$$\Lambda = E \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} [y_1' \; y_2'] = \begin{bmatrix} \Lambda_1 & \Lambda_{12} \\ \Lambda_{12}' & \Lambda_2 \end{bmatrix} \tag{2.3}$$

and denote by n the rank of the cross covariance Λ_{12}. Notice that $n \leq \min(m_1, m_2)$. Since Λ is strictly positive definite, both matrices

$$Q_1 := \Lambda_1 - \Lambda_{12} \Lambda_2^{-1} \Lambda_{21} \tag{2.4}$$

$$Q_2 := \Lambda_2 - \Lambda_{21} \Lambda_1^{-1} \Lambda_{12} \tag{2.5}$$

are (symmetric and) strictly positive definite. Actually Q_1 and Q_2 are the covariance matrices of the "prediction errors"

$$\tilde{y}_{1|2} := y_1 - E(y_1|y_2) \tag{2.6}$$

$$\tilde{y}_{2|1} := y_2 - E(y_2|y_1) \tag{2.7}$$

Let $\underline{\underline{P}}_1$ be the class of all Q_1-orthogonal projectors $\Pi_1 \in R^{m_1 \times m_1}$, mapping onto a subspace of the range space $\underline{\underline{R}}(\Lambda_{12})$ of Λ_{12}. Dually, let $\underline{\underline{P}}_2$ be the set of all Q_2-orthogonal projectors $\Pi_2 \in R^{m_2 \times m_2}$ mapping onto a subspace of $\underline{\underline{R}}(\Lambda_{21})$.

There are as many Π_1 in $\underline{\underline{P}}_1$ (in a fixed basis) as many subspaces of $\underline{\underline{R}}(\Lambda_{12})$. The next theorem states that this is exactly how many minimal splitting subspaces $X \subset Y$ there are.

THEOREM 2.1 ([12])

There is a bijective mapping between the set $\underline{P}_1$ *and the family of all minimal splitting subspaces* X *contained in* $Y = Y_1 \vee Y_2$. *For each* $\Pi_1 \in \underline{P}_1$ *the components of the* m_1*- dimensional random vector*

$$\hat{y}_1 := \Pi_1 \, y_1 + (I - \Pi_1) \, E \, (y_1 | y_2) \tag{2.8}$$

span a minimal splitting subspace $X \subset Y$ *and each minimal* $X \subset Y$ *is obtained in this way for a unique* $\Pi_1 \in \underline{P}_1$. *Moreover* $\hat{y}_1 = E^X y_1$, X *being the splitting subspace corresponding to* Π_1.

The same statement holds for the class $\underline{P}_2$. *For each* $\Pi_2 \in \underline{P}_2$, *the components of the* m_2*-dimensional random vector*

$$\hat{y}_2 := \Pi_2 y_2 + (I - \Pi_2) \, E \, (y_2 | y_1) \tag{2.9}$$

span a minimal splitting subspace $X \subset Y$ *and vice versa, given any minimal* X *in* Y *there is a unique projector* $\Pi_2 \in \underline{P}_2$ *such that* $\hat{y}_2$ *given by* (2.9) *spans* X. *Moreover* $\hat{y}_2 = E^X y_2$.

□

We would like to comment briefly on the significance of this result.

Pick any $\Pi_1 \in \underline{P}_1$ and $\Pi_2 \in \underline{P}_2$, then from (2.8),(2.9) we can express y_1 and y_2 as

$$y_1 = \hat{y}_1 + w_1 \tag{2.10}$$

$$y_2 = \hat{y}_2 + w_2 \tag{2.11}$$

where the random vectors w_1, w_2 are given by

$$w_1 = (I - \Pi_1)\, \tilde{y}_{1|2}, \qquad w_2 = (I - \Pi_2)\, \tilde{y}_{2|1} \tag{2.12}$$

Note that w_1 and $\hat{y}_1$ are uncorrelated (similarly, w_2 and $\hat{y}_2$ are uncorrelated) because $E(y_1|y_2)$ and $\tilde{y}_{1|2}$ are, $E\,(y_1\ \tilde{y}'_{1|2}) = Q_1$ and Π_1 is Q_1-orthogonal. Thus if $\hat{y}_1$ spans a minimal splitting subspace X, then $w_1 \perp X$ and $\hat{y}_1$ necessarily coincides with the projection (conditional expectation in the Gaussian case) $E^X y_1$ - Exactly the same argument applies to $\hat{y}_2$.

We shall call $\hat{y}_1$ <u>the component of</u> y_1 <u>explained by the minimal splitting subspace</u> X. This component is in turn composed of two parts

-An "exact" part $\Pi_1 y_1$

-A "regression" part $(I - \Pi_1)\ E(y_1|y_2)$

While the "exact" part of y_1 is explained by X with no modelling error (i.e. $\Pi_1 y_1 = E^X\ \Pi_1 y_1$), the regression part, which uses y_2 to model the remaining piece $(I - \Pi_1)\ y_1$, of y_1, can describe $(I - \Pi_1) y_1$ only up to some random modelling misfit (error). Note that the modelling misfit vector is precisely w_1 in (2.12). It is evident that choosing Π_1 (and this can be dome arbitrarily in $\underline{P}_1$) means in essence deciding what part of y_1 will be described <u>exactly</u>, i.e. with no modelling misfit, by the corresponding model. Note that, since rank Π_1 can at most be chosen equal to n = rank Λ_{12}, there are a maximum of n linearly independent linear combinations of the scalar components of y_1 which can be described exactly. We introduce now the following definition.

DEFINITION 2.1

<u>Two projectors</u> $\Pi_1 \in \underline{P}_1$, $\Pi_2 \in \underline{P}_2$ <u>are called "conjugate" if they generate the same splitting subspace</u> X <u>(i.e. if</u> $\hat{y}_1$ <u>and</u> $\hat{y}_2$ <u>span the same</u> X <u>)</u>.

Given $\Pi_1 \varepsilon \underline{P}_1$ there is, by Theorem 2.1, an unique projector $\Pi_2 \varepsilon \underline{P}_2$ conjugate to Π_1. This projector determines a corresponding decomposition of the second random vector y_2,

$$y_2 = \Pi_2 y_2 + (I - \Pi_2) E (y_2 | y_1) + w_2 \tag{2.13}$$

where $\Pi_2 y_2$ is now the part of y_2 which is described exactly by the model. Note that $\Pi_2 y_2$ is uniquely determined once the exact part $\Pi_1 y_1$ has been chosen by selecting Π_1. The main question here is to discover how conjugate projectors are related to each other.

THEOREM 2.2 ([12])

A projector $\Pi_2 \varepsilon \underline{P}_2$ _is conjugate to_ $\Pi_1 \varepsilon \underline{P}_1$ _if and only if it maps_ R^{m_2} _exactly onto the range space of_ $\Lambda_{21} \Lambda_1^{-1} (I - \Pi_1)$. _Dually,_ $\Pi_1 \varepsilon \underline{P}_1$ _is conjugate to_ $\Pi_2 \varepsilon \underline{P}_2$ _if and only if the range space of_ Π_1 _is equal to_ $\underline{R}(\Lambda_{12}\Lambda_2^{-1}(I - \Pi_2))$.

□

This characterization of conjugate projectors provides the rule for computing Π_2 or equivalently, the companion representation of y_2 in (2.13), once Π_1 has been selected. The procedure could in principle be adapted to sample covariance matrices and hence used directly in estimation problems.

Let Π_{2MAX} be the Q_2-orthogonal projection onto the range space of Λ_{21}, or, equivalently, onto the range space of the _regression matrix_

$$A := \Lambda_{21}\Lambda_1^{-1} : R^{m_1} \to R^{m_2} \tag{2.14}$$

Dually, let Π_{1MAX} be the Q_1-orthogonal projector onto the range space of the adjoint regression matrix

$$A^* := \Lambda_{12}\Lambda_2^{-1} : R^{m_2} \to R^{m_1} \tag{2.15}$$

(it is easy to check that A and A* are really adjoint operators with respect to the inner products $\langle x, y\rangle_1 = x'Q_1^{-1} y$ and $\langle x, y\rangle_2 = x'Q_2^{-1} y$ in R^{m_1} and R^{m_2}. The verification follows from the identity

$$Q_1\Lambda_1^{-1}\Lambda_{12} = \Lambda_{12}\Lambda_2^{-1}Q_2 \tag{2.16}$$

Notice that by definition $\underline{P}_1$ and $\underline{P}_2$ consist precisely of all Q_1-orthogonal Π_1 and, respectively, all Q_2-orthogonal Π_2 satisfying

$$\underline{P}_1 : \quad \Pi_{1\ MAX} \geq \Pi_1 \geq 0 \tag{2.17}$$

$$\underline{P}_2 : \quad \Pi_{2\ MAX} \geq \Pi_2 \geq 0 \tag{2.18}$$

where $\geq$ is the partial ordering of projections induced by subspace inclusion. From this it is seen that the decompositions

$$I - \Pi_1 = (I - \Pi_{1\ MAX}) + (\Pi_{1\ MAX} - \Pi_1) \tag{2.19}$$

$$(I - \Pi_2) = (I - \Pi_{2\ MAX}) + (\Pi_{2\ MAX} - \Pi_2) \tag{2.20}$$

are (respectively) Q_1- and Q_2-orthogonal.

REMARK:

The nullspaces of the regression matrix $\Lambda_{21}\Lambda_1^{-1}$ *and of its adjoint* $\Lambda_{12}\Lambda_2^{-1}$ *are the range spaces of* $I - \Pi_{1\ MAX}$ *and of* $I - \Pi_{2\ MAX}$*, respectively*.

For, from $\underline{R}\,(\Pi_{1\ MAX}) = \underline{R}(A^*)$ it follows that the nullspace of A is the range of the complementary projector $I - \Pi_{1\ MAX}$. Thus

$$\Lambda_{21}\Lambda_1^{-1}\ (I - \Pi_{1\ MAX}) = 0, \quad \Lambda_{12}\Lambda_2^{-1}\ (I - \Pi_{2\ MAX}) = 0 \tag{2.21}$$

□

On the orthogonal complement of its nullspace the map A is injective. Hence Theorem 2.2 together with (2.19), (2.21) implies that *the conjugate projector* Π_2 *maps onto a subspace of exactly the same dimension of the range space of* $\Pi_{1\ MAX} - \Pi_1$. In other words,

$$\text{rank } \Pi_2 = \text{rank } (\Pi_{1\ MAX} - \Pi_1) = n - \text{rank } \Pi_1 \tag{2.22}$$

This, in turn, implies that *there is always an n-dimensional subspace of the data space* $Y = \text{span }\{y_1, y_2\}$ *which is described exactly by any minimal Factor Analysis model*. Obviously, since a random variable $\eta \in Y$ is described exactly by X (i.e. $\eta = E^X\eta$) if and only if $\eta \in X$ we see that the exact subspace of the model is nothing else but the minimal splitting subspace X attached to it. Then,

COROLLARY 2.3

The minimal splitting subspaces X Y *are just the linear hulls of the exact parts of* y_1 *and* y_2 *i.e.*

$$X = \text{span }\{\Pi_1\ y_1,\ \Pi_2\ y_2\} \tag{2.23}$$

where $\Pi_1 \varepsilon \underline{P}_1$ *and* $\Pi_2 \varepsilon \underline{P}_2$ *are conjugate projectors.*

The notion of conjugacy has an interesting probabilistic meaning. It is based on the relations

$$E((I - \Pi_1)y_1 \mid y_2) = E((I - \Pi_1)y_1 | \Pi_2 y_2) \tag{2.24}$$

$$E((I - \Pi_2)y_2 \mid y_1) = E((I - \Pi_2)y_2 | \Pi_1 y_1) \tag{2.25}$$

Theorem 2.2 can be restated to say that the projector conjugate to Π_1 is just the minimal $\Pi_2 \varepsilon \underline{P}_2$ for which (2.24) holds. In other words, $\Pi_2\ y_2$ *is a minimal sufficient statistic in* Y_2 *for predicting the "non exact part"* $(I - \Pi_1)y_1$ *of* y_1, *on the basis of the observation* y_2. A similar interpretation can be given to Π_1.

Given any pair of conjugate projectors we can then represent y_1, y_2 by the F.A. model

$$y_1 = \Pi_1 y_1 + E((I - \Pi_1)y_1 | \Pi_2 y_2) + w_1 \tag{2.26}$$

$$y_2 = \Pi_2 y_2 + E((I - \Pi_2)y_2 | \Pi_1 y_1) + w_2 \tag{2.27}$$

where the "noise" terms w_1 and w_2 given by (2.12), are uncorrelated of $(\Pi_1 y_1, \Pi_2 y_2)$. In matrix terms $E((I - \Pi_1)y_1 | \Pi_2 y_2) = (I - \Pi_1)\ E(y_1 | \Pi_2 y_2) = (I - \Pi_1)\ E[E(y_1 | y_2) | \Pi_2 y_2] = (I - \Pi_1)\ \Lambda_{12}\Lambda_2^{-1}\Lambda_1 y_2$ and similarly, $E((I - \Pi_2)y_2 | \Pi_1\ y_1) = (I - \Pi_2)\ \Lambda_{21}\Lambda_1^{-1}\Pi_1 y_1$. If we now bring in the orthogonal decompositions (2.19), (2.20) and recall that $I - \Pi_{1\ MAX}$ annihilates the range space of $\Lambda_{12}\Lambda_2^{-1}$ (similarly $(I - \Pi_{2\ MAX})\ \Lambda_{21}\ \Lambda_1^{-1} = 0$)

we can rewrite the various components of the model (2.26), (2.27) in a more explicit form as

$$\begin{aligned} \hat{y}_1 &= \Pi_1 y_1 + (\Pi_{1\ \mathrm{MAX}} - \Pi_1)\, \Lambda_{12}\, \Lambda_2^{-1}\, \Pi_2\, y_2 \\ \hat{y}_2 &= \Pi_2 y_2 + (\Pi_{2\ \mathrm{MAX}} - \Pi_2)\, \Lambda_{21}\, \Lambda_1^{-1}\, \Pi_1\, y_1 \end{aligned} \tag{2.28}$$

and,

$$\begin{aligned} w_1 &= (\Pi_{1\ \mathrm{MAX}} - \Pi_1)(y_1 - E(y_1 | \Pi_2\, y_2)) + w_{1\ \mathrm{MIN}} \\ w_2 &= (\Pi_{2\ \mathrm{MAX}} - \Pi_2)(y_2 - E(y_2 | \Pi_1\, y_1)) + w_{2\ \mathrm{MIN}} \end{aligned} \tag{2.29}$$

where

$$w_{1\ \mathrm{MIN}} = (I - \Pi_{1\ \mathrm{MAX}})\, y_1, \quad w_{2\ \mathrm{MIN}} = (I - \Pi_{2\ \mathrm{MAX}})\, y_2 \tag{2.30}$$

Observe that the random vector $\hat{y}_1$ takes its values in the n-dimensional subspace $\underline{\underline{R}}(\Lambda_{12}) \subset R^{m_1}$. Its sample values are the Q_1-orthogonal sum of the exact and regression parts of y_1. The noise vector w_1 is the sum of the regression error incurred when estimating $(\Pi_{1\ \mathrm{MAX}} - \Pi_1)y_1$ by y_2 (this is the first summand in (2.29)) plus a term, $w_{1\ \mathrm{MIN}}$, which is the (maximal) component of y_1 uncorrelated of y_2. This last term is of course not dependent on the choice of the splitting subspace. Similar comments apply to $\hat{y}_2$ and w_2.

As a last issue we shall briefly address the question of model choice. In the present context all models represent the data exactly (i.e. equally well!). However, different choices of the splitting subspace X originate different noise vectors w_1, w_2 in (1.2). These noise vectors are infact representation (or prediction) errors of the two random vectors y_1 and y_2. Once X is selected we will be able to predict y_1 and y_2 on the basis of X alone as $\hat{y}_1 = H_1 x$ and $\hat{y}_2 = H_2 x$. In doing so we commit the errors $w_1 = y_1 - \hat{y}_1$ and $w_2 = y_2 - \hat{y}_2$. (These errors are the same irrespective of the choice of basis in X). Hence the covariance matrices of w_1 and w_2, which we shall denote by R_1 and R_2, measure how well a model is doing in predicting y_1 and y_2. From the general expressions (2.12) we compute R_1 and R_2 as

$$R_1 = (I - \Pi_1)\, Q_1, \qquad R_2 = (I - \Pi_2)\, Q_2 \tag{2.31}$$

Note that we have been using the Q-orthogonality property (2.1). In (2.31), Π_1 <u>and</u> Π_2 <u>are conjugate projectors</u>.

It is a consequence of the characterization given in Theorem 2.2 that the ordering ≥ between projectors in $\underline{P}_1$ gets <u>reversed</u> when we pass to the conjugates i.e. if $\Pi_1^1 \geq \Pi_1^2$ in $\underline{P}_1$, then the conjugate projectors satisfy

$$\Pi_2^2 \geq \Pi_2^1$$

(Compare the argument given to derive (2.22)). For example, the conjugate of $\Pi_1 = 0$ is $\Pi_{2\ MAX}$ and the conjugate of $\Pi_{1\ MAX}$ is $\Pi_2 = 0$. This fact implies that a "good" description say of y_1, producing a small prediction error covariance matrix R_1, will automatically have to be paid with a "bad" representation of y_2 which will instead have a big error covariance matrix

As a last issue we shall briefly address the question of model choice. In the present context all models represent the data exactly (i.e. equally well!). However, different choices of the splitting subspace X originate different noise vectors w_1, w_2 in (1.2). These noise vectors are infact representation (or prediction) errors of the two random vectors y_1 and y_2. Once X is selected we will be able to predict y_1 and y_2 on the basis of X alone as $\hat{y}_1 = H_1 x$ and $\hat{y}_2 = H_2 x$. In doing so we commit the errors $w_1 = y_1 - \hat{y}_1$ and $w_2 = y_2 - \hat{y}_2$. (These errors are the same irrespective of the choice of basis in X). Hence the covariance matrices of w_1 and w_2, which we shall denote by R_1 and R_2, measure how well a model is doing in predicting y_1 and y_2. From the general expressions (2.12) we compute R_1 and R_2 as

$$R_1 = (I - \Pi_1)\, Q_1, \qquad R_2 = (I - \Pi_2)\, Q_2 \tag{2.31}$$

Note that we have been using the Q-orthogonality property (2.1). In (2.31), Π_1 <u>and</u> Π_2 <u>are conjugate projectors</u>.

It is a consequence of the characterization given in Theorem 2.2 that the ordering ≥ between projectors in $\underline{P}_1$ gets <u>reversed</u> when we pass to the conjugates i.e. if $\Pi_1^1 \geq \Pi_1^2$ in $\underline{P}_1$, then the conjugate projectors satisfy

$$\Pi_2^2 \geq \Pi_2^1$$

(Compare the argument given to derive (2.22)). For example, the conjugate of $\Pi_1 = 0$ is $\Pi_{2\ MAX}$ and the conjugate of $\Pi_{1\ MAX}$ is $\Pi_2 = 0$. This fact implies that a "good" description say of y_1, producing a small prediction error covariance matrix R_1, will automatically have to be paid with a "bad" representation of y_2 which will instead have a big error covariance matrix

R_2. (Here "big" and "small" are in the sense of the positive semidefinite ordering). We proceed to make this observation precise.

There are <u>minimum</u> values of R_1 and R_2 which are obtained for $\Pi_1 = \Pi_{1\ MAX}$ and $\Pi_2 = \Pi_{2\ MAX}$, respectively,

$$\begin{aligned} R_{1\ MIN} &= (I - \Pi_{1\ MAX})\, Q_1 = (I - \Pi_{1\ MAX})\, \Lambda_1 \\ R_{2\ MIN} &= (I - \Pi_{2\ MAX})\, Q_2 = (I - \Pi_{2\ MAX})\, \Lambda_2 \end{aligned} \qquad (2.32)$$

These correspond to the case where $\hat{y}_1$ and $\hat{y}_2$ have no regression parts. Actually $R_{1\ MIN}$ and $R_{2\ MIN}$ are just the covariance matrices of the "minimal" noise vectors $w_{1\ MIN}$, $w_{2\ MIN}$ in (2.29), (2.30). Clearly we cannot have at the same time $R_1 = R_{1\ MIN}$ and $R_2 = R_{2\ MIN}$. In fact, the case of minimal noise for the y_1 vector corresponds to the model with conjugate projectors $\Pi_1 = \Pi_{1\ MAX}$ and $\Pi_2 = 0$, i.e. to,

$$\begin{aligned} y_1 &= \Pi_{1\ MAX}\, y_1 + w_{1\ MIN} \\ y_2 &= E(y_2|y_1) + \tilde{y}_{2|1} \end{aligned} \qquad (2.33)$$

where w_2 is the full regression error term $w_2 = \tilde{y}_{2|1}$. Observe that the covariance of w_2 in (2.32) is actually the <u>maximum possible</u> value of the error covariance matrix R_2, namely $R_{2\ MAX} = Q_2$. Dually, the best model for describing y_2, i.e. the model with smallest prediction error covariance matrix ($R_2 = R_{2\ MIN}$) for the y_2 variable, is obtained when $\Pi_1 = 0$ and $\Pi_2 = \Pi_{2\ MAX}$

$$y_1 = E(y_1|y_2) + \tilde{y}_{1|2}$$

$$y_2 = \Pi_{2\ MAX}\, y_2 + w_{2\ MIN} \tag{2.34}$$

In this case the prediction error for y_1 is the full regression error $\tilde{y}_{1|2}$ with covariance $R_{1\ MAX} = Q_1$.

In general the prediction errors for y_1 and y_2 are composed of the two terms appearing at the right hand sides of (2.29). Setting

$$\Delta w_1 := w_1 - w_{1\ MIN}, \qquad \overline{\Delta w}_2 = \tilde{y}_{2|1} - w_2 \tag{2.35}$$

where $\tilde{y}_{2|1} = w_{2\ MAX}$ is the maximum variance prediction error for y_2, we compute the covariance matrices

$$\Delta R_1 = E(\Delta w_1, \Delta w_1'), \qquad \overline{\Delta R}_2 = E(\overline{\Delta w}_2\ \overline{\Delta w}_2').$$

From (2.29) it is easy to check that $\Delta w_1 = (\Pi_{1\ MAX} - \Pi_1)\ \tilde{y}_{1|2}$. (By Q_1-orthogonality of the projector, Δw_1 and $w_{1\ MIN}$ are actually uncorrelated). Moreover

$$\Delta R_1 = (\Pi_{1\ MAX} - \Pi_1)\ Q_1 = R_1 - R_{1\ MIN} \tag{2.36}$$

Similarly we find $\overline{\Delta w}_2 = \Pi_2\ \tilde{y}_{2|1}$ (uncorrelated of w_2) and so

$$\overline{\Delta R}_2 = \Pi_2\ Q_2 = R_{2\ MAX} - R_2 \tag{2.37}$$

We may thus conclude that for a general minimal F.A. model the normalized error variances satisfy

$$\begin{aligned}(R_1 - R_{1MIN})Q_1^{-1} &= \Pi_{1MAX} - \Pi_1 \\ (R_{2MAX} - R_2)Q_2^{-1} &= \Pi_2\end{aligned} \tag{2.38}$$

Although there is no simple explicit formula connecting Π_2 to $\Pi_{1MAX} - \Pi_1$, the range spaces of these two projectors are related by a fixed invertible transformation $\tilde{A} : \underline{R}(\Lambda_{12}) \to \underline{R}(\Lambda_{21})$ where $\tilde{A}$ is just the regression operator A of (2.14) restricted to $\underline{R}(\Lambda_{12}) = \underline{R}(\Pi_{1MAX})$. Thus we may say that the two relative errors in (2.38) are, roughly speaking, "proportional" and the extreme situations encountered with the models (2.33) and (2.34) are seen to be just particular instances of a general behavior.

To conclude, there are precisely n scalar variables that any F.A. model is capable of describing exactly. Choosing a model is the same as choosing these variables. If say only $k \leq n$ scalar components of y_1 are chosen to be "exact" (and if this corresponds to an admissible projector Π_1) then the conjugate projector Π_2 will fix the n-k linear combinations of y_2 that will also be described exactly. The resulting model will then describe the vectors y_1 and y_2 with prediction error variances given by (2.38).

REFERENCES

[1] Aigner, D. J., Hsiao, C., Kapteyn, A., Wansbeek, T., Latent Variable Models in Econometrics, in Grilliches, Z. and Jntrilligator, M. D., eds., Handbook of Econometrics, North Holland, Amsterdam, 1974.

[2] Bartholomew, D. J., The Foundations of Factor Analysis, Biometrika, 71, pp. 221-232, 1984.

[3] Finesso, L., Picci, G., Linear Statistical Models and Stochastic Realization Theory in Analysis and Optimization of Systems, Springer Verlag Lecture Notes in Control and Information Sciences, 62, pp. 445-470, 1984.

[4] Gini, F., Sulla interpolazione di una retta quando la variabile indipendente ē affetta da errore, Metron, 1, pp 1-36, 1921.

[5] Kalman, R. E., Identifiability and Modeling in Econometrics, in Developments in Statistics, Vol. 4, P. R. Krishnaiah ed., Ac. Press, N.Y., 1983.

[6] Lawley, D. N., Maxwell, A. E., Factor Analysis as a Statistical Method, 2nd ed. Butterworth, London, 1971.

[7] Lindquist, A., Picci, G., Realization Theory for Multivariable Stationary Gaussian Processes, SIAM J. Control and Optim, 23, pp. 809-857, 1985.

[8] Lindquist, A., Picci, G., Ruckebusch, G., On Minimal Splitting Subspaces and Markovian Representations, Math. Syst. Theory, 12, pp. 271-279, 1979.

[9] Picci, G., Pinzoni, S., Dynamic Factor Analysis Models for Stationary Processes, IMA J. Math. Control & Information, 3, pp. 185-210, 1986.

[10] Van Putten, C., Van Schuppen, J. H., The Weak and Strong Gaussian Probabilistic Realization Problem, J. Multivariate Anal., 13, pp. 118-137, 1983.

[11] Van Schuppen, J. H., Stochastic Realization Problems Motivated by Econometric Modelling, in Modelling, Identification and Robust Control, C. Byrnes and A. Lindquist, eds., North Holland, Amsterdam, 1986.

[12] Picci, G, Parametrization of Factor Analysis Models, Journal of Econometrics, to appear.

Adaptive Pole Assignment by State Feedback.

J.W. Polderman
Centre for Mathematics and Computer Science
P.O Box 4079, 1009 AB Amsterdam, The Netherlands

An algorithm for adaptive pole placement for a restricted class of systems is proposed. The asymptotic properties of the algorithm are analysed by studying the invariant points and the asymptotic active part of the state space. A weak form of self-tuning is derived.

1980 Mathematics Subject Classification: 93C40.
Key Words & Phrases: Adaptive pole-placement, self-tuning, certainty-equivalence.

1. Introduction.

This note is concerned with the problem of adaptive pole placement of deterministic systems without external excitation. We consider a linear plant with only one input and observed state. The problem we then want to study is the asymptotic assignment of the closed-loop poles in a pre-described configuration by means of adaptive state feedback. The proposed algorithm is based on direct estimation of the plant-parameters and the certainty- equivalence principle. Since identification takes place in closed-loop the true system cannot be identified without external excitation. However it will be shown that because of the chosen control objective, closed-loop identification causes no extra difficulties, which is in contrast with adaptive LQ control (see [4]).

Since no external excitation is added, it cannot be expected that the state trajectory will span the whole state-space. Therefore the concept of excitation subspace will be introduced to analyse the proposed algorithm.

This work is motivated by two approaches of adaptive stabilization that appeared in the literature. The first is the model reference adaptive control method (see for instance [5]). The other approach has been presented in a series of papers which culminated in [2]. The first method was developed for systems in input/output form, whereas the second works in state space. In both cases stability results are derived without imposing conditions on exciting signals. Here we make an attempt to derive a weak form of self-tuning.

A shorter version of this paper is [3].

A serious difficulty is caused by the fact that we consider systems in state space form and try to identify the (A,b) parameters. During the estimation procedure all estimates have to be reachable in order to be able to calculate the control law to be applied. This problem has not yet been solved and will be commented upon elsewhere in the paper.

We start with a description of the class of systems under consideration and of the control problem. Next we present our algorithm. We will then formulate our main theorem followed by its proof, which is distributed over several lemmata. We end with some concluding remarks.

Report OS-R86..
Centre for Mathematics and Computer Science
P O Box 4079, 1009 AB Amsterdam, The Netherlands

2. Preliminaries.

Consider the following system:

$$x_{k+1} = Ax_k + bu_k\ , \tag{2.1}$$

where $(A,b) \in E := \{(A,b) \in \mathbf{R}^{n \times n} \times \mathbf{R}^{n \times 1} \mid (A,b) \text{ reachable }\}$.

Let $\Lambda := \{\lambda_1,...,\lambda_n\} \subset \mathbf{C}$ be such that $\lambda \in \Lambda \Rightarrow \bar{\lambda} \in \Lambda$. Define $\sigma \in \mathbf{R}[X]$ by: $\sigma(X) = \prod_{i=1}^{n}(X - \lambda_i)$.

Define $f : E \rightarrow \mathbf{R}^{1 \times n}$ by:

$$f(A,b) := -[0...01]\,[b \vdots bA \vdots .... \vdots bA^{n-1}]^{-1}\sigma(A) \tag{2.2}$$

Then the characteristic polynomial of $A + bf(A,b)$ is exactly σ and moreover since the system is single-input $f(A,b)$ is the only feedback law with that property. (see [6]).

Suppose now that the true value, say (A_0,b_0) of the system-parameters is not known, then the question arises how 'much' we should know about them to be able to control the system as desired. Of course it will be enougb to know $f(A_0,b_0)$, but we will see that this is not the minimum of information we need.

In this paper we want to present an algorithm based on direct estimation of (A_0,b_0) and the certainty equivalence principle. This structure causes certain identification problems (see [4]). In the following theorem the best possible situation for an estimate (A,b) is studied.

Theorem 2.1 Let $(A,b) \in E$ and $\mathcal{V}$ a linear subspace of $\mathbf{R}^{n \times n}$ such that:

i) For all $v \in \mathcal{V}$: $(A_0 + b_0 f(A,b))v \in \mathcal{V}$

ii) For all $v \in \mathcal{V}$: $(A_0 + b_0 f(A,b))v = (A + bf(A,b))v$

Then:

For all $v \in \mathcal{V}$: $f(A,b)v = f(A_0,b_0)v$.

Proof Suppose that $\Lambda \subset \mathbf{R}$ and that $\lambda_i \neq \lambda_j$ for all $i \neq j$. Let $\mathcal{V}$ be one-dimensional. Then $\mathcal{V}$ is generated by an eigenvector v of $(A + bf(A,b))$ corresponding to let's say $\lambda := \lambda_i$. Hence $(A_0 + b_0 f(A,b))v = \lambda v$. Suppose (A_0,b_0) is in standard controllable form. Then $v = [1,\lambda,..,\lambda^{n-1}]^T$. Since λ is an eigenvalue of $(A_0 + b_0 f(A_0,b_0))$, there exists $\tilde{v}$ such that $(A_0 + b_0 f(A_0,b_0))\tilde{v} = \lambda\tilde{v}$. It is easy to see that $v = \mu\tilde{v}$, for some $\mu \neq 0$. Hence $(A_0 + b_0 f(A_0,b_0))v = (A_0 + b_0 f(A,b))v$. Since $b_0 \neq 0$, we conclude that $f(A,b)v = f(A_0,b_0)v$.

If $\dim\mathcal{V} > 1$, then $\mathcal{V}$ has a basis of eigenvectors and the above reasoning gives the result. For general Λ the proof goes along the same lines, but then one has to study several different cases. We skip the details.

Comment. Suppose we have an estimate (A,b) of (A_0,b_0), according to the certainty equivalence principle we will then apply $u_k = f(A,b)x_k$. The resulting closed -loop system is:

$$x_{k+1} = A_0 + b_0 f(A,b))x_k$$

Whereas on the basis of our guess we would predict:

$$\hat{x}_{k+1} = (A + bf(A,b))x_k$$

Suppose now that for all k we have $\hat{x}_{k+1} = x_{k+1}$, this is in some sense the best situation we could have. For once we have an estimate (A,b) with that property, the observed data will not give rise to any update of the parameter estimates. Define $V := \mathrm{span}\{x_k\}$, then it can be checked that V satisfies the conditions of Theorem 2.1 and hence we conclude that for all $v \in V$, $f(A,b)v = f(A_0,b_0)v$. In particular: $f(A,b)x_k = f(A_0,b_0)x_k$, for all k, or otherwise stated the applied input equals the desired input.

Summarizing: In order to control the system (2.1) as desired, it is not necessary to know (A_0,b_0), nor is it necessary to know $f(A_0,b_0)$, we only need to know the action of $f(A_0,b_0)$ on the active part of the state space.

3. Description of the algorithm.

We will introduce the algorithm inductively. Choose the initial guess $(\hat A_0,\hat b_0)\in E$ of (A_0,b_0) arbitrarily. Suppose the k –th guess $(\hat A_k,\hat b_k)$ has been calculated. Then take $u_k=f(\hat A_k,\hat b_k)x_k$. This gives :

$$x_{k+1}=(A_0+b_0f(\hat A_k,\hat b_k))x_k$$

Define

$$\hat G_{k+1}:=\{(A,b)\mid(A+bf(\hat A_k,\hat b_k))x_k=x_{k+1}\}\tag{3.1}$$

$\hat G_{k+1}$ is an affine subvariety of $\mathbf{R}^{n\times n+n\times 1}$. Hence we can take $(\hat A_{k+1},\hat b_{k+1})$ to be the orthogonal projection of $(\hat A_k,\hat b_k)$ on $\hat G_{k+1}$ in $\mathbf{R}^{n\times n+n\times 1}$. This procedure is equivalent to the following recursion for $(\hat A_k,\hat b_k)$:

$$\hat A_{k+1}=\hat A_k+(\|u_k\|^2+\|x_k\|^2)^{-1}(x_{k+1}-\hat x_{k+1})x_k^T\tag{3.2a}$$

$$\hat b_{k+1}=\hat b_k+(\|u_k\|^2+\|x_k\|^2)^{-1}(x_{k+1}-\hat x_{k+1})u_k\tag{3.2b}$$

$$u_k=f(\hat A_k,\hat b_k)x_k\tag{3.2c}$$

$$\hat x_{k+1}=(\hat A_k+\hat b_kf(\hat A_k,\hat b_k))x_k\tag{3.2d}$$

Comment. The algorithm is based on two ideas. The first is concerned with the analysis of the invariant points of the algorithm. From the above description it follows that $(\hat A_{k+1},\hat b_{k+1})=(\hat A_k,\hat b_k)$ if and only if $(\hat A_k,\hat b_k)\in\hat G_{k+1}$. Define $G:=\{(A,b)\mid A+bf(A,b)=A_0+b_0f(A,b)\}$. Then certainly every element of G is an invariant point of 3.2. It follows from Theorem 2.1 that $(A,b)\in G$ implies $f(A,b)=f(A_0,b_0)$. Hence if all the limit points of $\{(\hat A_k,\hat b_k)\}_{k\in\mathbf{N}}$ are in G, then we have achieved our control objective.

The second motivation is the following. Suppose at time k we have the estimate $(\hat A_k,\hat b_k)$ of (A_0,b_0). The certainty equivalence principle tells us to act as if we were sure about (A_0,b_0) and hence we should apply $u_k=f(\hat A_k,\hat b_k)x_k$ to the real system. After having done so we observe the new state x_{k+1}. Now $\hat G_{k+1}$ is exactly the set of those parameters (A,b) that are able to explain the observed data (x_k,x_{k+1},u_k). Since obviously $(A_0,b_0)\in\hat G_{k+1}$, it is natural to choose $(\hat A_{k+1},\hat b_{k+1})$ somewhere in $\hat G_{k+1}$. The reason that we take the orthogonal projection of $(\hat A_k,\hat b_k)$ on $\hat G_{k+1}$ is that as a direct consequence $\|(A_0,b_0)-(\hat A_k,\hat b_k)\|$ converges. The idea of orthogonality was already used in [1], where it was derived from a certain stochastic approximation algorithm. Here we choose it as a starting point rather than as a consequence.

One further remark has to be made. The algorithm 3.2 only makes sense if $(\hat A_k,\hat b_k)$ is reachable for every $k\in\mathbf{N}$. Throughout the paper we will hence make the following assumptions: For all $k\in\mathbf{N}$, $(\hat A_k,\hat b_k)\in E$, and also all limit points of $\{(\hat A_k,\hat b_k)\}_{k\in\mathbf{N}}$ are in E. The first assumption is not really a limitation, for it is not difficult to see that for a generic choice of $(\hat A_0,\hat b_0)\in E$, $(\hat A_k,\hat b_k)\in E$ for all k. The condition on the limit points however is undesirable and should follow as a consequence of the first. This point is still under investigation.

4. Analysis of the algorithm.

The properties of the algorithm will be derived in several steps. We will need some definitions and lemmata before we can draw asymptotic conclusions. First we shall state our main result.

THEOREM 4.1 Consider the (controlled) system (2.1,3.2), there exists a sequence of matrices $\{\Delta_k\}_{k,\mathbf{N}}$, such that:

$$i)\ x_{k+1} = (A_0 + b_0 f(\hat{A}_k, \hat{b}_k))x_k$$
$$= (A_0 + b_0 f(A_0, b_0) + \Delta_k)x_k)$$
$$ii)\ \lim_{k\to\infty} \Delta_k = 0$$

COMMENT. Theorem 4.1 tells us that asymptotically the action of the closed-loop matrix is identical to that of the optimal closed-loop matrix. It should be noticed that we do not claim that the real closed-loop matrix converges to the optimal one, but only as far as the action on the real state-trajectory is concerned. This weaker form of convergence is not surprising if we realise the fact that the estimation procedure only receives information about the action of the real closed-loop matrix on the state-trajectory. We propose the term 'weak self-tuning' for this kind of behaviour. Self-tuning would have implied that $\lim_{k\to\infty} (A_0 + b_0 f(\hat{A}_k, \hat{b}_k) = A_0 + B_0 f(A_0, b_0)$, which we do not claim.

We shall now state two technical lemmata which we will need in the proof of Theorem 4.1.

LEMMA 4.2 Let $K \subset \mathbf{R}^{n\times n}$ be compact and let $\epsilon > 0$. Then there exists $\gamma > 0$ such that for all $A \in K$ and for all $x \in \mathbf{R}^n$ with $\|Ax\| \geqslant \epsilon$ and $x^T x = 1$: $\|Axx^T\| \geqslant \gamma$.

PROOF Suppose the claim is not true. Then there exist $A \in K$ and $x \in \mathbf{R}^n$ with $\|Ax\| \geqslant \epsilon$, $x^T x = 1$ and $\|Axx^T\| = 0$. This implies that $Axx^T = 0$, which means that either Ax or $x^T = 0$, which are both contradictions.

LEMMA 4.3 Let $\{M_k\}_{k,\mathbf{N}}$ be a bounded sequence of matrices in $\mathbf{R}^{n\times n}$, such that $\lim_{k\to\infty} \|M_{k+1} - M_k\| = 0$. Let $x_0 \in \mathbf{R}^n$ be given and define the sequence $\{x_k\}$ by putting: $x_{k+1} = M_k x_k$. Suppose $\lim_{k\to\infty} M_{t_k} = M$, define $\mathcal{X}$ as the linear subspace generated by the limit points of $x^{\bullet}_{l+t_k}$, where l ranges from 0 to infinity. Then $M\mathcal{X} \subset \mathcal{X}$.

PROOF Suppose $x^{\bullet}$ is a limit point of $\{x^{\bullet}_{l+t_k}\}$ for some l. Say $\lim_{k\to\infty} x^{\bullet}_{l+s_k} = x^{\bullet}$, for some subsequence $\{s_k\}$ of $\{t_k\}$. Then:

$$Mx^{\bullet} = \lim_{k\to\infty} M_{l+s_k} x^{\bullet}_{l+t_k} = \lim_{k\to\infty} \frac{1}{\|x_{l+s_k}\|} M_{l+s_k} x_{l+s_k} = \lim_{k\to\infty} \frac{1}{\|x_{l+s_k}\|} x_{l+1+s_k}$$
$$= \lim_{k\to\infty} \frac{\|x_{l+1+s_k}\|}{\|x_{l+s_k}\|} x^{\bullet}_{1+l+s_k} = \lim_{k\to\infty} \frac{\|M_{l+s_k} x_{l+s_k}\|}{\|x_{l+s_k}\|} x^{\bullet}_{1+l+s_k} = \lim_{k\to\infty} \|M_{l+s_k} x^{\bullet}_{l+s_k}\| x^{\bullet}_{1+l+s_k}$$
$$= \|Mx^{\bullet}\| \lim_{k\to\infty} x^{\bullet}_{1+l+s_k}.$$

Hence $Mx^{\bullet} \in \mathcal{X}$. By linearity the result follows.

LEMMA 4.4 $\|(\hat{A}_k, \hat{b}_k) - (A_0, b_0)\|$ is a decreasing sequence, hence it converges to some real constant $R \geqslant 0$.

PROOF This a direct consequence of the orthogonal projection feature which assures that $\|(\hat{A}_k, \hat{b}_k) - (A_0, b_0)\| \geqslant \|(\hat{A}_{k+1}, \hat{b}_{k+1}) - (A_0, b_0)\|$.

Although Lemma 4.4 is very simple not to say trivial, it is the central feature of our algorithm. A direct consequence of 4.4 is that $(\hat{A}_k,\hat{b}_k)$ converges to a sphere with centre (A_0,b_0) and radius R. If $R=0$ then $(\hat{A}_k,\hat{b}_k)\rightarrow(A_0,b_0)$ and we are done. In the sequel we shall hence assume that $R>0$.

DEFINITION 4.5

i) Denote by $\{(\bar{A}_i,\bar{b}_i)\}_{i\in I}$ the set of limit points of $\{(\hat{A}_k,\hat{b}_k)\}_{k\in\mathbf{N}}$. Assume that for every $i\in I$ $\lim_{k\to\infty}(\hat{A}_{t_k^i},\hat{b}_{t_k^i})=(\bar{A}_i,\bar{b}_i)$. Since $(\hat{A}_k,\hat{b}_k)$ cannot make positive jumps bounded from below infinitely often without penetrating the sphere to which it was supposed to converge from the outside, I is either a singleton or an infinite set.

ii) Let $\{x_k\}_{k\in\mathbf{N}}\subset\mathbf{R}^n$ be the state trajectory of the real closed-loop system. Define for every $x\in\mathbf{R}^n$: $x^\bullet:=x/\|x\|$ if $x\neq 0$, and $0^\bullet:=0$. Denote by $\mathcal{X}$ the linear subspace of $\mathbf{R}^n$ generated by the limit points of $\{x_k^\bullet\}$.

iii) Let for every $i\in I$, $\mathcal{X}_i$ be the subspace generated by the limit points of $x^\bullet_{l+t_k^i}$, where l ranges from zero to infinity.

The space $\mathcal{X}$ can be viewed as the excitation subspace of the state space. It reveals the separation between fast and slow convergence/divergence. Since it is easy to see that 3.2 depends only on $x_k^\bullet$ rather then on x_k itself, it will appear that $\mathcal{X}$ will be very helpful in the analysis of the algorithm. It will be supposed that dim $\mathcal{X}>0$, since if dim $\mathcal{X}=0$ then $x_k^\bullet=0$ for $k\geqslant k_0$ for some k_0 and then there is very little to analyse.

$\mathcal{X}_i$ can be interpreted as the excitation space belonging to $(\bar{A}_i,\bar{b}_i)$. The reason that we take the union over all l, is that since $\|(\hat{A}_{k+1},\hat{b}_{k+1})-(\hat{A}_k,\hat{b}_k)\|\rightarrow 0$, $\lim_{k\to\infty}(\hat{A}_{l+t_k^i},\hat{b}_{l+t_k^i})=(\bar{A}_i,\bar{b}_i)$, for all l.

LEMMA 4.6 $\sum_{i\in I}\mathcal{X}_i=\mathcal{X}$.

PROOF It is obvious that the left hand side is contained in the right hand side.
Suppose $x^\bullet$ is a limit point of $\{x_k^\bullet\}$, say, $\lim_{k\to\infty}x^\bullet_{s_k}=x^\bullet$. Let for some $i\in I$, $(\bar{A}_i,\bar{b}_i)$ be a limit point of $(\hat{A}_{s_k},\hat{b}_{s_k})$, then $x^\bullet\in\mathcal{X}_i$. For arbitrary $x\in\mathcal{X}$ the result follows by writing x as a linear combination of limit points.

LEMMA 4.7 $\lim_{k\to\infty}[(\hat{A}_k+\hat{b}_kf(\hat{A}_k,\hat{b}_k))-(A_0+b_0f(\hat{A}_k,\hat{b}_k))]x_k^\bullet=0.$

PROOF Define $M_k:=[(\hat{A}_k+\hat{b}_kf(\hat{A}_k,\hat{b}_k))-(A_0+b_0f(\hat{A}_k,\hat{b}_k))]$. Suppose the claim is not true. Then there exists $\epsilon>0$ and a sequence $\{s_k\}$ such that: $\|M_{s_k}\|\geqslant\epsilon$ for all k. From 3.2a we see that:

$$\|(\hat{A}_{1+s_k},\hat{b}_{1+s_k})-(\hat{A}_{s_k},\hat{b}_{s_k})\|=(\|f(\hat{A}_{s_k},\hat{b}_{s_k})x^\bullet_{s_k}\|^2+\|x^\bullet_{s_k}\|^2)^{-1}\|M_{s_k}x^\bullet_{s_k}x^{\bullet T}_{s_k}\|\geqslant C_1>0.$$

This follows from the facts that $(\hat{A}_k,\hat{b}_k)$ is bounded and reachable, the continuity of f on E, the reachability of $(\bar{A}_i,\bar{b}_i)$ and Lemma 4.2. Now denote $\|(\hat{A}_k,\hat{b}_k)-(A_0,b_0)\|$ by r_k. Choose $\delta>0$ and let k_0 be such that $R\leqslant r_{s_k}\leqslant R+\delta$ for all $k\geqslant k_0$. Using Pythagoras' theorem we see that for all $k\geqslant k_0$:

$$r_{s_k}-r_{1+s_k}\geqslant r_{s_k}-({r_{s_k}}^2-C_1^2)^{1/2}\geqslant R(1-(1-(\frac{C_1}{R+\delta})^2)^{1/2})\geqslant C_2>0.$$

Since r_k is non-increasing we have $r_{s_k}-r_{s_{k+1}}\geqslant C_2$, which yields:

$$r_{s_k}<r_{s_{k_0}}-C_2(k-k_0)\leqslant R+\delta-C_2(k-k_0).$$

Hence there exists k such that $r_{s_k}<R$, which is a contradiction.

LEMMA 4.8 For every $i \in I$ and $x \in \mathcal{X}_i$: $[(\overline{A}_i + \overline{b}_i f(\overline{A}_i, \overline{b}_i)) - (A_0 + b_0 f(\overline{A}_i, \overline{b}_i))]x = 0$.

PROOF Suppose $x^\bullet$ is a limit point of $x^\bullet_{l+t^i_k}$, say $\lim\limits_{k \to \infty} x^\bullet_{l+s^i_k} = x^\bullet$, then from Lemma 4.7 we deduce: $[(\overline{A}_i + \overline{b}_i f(\overline{A}_i, \overline{b}_i)) - (A_0 + b_0 f(\overline{A}_i, \overline{b}_i))]x^\bullet = 0$. For arbitrary $x \in \mathcal{X}$ the statement follows by writing x as a (finite) linear combination of limit points.

COROLLARY 4.9

i) For every $i \in I$: $(A_0 + b_0 f(\overline{A}_i, \overline{b}_i))\mathcal{X}_i \subset \mathcal{X}_i$.

ii) For every $i \in I$: $f(\overline{A}_i, \overline{b}_i)|_{\mathcal{X}_i} = f(A_0, b_0)|_{\mathcal{X}_i}$.

PROOF

i) Take $M_k = A_0 + b_0 f(\hat{A}_k, \hat{b}_k)$ in lemma 4.3. Since $\|(\hat{A}_{k+1}, \hat{b}_{k+1}) - (\hat{A}_k, \hat{b}_k)\| \to 0$, we have by the continuity of f on E that $\|M_{k+1} - M_k\| \to 0$.

ii) From Lemma 4.9 we deduce that $(\overline{A}_i + \overline{b}_i f(\overline{A}_i, \overline{b}_i))|_{\mathcal{X}_i} = (A_0 + b_0 f(\overline{A}_i, \overline{b}_i))|_{\mathcal{X}_i}$. From i) we deduce that $(A_0 + b_0 f(\overline{A}_i, \overline{b}_i))\mathcal{X}_i \subset \mathcal{X}_i$. The result now follows from Theorem 2.1.

THEOREM 4.10

i) $\lim\limits_{k \to \infty} \|(f(\hat{A}_k, \hat{b}_k) - f(A_0, b_0))x^\bullet_k\| = 0$.

ii) $\lim\limits_{k \to \infty} [(A_0 + b_0 f(\hat{A}_k, \hat{b}_k)) - (A_0 + b_0 f(A_0, b_0))]x^\bullet_k = 0$.

PROOF

i) Suppose the claim is not true. Then there exist $\epsilon > 0$ and a subsequence $\{s_k\}$ such that $\|(f(\hat{A}_{s_k}, \hat{b}_{s_k}) - f(A_0, b_0))x^\bullet_{s_k}\| \geqslant \epsilon$, for all k. Choose a subsequence $\{\tilde{s}_k\}$ of $\{s_k\}$ such that $\lim\limits_{k \to \infty} (A_{\tilde{s}_k}, b_{\tilde{s}_k}) = (\overline{A}_i, \overline{b}_i)$ for some $i \in I$ and $\lim\limits_{k \to \infty} x^\bullet_{\tilde{s}_k} = x^\bullet \in \mathcal{X}_i$. Then by Corollary 4.9*ii*: $\lim\limits_{k \to \infty} \|(f(\hat{A}_{\tilde{s}_k}, \hat{b}_{\tilde{s}_k}) - f(A_0, b_0))x^\bullet_{\tilde{s}_k}\| = \|(f(\overline{A}_i, \overline{b}_i) - f(A_0, b_0))x^\bullet\| = 0$, which is a contradiction. The result follows.

ii) This is now trivial.

We will now prove Theorem 4.1:

PROOF of THEOREM 4.1 : Choose $\epsilon > 0$. Denote by S^{n-1} the boundary of the unit sphere in $\mathbb{R}^n$, and define $B(x, \delta) := \{y \in \mathbb{R}^n \mid |x - y| < \delta\}$. Let for every $x^\bullet \in S^{n-1}$, $g_{x^\bullet} \in \mathbb{R}^{1 \times n}$ be such that : $|g_{x^\bullet} x^\bullet| = > 2\epsilon$. Define:

$$O_{x^\bullet} := S^{n-1} \cap B(x^\bullet, \delta)$$

Where $\delta > 0$ (depending on $x^\bullet$) is such that:

$$x \in O_{x^\bullet} \Rightarrow |g_{x^\bullet} x| > \epsilon$$

Then $\{O_{x^\bullet}\}_{x^\bullet \in S^{n-1}}$ forms an open covering of S^{n-1}. Since S^{n-1} is compact we conclude that there exist $x^\bullet_{(1)}, \ldots, x^\bullet_{(p)} \in S^{n-1}$, such that $\{O_{x^\bullet_{(i)}}\}_{i=1,\ldots,p}$ covers S^{n-1}. Define K_i as the closure of $O_{x^\bullet_{(i)}}$. Choose subsequences $\{s^i_k\}$ of $\mathbb{N}$ with the following properties:

$$a)\ \bigcup_{i=1}^{p} \bigcup_{k=0}^{\infty} \{s^i_k\} = \mathbb{N}$$

$$b)\ i \neq j \Rightarrow \{s^i_k\}_{k \in \mathbb{N}} \cap \{s^j_k\}_{k \in \mathbb{N}} = \emptyset$$

$$c)\ \{\text{limit points of } x^\bullet_{s^i_k}\} \subset K_i$$

Define $g_{(i)} := g^\bullet_{s^i_k}$.

Choose $(A_i, b_i) \in E$ such that:

$$A_0 + b_0 f(A_0, b_0) = A_i + b_i g_{(i)}$$

Define:

$$g_k^i := b_i^{\#}(A_0 - A_i + b_0 f(\hat{A}_k, \hat{b}_k))$$

Where # denotes left inverse. $b_i^{\#}$ should be chosen such that $b_i^{\#} b_0 \neq 0$. Then:

$$f(\hat{A}_k, \hat{b}_k) = \frac{1}{b_i^{\#} b_0}(g_k^i + b_i^{\#}(A_i - A_0))$$

Then for every $i \in \{1,..,p\}$ we have by Theorem 4.10 that:

$$\lim_{k \to \infty} \|(g_{k_i} - g^i)x_k^{\bullet}\| = \lim_{k \to \infty} \|[b_i^{\#}(A_0 - A_i + b_0 f(\hat{A}_k, \hat{b}_k) - (A_0 - A_i + b_0 f(A_0, b_0))]x_k^{\bullet}\|$$

$$= \lim_{k \to \infty} \|b_i^{\#} b_0 (f(\hat{A}_k, \hat{b}_k) - f(A_0, b_0))x_k^{\bullet}\| = 0.$$

Since by construction $|g_i x_{s_k^i}^{\bullet}| > \epsilon$, for k sufficiently large, we conclude that:

$$\lim_{k \to \infty} \left| \frac{g_{s_k^i}^i}{g^i x_{s_k^i}} - 1 \right| = \lim_{k \to \infty} \left| \frac{(g_{s_k^i}^i - g^i)x_{s_k^i}^{\bullet}}{g^i x_{s_k^i}^{\bullet}} \right| = 0.$$

Define:

$$\alpha_{s_k^i}^i := \frac{g_{s_k^i}^i x_{s_k^i}}{g^i x_{s_k^i}} \quad \text{then:} \quad \lim_{k \to \infty} \alpha_{s_k^i}^i = 1.$$

Hence:

$$\begin{aligned} f(\hat{A}_{s_k^i}, \hat{b}_{s_k^i})x_{s_k^i} &= \frac{1}{b_i^{\#} b_0}(g_{s_k^i}^i + b_i^{\#}(A_i - A_0))x_{s_k^i} \\ &= \frac{1}{b_i^{\#} b_0}(\alpha_{s_k^i}^i g^i + b_i^{\#}(A_i - A_0))x_{s_k^i} \\ &= \frac{1}{b_i^{\#} b_0}(\alpha_{s_k^i}^i b_i^{\#}(A_0 + b_0 f(A_0, b_0) - A_i) + b_i^{\#}(A_i - A_0))x_{s_k^i} \\ &= \frac{1}{b_i^{\#} b_0}((\alpha_{s_k^i}^i - 1)(b_i^{\#}[A_0 - A_i + b_0 f(A_0, b_0)] + b_i^{\bullet} b_0 f(A_0, b_0))x_{s_k^i} \end{aligned}$$

Hence define:

$$\Delta_{s_k^i} := b_0 b_0^{\#}(1 - \alpha_{s_k^i}^i)(A_0 - A_i) + (1 - \alpha_{s_k^i}^i)f(A_0, b_0)$$

Because of the properties a,b of the sequences $\{s_k^i\}$, Δ_k is now well defined for every k. Since

$$\lim_{k \to \infty} \Delta_{s_k^i} = 0, \text{ for } i \in \{1,..,p\} \text{we also have:}$$

$$\lim_{k \to \infty} \Delta_k = 0$$

Moreover:

$$x_{k+1} = (A_0 + b_0 f(A_0, b_0) + \Delta_k)x_k$$

This completes the proof.

COROLLARY 4.11 For all $x \in X$ we have:

$$(A_0 + b_0 f(A_0, b_0))x \in \mathcal{X}$$

PROOF This follows immediately from Theorem 4.1 and Lemma 4.3.

Note that the above results are valid whether or not Λ is contained in the unit disk. But of course for stability of the closed-loop system it is needed that Λ is contained in the unit disk.

The theory as presented does not exclude the possibility that $(\hat{A}_k, \hat{b}_k)$ or even $f(\hat{A}_k, \hat{b}_k)$ does not converge. We have only derived results about their limit points. Indeed it could happen that $(\hat{A}_k, \hat{b}_k)$ keeps drifting along a subset of the sphere to which it converges. However this drifting behaviour requires very rare properties of the sequence of estimates. For if it moves too fast it enters the sphere and if moves too slowly it converges. But the question of convergence versus eternal drifting remains relatively unimportant considering Theorem 4.1.

SIMULATIONS. Extensive simulations have been done for low order systems ($n \leqslant 6$). As could be expected convergence gets slower as n increases. Problems with the reachability of limit points have not been observed and hence it can be expected that the imposed condition is superfluous.

4. CONCLUSIONS.

An algorithm has been proposed and analysed for adaptive pole placement. A weak form of self-tuning has been derived under the reachability condition on the limit points of the estimates. In a forthcoming paper the presented ideas will be applied to a more realistic class of systems, namely SISO systems with unobserved states. There we will also investigate the state trajectory of the controlled system.

REFERENCES

1. A. BECKER, P.R. KUMAR, C.Z. WEI (1985). Adaptive control with the stochastic approximation algorithm: geometry and convergence. *IEEE Trans. Aut. Contr. AC-30*, 330-338.
2. B. MARTENSSON (1985). The order of any stabilizing regulator is sufficient a priori information for adaptive stabilization. *Systems & Control Letters 6*, 87-91
3. J.W. POLDERMAN (1986). Adaptive Pole Assignment by State Feedback. *Proceedings of the 25th IEEE Conference on Decision and Control*, Athens, Greece.
4. J.W. POLDERMAN (1985). *On the necessity of identifying the true system in adaptive LQ control*, to appear in *Systems and Control Letters.*
5. S.S. SASTRY (1984). Model-Reference Adaptive Control-Stability,Parameter Convergence, and Robustness. *IMA Journal of Mathematical Control & Information 1*, 27-66.
6. W.M. WONHAM (1979). *Linear Multivariable Control: a Geometric Approach.* Springer, New York.

QUALITATIVE THEORY AND COMPUTATIONAL ASPECTS OF LARGE SCALE DYNAMIC NONLINEAR SYSTEMS - A REVIEW AND SOME OPEN QUESTIONS.

by T.Roska
Computer and Automation Institute, Hungarian Academy of Sciences
H-1014 Budapest Uri-u 49

Abstract

Deterministic nonlinear large scale dynamical systems are considered as the interconnection of subsystems described by n-ports having nonlinear memoryless and/or memory-type characteristics.

Statements and conditions are given in terms of interconnection and subsystem characteristics. The particular questions considered, mostly in the time domain, are :

the uniqueness of the solution,

the qualitative properties of the solution,

the uniqueness of the computed solution,

the convergence of the computations,

the complexity of the computations in terms of the complexity of the systems and the computing machines and

the choice of the time step in large systems.

Furthermore, the role of the choice of the state variables, the role of the qualitative properties of the subsystem characteristics, the choice of the nonlinear models in large systems and the controversial role of asymptotic results in guessing the complexity of computations are discussed.

1 Introduction

The main question to be answered is as follows. Given a large scale lumped dynamical system i.e. the number of the subsystems are large (e.g. 1000 or greater) and the interconnection is simpler (linear, memoryless and sparse) than the subsystems (they are nonlinear and dynamic). Under what conditions are the time domain results of the analysis qualitatively correct and within a prescribed accuracy with a reasonable computational complexity.

The various aspects of the dynamical circuits, networks and systems have been considered recently in review papers (e.g. (1,2)), in several books (e.g. (3,4)) and research papers are published continuously in the field.

The specific aspects of the investigations of the present paper are as follows:

(i) non-asymptotic results are preferred,

(ii) the complexity of the computations are considered also in terms of the complexity of the computing devices and

(iii) the limits of the results are emphasized in view of the practical analysis of the large scale dynamical systems.

These questions seem to be important because , in spite of the many advances,

(i) asymptotic results are frequently misleading in actual computations

(ii) the possibilities due to the rapid and continuous development of the computing devices (both in speed and complexity) are partly neglected or overestimated and

(iii) to improve the computational efficiency nonconvergent and inaccurate numerical methods are used sometimes.

Besides reviewing the relevant results and emphasizing some open questions it is shown that

- some recently introduced passivity properties and the algorithmic passivity are useful concepts assuring the qualitatively correct computations,

- well known, frequently used algorithms for numerical integration have inherent defects as the number of variables becomes large,

- a definite limit of the complexity of the systems to be reasonable analysed are given in terms of the complexity of the computing devices,

- a memory type discrete time realization of lumped dynamic nonlinear systems is proposed and

- it is pointed out that, except simple cases, the minimal complexity digital simulators can not be found.

2 The uniqueness of the solution

Before any analysis or computations the prerequisite is that the model of the system should be causal (outputs are unique and are zero until the inputs are zero e.g.at t=0).In case of lumped systems this means the uniqueness of the solution in the time domain starting at a unique initial condition. Models of dynamical systems do not posses always a unique solution, they are not always causal.

The basic result on this topic (5) asserts that as far as the elements (including the interconnective ones) are linear and passive (they could be lumped and distributed) the interconnected network will be causal.If there are linear active elements in the network too, the key theorems and the various conditions checking causality have also been given (6,7).In these results a crucial point is that t=0 should be an essential point of all non-zero principal minors of the convolution operator and a checking condition has been given in terms of a multivariable real rational matrix.

In case of nonlinear networks and systems such complete results are not known. If the input-output operator is known then it has been shown (8) that a type of local passivity ensures causality. However , while most of the elements of the physical systems are globally pasive only a few of them are locally passive.

In case of lumped nonlinear systems it is not true that global passivity implies the uniqueness of the time domain solution on the other hand under reasonable conditions local passivity of the elements and strict passivity of the linear interconnections does it (9). If the nonlinear state equation has the Lipschitz -.property (L-property) then according to the well known sufficient condition uniqueness is guaranteed. The problem is that the L-property of the subsystems does not imply the L-property of the interconnected system and , on the other hand, simple non-L systems have unique solution. In (9) it has been shown that under reasonable assumptions if there are unbounded elements in the diagonal entries of the Jacobian then, if they are negative, the time domain solution will be unique. The generalization of this result to the case of interconnected subsystems can be found in (10).

In case of systems with variables having positive values only (e.g.prices, commodities,etc.) several results were also published (2).

An important question is the representation invariance of the uniqueness. In case of nonlinear circuits it is true that if the solution is unique(in the sense of local solvability) in the canonical representation using the charge and flux variables of the lossless element then it is unique in all other representations and if it is not unique in the canonical representation then it is not possible to find another representation having uniqueness(35).

3 The qualitative properties of the exact time domain solution

Without going into the details of the vast amount of literature only the main questions relating to our specific goal are considered. Namely, we try to compute the solution of the system in a finite time domain and the question is how long to integrate for getting a complete information about the system.

Hence, supposed there exists a unique solution, the presence of impasse points and the finite escape time should be excluded (1) and the bounded input bounded output (BIBO) stability is required. Next, logically, there are three possibilities:(i) the solution tends to a finite state vector, (ii) the solution is (almost) periodic or (iii) the solution is chaotic. The latter case is far not being a result of complicated systems. On the contrary, the most simple population model represented by the simple nonlinear difference equation (12) or simple electronic circuits (11) e.g.a 5 element RLC circuit containing only one nonlinear element (piecewise linear) result chaotic solutions.

The problem is that if the qualitative properties of the solution are not known the relevant finite time domain can not be determined. To determine these properties in case of the large scale systems only those conditions are useful which can be algorithmically evaluated in terms of the subsystem characteristics and the interconnection properties. The general mathematical conditions are often useless, some specific properties based on the very nature of the subsystems , on the other hand, could be useful. Such an example is the monotone, isotone and antitone characteristics of the mappings (13). These properties of the Jacobian of a dynamic system state equation are called cooperative or competitive(14). It turned out that in the most different areas like e.g.in medicine and biology (the compartmental system models (e.g.16)), in economic system models(15), in electronic circuit models (17), etc. these properties of the mappings of the relevant models are derived from the nature of the subsystems.

Exploiting these properties leads to strong and algorithmically testable conditions (e.g.15-17). On the other hand, despite the many nice results of system theory it seems that without the relevant knowledge of the area of application i.e. without realizing the essence of the nature and the generic properties of the objects strong results can not be obtained.

4 Uniqueness of the computed solution

At a first glance it seems that if the system (model) has a unique solution in the time domain then the computed solution (e.g. by numerical integration) will also be unique if the computations converge. This is true for open type numerical integration formulas. Unfortunately, however, this is not the case for the most important closed type integration formulas.It depends very much on the structure of the state equations and on the characteristics of the subsystems. Many results in the literature on numerical integration refer mainly to cases where the time step h goes to zero or the number of steps become very large. Hier, however, the most important case is when h is finite and sometimes as large as possible. For a practically important class of nonlinear networks conditions have been given to ensure the uniqueness of the solution of the multistep implicite (closed type) integration formula (18). Furthermore, quite surprisingly, it turned out (19) that in case of nonlinearities having negative slope it could happen that even for finite but arbitrarily small time steps the solution will not be unique. Hence, only an explicite (open type) integration formula can be used. The conditions of the uniqueness in terms of the interconnection and the element characteristics has also been given (19).

Summarizing the qualitative conditions discussed in sections 2,3 and 4 we are in a position to define the notion of a "well posed circuit or system analysis problem"(22). Namely, the following conditions are to be satisfied:

(i) there exists at least one locally stable initial condition,

(ii) the solution exists, it is bounded and unique in any finite time domain and

(iii) the computed solution is unique (if the computational process is convergent).

Unfortunately, in case of many systems having subsystems of well posed system analysis problem the interconnected system fails to have this property. For a fairly broad class of networks having passive linear interconnections and eventually passive memoryless elements the conditions of well posedness have been given (22).

5 The convergence of the computations

Suppose, given a well posed system analysis problem. The next question is: does the iterative algorithm for finding the solution converge (and stable)? Basically two types of problems are considered. The stability of the integration formula and the convergence of the algorithm used for solving the system of nonlinear algebraic equations. Furthermore, sometimes the former problem is inherently interconnected with the latter. This has been the famous case with a stiffly stable numerical algorithm where the implicite formula has been degraded by the predictor corrector iteration algorithm (23,pp.516). Since 1968 the importance of the implicite integration formula (23,24) has benn fully acknowledged and special attention is devoted for the stable implementation of it. Another important way of investigating the convergence and stability of the integration process is by the use of the notion of algorithmic passivity (21). Its limited applicability in circuit analysis can be generalized as follows.

Passive systems have a strong interconnection invariant property. If an algorithm is designed in such a way that first it is applied for the subsystems and next the algorithms are interconnected then if the subsystem algorithms are passive the whole algorithm inherits this property. The passivity of an algorithm can be defined either by the circuit equivalent or by applying the direct scalar product passivity condition.

General conditions for checking the convergence of iteration schemes are numerous (e.g.13,20). The problem using these are twofolds. First, they are sufficient conditions only, secondly, in case of large systems they are very time consuming. For large systems instead of the Newton type process relaxation type algorithms are frequently preferred (e.g.Gauss-Seidel, Gauss-Jacobi etc.) despite the fact of the slower convergence. A temptation in large systems is that only one or two relaxation sweeps are carried out. Therefore the nice properties concerning stability and accuracy are no longer hold. Hence, even standard, famous programs are not working always correctly (25,26). In case of a general partitioned relaxation process the conditions of convergence have been found (26). It turned out again that certain consistency conditions and passivity constraints (inherently in the subsystems, e.g. in some MOS device models) could play a central role.

In case when the subsystems have certain isotone and antitone (13) characteristics and the interconnection is simple then the convergence and stability properties of well known relaxation procedures have been proved (15,17). It turned out that certain class of subsystems (n-ports) which are between the local and global passivity properties play a crucial role.

6 The complexity of the computations

The complexity and the speed of computing machines increasing spectacularly and due to the scaling down process in electronic devices this development proceeds further (27). New questions of designing information processing circuits and systems arise (28). The joint consideration of areas of information theory, physics (e.g. thermodynamics) as well as the circuit and system theory (29) promise new dimensions of understanding the highly complex systems including the computing machines. Due to the very high complexity of these machines(e.g. one million elements per device being the building blocks of the computing machines) the representation, the simulation algorithm, the design of a hardware simulator and the electronic realization of analgorithm or a dynamical system are becoming eventually the same problem, the four areas are inherently coupled.

What the simulation of large scale dynamic circuits and systems is concerned two conflicting tendencies are competing: the complexity and speed of the simulators increase, however, at the same time the complexity of the systems to be simulated increase too. What is the balance? Based on the above ideas some partial results were published (30) which indicate that if the rate of increase of the complexity of the simulator does not exceeds the rate of increase of the complexity of the circuit or system to be simulated then the simulation comlexity will not decrease (even when taking into account the speed increase of the computing devices due to the scaling down effect). More precisely, in the line of these investigations the following statement can be proved (30,pp.459).

Consider a dynamic system composed of K subsystems of identical structures. Suppose a conceptual digital simulator (CDS) is used having a combined (in some sense optimal) use of the time parallel, time series (pipelined) and time iterative mode of operation including the memory elements for realizing the nonlinear I/O operators. If the complexity of the simulator (proportional to the gate count or the relative area of the comuting machine) is increasing as fast as K then the simulation complexity tc (measured in basic operation steps or relative time) is increasing as follows;

$$tc= ko + kl\ w + k2\ w^2$$

where w is the bandwidth of the interconnection matrix and ko, kl, k2 are constants (independent of K).

A natural question or objection concerning of this reasoning is that how can the minimal complexity of the simulator be determined. The answer is that, in principle, the minimal complexity of the simulator can not be determined. The reasoning uses a basic result of complexity theory (34,Theorem 1) and the algorithms of the subsystems are considered as partial recursions.

The complexity of the realization of a nonlinear operator (the number of memory cells) depends on the total sum of the input and output bits. In case of multivariable nonlinear operators this can be greatly reduced by using the nonlinear approximation theorem due to Kolmogorov (36) applied also in nonlinear synthesis for device modeling (37). According to this approximation any multivariable function can be approximated by a finite number (not greater than $2n^2$) of one variable functions. Our new method is that these approximating diagonal mappings are realized by single input single output memories, additions are carried out directly and the chain functions are realized by cascading the memories. Using this idea the memory type realization of any discrete time nonlinear dynamical systems can be carried out.

7 The choice of the time step in large systems

Practical experiences show that increasing the number of variables in solving large scale dynamic systems the time step predicted by the accuracy formula (e.g. (33) p. 497) is shrinking unnecessarily. The well known equation for a k-th order multistep formula is (see e.g.(33))

$$emax = h^{k} \, abs(\, ck \; x^{(k+1)})$$

where emax = Emax/T; Emax being the prescribed maximum truncation error in the time interval (to,to+T); ck is a constant of the integration formula; the (k+1)-th derivative is defined somewhere within the time step. Considering the simple case of a cascade connection of K identical subsystems of order 1 (the same reasoning can be applied for any other order) having about the same (k+1)-th derivative we get the time steps h1 and hK for the case of systems containing one and K subsystems respectively:

$$h1 = \sqrt[k]{emax/ck} \; / \sqrt[2k]{(x1^{(k+1)})^{2}}$$

$$hK \simeq h1 \; / \sqrt[2k]{K}$$

(xi being the i-th element of the vector x).
Hence in case of a backward Euler formula (k=1) it means that the number of time steps increases unnecessarily by sqrt(K).

Acknowledgement

The constant help of and the important discussions with Dr. A. Csurgay is gratefully acknowledged.

References

(1) L.O.Chua, "Dynamic nonlinear networks:state-of-the-art", IEEE Trans Circuits and Systems, Vol.CAS-27,pp.1059-1087 (1980)

(2) I.W.Sandberg, "A perspective on system theory",ibid.,Vol.CAS-31, pp.88-103 (1984)

(3) C.A.Desoer and M.Vidyasagar, Feedback systems: input-output properties, New York:Academic Press, 1975

(4) A.N.Michel and R.K.Miller, Qualitative analysis of large scale dynamical systems, New York:Academic Press,1977

(5) D.C.Youla, L.J.Castriota and H.J.Carlin, "Bounded real scattering matrices and the foundations of linear passive network theory", IRE Trans. Circuit Theory, Vol.CT-6,pp.102-124 (1959)

(6) A.Csurgay and D.C.Youla ,"On the postulational approach to active networks", Polytechnic Inst.Brooklyn, Memo.PIB-MRI-1384-67 (1967)

(7) A.Csurgay, "Multivariable realizability criteria", Proc. Fourth Coll. Microwave Comm.,pp.CT-6/1-9,Budapest:Akadémiai Kiadó, 1970

(8) I.W.Sandberg, "Conditions for the causality of nonlinear operators defined on a function space", Quart.Appl.Math., Vol.XXIII,pp.87-91 (1965)

(9) T.Roska, "On the uniqueness of solutions of nonlinear dynamic networks and systems", IEEE Trans.Circuits and Systems,Vol.CAS-25,pp.161-169 (1978)

(10) T.Roska, "On some qualitative properties of large scale nonlinear circuits and systems",Proc. IEEE ISCAS-82,pp.1062-1065 (1982)

(11) T. Matsumoto, L.O.Chua and M.Komuro,"The double scroll bifurcations",Int. J.Circuit Theory and its Appl.,Vol.14,pp.117-146 (1986)

(12) R.M.May,"Simple mathematical models with very complicated dynamics", Nature, Vol.261, pp.459-467 (1976)

(13) J.M.Ortega and W.C.Rheinboldt, Iterative solution of nonlinear equations in several variables, New York: Academic Press, 1970

(14) M.W.Hirsch, "Systems of differential equations which are competitive or cooperative I.: Limit sets",SIAM J.Math. Anal., Vol.15,pp.167-179 (1982)

(15) I.W.Sandberg, "A criterion for the global stability of a price adjustment process", J.Economic Theory, Vol.19,pp.192-199 (1978)

(16) Y.Ohta, "Stability criteria for off-diagonally monotone nonlinear dynamic systems",Proc.IEEE ISCAS-79,pp.404-407 (1979)

(17) T.Roska "A possibility of filling the gap between local and global passivity of nonlinear networks and some of its consequences", Int. J. Circuit Theory and its Appl.,Vol.9,pp.393-399 (1981)

(18) I.W.Sandberg, "Theorems of the analysis of nonlinear transistor networks", BSTJ,Vol.49, pp.95-114 (1970)

(19) T.Roska and J.Klimó,"On the solvability of DC equations and the implicite integration formula",Int.J.Circuit Theory and its Appl.,Vol.1,pp.273-280 (1973)

(20) I.W.Sandberg, "Diffeomorphisms and Newton direction algo-rithms", BSTJ,Vol.59,pp.1721-1733 (1980)

(21) R.Rohrer and H.Nosraty,"Passivity and stability of single step integration algorithms",Proc.IEEE ISCAS-80,pp.894-896

(22) T.Roska,"The limits of modeling of nonlinear circuits",IEEE Trans.Circuits and Systems,Vol.CAS-28,pp.212-217(1980)

(23) I.W.Sandberg and H.Shichman, "Numerical integration of systems of stiff nonlinear differential equations", BSTJ, Vol.47,pp.511-527 (1968)

(24) C.W.Gier,"The control of parameters in the automatic integration of ordinary differential equations",Int.Rep.757, Dep.Comp.Sci. Univ.Illinois, Urbana , 1968

(25) G.DeMicheli and A.L.Sangiovanni-Vincentelli, "Characterization of integration algorithms for the timing analysis of MOS VLSI circuits", Int.J.Circuit Theory and its Appl.,Vol.10, pp.299-309 (1982)

(26) E.Lelarasmee, A.E.Ruehli and A.L.Sangiovanni-Vincentelli, "The waveform relaxation method for time-domain analysis of large-scale integrated circuits", IEEE Trans. CAD of Integrated Circuits and Systems, Vol.CAD-1, pp.131-145 (1982)

(27) C.Mead and L.Conway, Introduction to VLSI systems, London: Addison Wesley, 1980

(28) O.Wing, "The VLSI-theoretic challenge", IEEE ISCAS-82,Suppl.

(29) A.Csurgay, "Fundamental limits in large scale circuit modelling", Proc. ECCTD-83, pp.454-456

(30) T.Roska, "Complexity of digital simulators used for the analysis of large scale circuit dynamics", Proc. ECCTD-83, pp.457-459

(31) H.T.Kung and C.E.Leierson, "Algorithms for VLSI processor arrays", Section 8.3 in Reference 27

(32) G.D.Hachtel and A.L.Sangiovanni-Vincentelli,· "A survey of third generation simulation techniques", Proc.IEEE, Vol.69, pp.1264-1280 (1981)

(33) L.O.Chua and P-M.Lin, Computer aided analysis of electronic circuits, Englewood Cliffs:Prentice Hall, 1975

(34) G.J.Chaitin, "Information-theoretic computational complexity", IEEE Trans. Information Theory, Vol.IT-20, pp.10-15

(35) T.Matsumoto, L.O.Chua, H.Kawakami and S.Ichiraku, "Geometric properties of dynamic nonlinear networks: transversality, local solvability and eventual passivity", IEEE Trans. Circuits and Systems, Vol.CAS-28, pp.406-428 (1981)

(36) A.N.Kolmogorov,"On the representation of continuous functions of several variables by superposition of continuous functions of one variable and additions" Dokl.Akad.Nauk SSSR, Vol.114, pp.953-956 (1957)

(37) L.O.Chua, "Device modeling via basic nonlinear circuit elements", IEEE Trans Circuits and Systems, Vol.CAS-27, pp.1014-1044 (1980)

SEQUENTIAL AND STABLE METHODS FOR THE SOLUTION OF MASS RECOVERY PROBLEMS (ESTIMATION OF THE SPECTRUM AND OF THE IMPENDANCE FUNCTION)

Gy. Sonnevend
Dept. of Numerical Analysis
Inst. of Math., Eötvös University
1088, Budapest, Muzeum krt. 6-8. Hungary

Introduction

The aim of the present paper is to provide some new tools and methods for the theory of (recursive) identification of a class of linear, causal infinite dimensional input-output systems, where the unknown of the identification problem is (or can be uniquely associated to) a nonnegative mass distribution. Important special cases to be studied are the following (closely related) problems

1) estimation of the spectrum of a stationary (gaussian) stochastic process

2) identification of coefficients in onedimensional hyperbolic systems (describing waves in horizontally layered media,whose impendence or reflectance function is to be recovered).

As it is well known,see e.g. [14], associated to 2, is a coefficient (potential) recovery problem for a corresponding Sturm-Liouville (Schrödinger) equation, in the latter problems however more smoothness needs to be assumed for the impendance function. We are interested in assuming no smoothness on the impendance functions except its positivity and bounded variation (so that two positive measures correspond to it); the connection beetween the smoothness properties of the impendance function and those of the corresponding spectral function (measure) are not yet understood, see [23].

In a more abstract level problems 1) and 2) are about the identification of self adjoint input-output systems i.e. those which are realizable by state space triples (A,b,c) where $A=A^*:H\to H$ (or $AA^* = I$, i.e. A unitary), as usual we denote conjugation for operators by $*$; $b=c$, (for simplicity we restrict ourselves to the scalar case, i.e. that of single input - single output systems).

The association between mass distributions and self adjoint or unitary operators is via the spectral functions of the latters

$$\mu(S) = <E_S b, b>, \qquad A = \int_S s\, d(E_s),$$

where E_s, $s \in R^1$ are the spectral projectors in the spectral decomposition of A (we shall assume in the sequel that all monoton nondecreasing functions μ are defined to be continuous from the right).

The motivation for proposing (thus expecting superior performance from) sequential identification, i.e. mass recovery methods - contrasted to passive or open loop procedures like the Levinson (fast Cholesky) algorithm or "layer peeling", downward continuation methods - comes from the intuitive idea that "measurements" (on the available data or within the given experimental setup) should be more concentrated, where the unknown measure is more concentrated; this should provide not only a more exact recovery (for many, natural definitions of distance between measures) but also assure a more stable recovery, i.e. one in which the recovery errors due to measurement errors are kept smaller. Precize elaboration and justification of this "expectation" will be given below partly based on earlier results(experience) concerning analogous problems, see [18] - [22].

In (deterministic) moment problems concerning mass distributions $d\mu(s)$ on a set S we (potentially) have the data $\{\tilde{c}(t),\ t \in T_M\}$, where $T_M \subseteq T$, the set of all possible measurements (nodes) and

$$c(t) = \int_S K(t,s)\mu(ds),\ t \in T_M,\ |c(t) - \tilde{c}(t)| \le \varepsilon_0 d(t), \qquad (1.1)$$

from which μ is to be recovered. Here K(.,.) is a known continuous Kernel function defined on the product set T×S and d(.) is a known positive function ε_0 is a known (measurement error) scaling parameter. In many cases only N values $c_i = \tilde{c}(t_i)$, $i=1,\dots,N$ can be measured, thus $T_M := (t_1,\dots,t_N)$, since each measurement $t \to \tilde{c}(t)$ is a costly operation (in the case of cheap measurement we usually have $T_M=T$). In the above identification problems these may correspond to measuring (or evaluating) the value of the transfer or impendance function at specific complex numbers ("frequencies") $z \Longleftrightarrow t$, inside the unit circle or the left half plane for discrete resp. continuous time systems.

In the problems of spectrum estimation one finite length realization of the process is used to estimate the values of the "positive real" (impendance) function associated to the spectral measure, thus we always have a nonzero function d(.) and ε_0 in (1.1).

There are two main problems associated to the recovery based on the information (1.1).

The first problem is to find one "nice" solution $\bar{\mu}=\bar{\mu}(T_M)$ of (1.1).

In Section 2 we show that the solution $\bar{\mu}$ (called "the analytic centre" of (1.1)) - defined (uniquely, if it exists) as the solution of the "convex analytic" extremal problem

$$\sup\{\int_S \log \mu'(s)ds \;\Big|\; |\tilde{c}(t) - \int_S K(t,s)\mu'(s)ds| \leq \varepsilon_0 d(t)\} \qquad (1.2)$$

has many "nice" (desirable) properties; these are:

1) stability with respect to perturbations, i.e. errors in the values of $\tilde{c}_i$, t_i oı $K(.,.)$

2) low complexity, i.e. $\bar{\mu}'(s)$, $s \in S$ can be computed "easily" this means that for $\varepsilon_0=0$ and some important cases, i.e. choices of the kernel function in (1.1) (corresponding to the Nevanlinna-Pick type moment problems for the impendance functions c(.) see below (1.7)) the solution of (1.3), $\bar{\mu}(T_M)$ can be computed exactly in $O(N^2)$ arith.op.-s. In analogons (and approximations) of (1.3) (for $\varepsilon_0=0$)

$$\sup\{\sum_{i=1}^{m} \log \mu_i | <k_i,\mu> = c_i,\ i=1,\dots,N,\ \mu \in R_+^m\} \qquad (1.3)$$

fast numerical algorithms (combining Newton'-s method with special globalization techniques using homotapy and rational extrapolation, see [20] and below) can be constructed for the solution of (1.3).

3) Invariance with respect to affine transformations of the "polyhedral" set $K(t^N,c^N)$ (1.1), more precizely of $K(\mu^N,c^N)$ in (1.3).

4) Existence and easy computability of inner and outer ellipsoidal approximations (for $\varepsilon_0=0$)

$$\bar{\mu} + E(\mu^N,c^N) \subseteq K(\mu^N,c^N) \subseteq \bar{\mu} + mE(\mu^N,c^N) \qquad (1.4)$$

where

$$E(\mu^N,t^N) = \{w | w \in R^{m-N}, <L^{-1}w,w> \leq 1\}\ ,\ L=L^*>0. \qquad (1.5)$$

The solution $\bar{\mu}$ of (1.4) is known (for $\varepsilon_0=0$) as the maximum entropy (and asymptotically maximum likelihood) solution in the theory of statistical spectrum estimation (using generalized covariance data, see [5]).

Below we shall present some results also concerning the minimal atomic solutions of (1.1), i.e. those in which $d\mu$ is concentrated on a minimal number of (distinct) points in S, since these solutions can also be used for devising sequential recovery methods. While - for the concrete moment problems (1.6)-(1.7) the latter solutions μ_{ma} always exists (for $T_M=(z_1,\dots,z_N)$ and the solution (1.2) may not exist) the point in favor of the solutions (1.2) is that they can be computed more easily and recursively (in N).

Of special inteest for us will be the following Nevanlinna-Pick type moment problems for the impendance function $c(t) \Longleftrightarrow h(z)$

$$h(z) = \int_S \frac{1}{z-s} \mu(ds), \qquad S \subseteq (-\infty, \infty). \tag{1.6}$$

Notice that $h(z)$ is a transfer function $<(z\,I-A)^{-1}b,b>$ for $A=A^*$, Spectrum $A=S$. Notice that - up to a simple change of variables - this is the class of continuous time, positive real functions (i.e. those analytic in the right half plane and having positive real part there) associated to stationary, continuous time stochastic processes (note that $\mathrm{Im}\text{-}h(z) \geq 0$ for $\mathrm{Im}\, z \geq 0$); the same class of functions arise as transfer functions of hyperbolic (purely oscillating, energy preserving) systems. The discrete time analogon is the class of Caratheodorey functions defined over the unit disc (corresponding to measures on the unit circle which are symmetrical with respect to the real line)

$$\Omega(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i\theta}-z}{e^{i\theta}+z} \mu(d\theta). \tag{1.7}$$

An other interesting class of moment problems is given by

$$c(t) = \int_S e^{its} d\mu(s) \qquad t \in [\omega_0, \omega_\infty], \; S \subseteq (-\infty, \infty), \tag{1.8}$$

where we set $T_M = T$, assuming thus that all (error contaminated) values of $c(t)$ are available for recovery. Of course, if we have further, a priori knowledge concerning the unknown measure or its derivative this could be included - as side conditions - in the problem (1.2). The contrast between the two solutions $\bar{\mu}$ and μ_{ma} is geometrically that of a "central" and of an extremepoint of a polyhedron. Therefore if we have - for the unknown $\mu \in R^m$ - a possibly nonlinear (convex) inequality constraint $g(\mu) \leq 0$, then this is not added as a condition in (1.3) but is included by adding $\log(-g(\mu))$ to the function to be maximized, see [22].

The second, more difficult problem concerns the sequential choice of the sequence of measurement nodes $t_1, \ldots, t_N$ or in the case $T_M = T$, the linear information functionals

$$\tilde{c}_k = \tilde{f}_k(\mu) := \int \ell_k(t) \tilde{c}(t) dt, \; k=1,\ldots,N, \tag{1.9}$$

where ℓ_k is a sequence of scalar functions defined on T. Here the choice of the right, i.e. suitable family F of functions ℓ, $\ell \in F$ should depend on the final aim of the recovery problem, i.e. on the notion of distance (see below) over the class of objects to be identified (measures, impendances) and is a rather nontrivial, delicate problem. We propose algorithms in which the family of information functionals F (which for the case of costly measurements are defined to be $\ell_k(.) = K(t_k,.)$,

$t_k \in F(T)$) has the structure of a regular binary tree B. This structure for F allows to generate 2^N different sets of possible evaluation patterns $f_1,\dots,f_N$ in an N-step algorithm

$$f_{k+1} = A_{k+1}(f_1,\dots,f_k, \tilde{c}_1,\dots,\tilde{c}_k),\ k=0,\dots,N-1\ , \qquad (1.10)$$

so that - at each step k - one (or both: then $k=2j$, $j=0,1,\dots,$) successors of an already selected element f_j, $j\leq k$ of $F\equiv B$ will be selected as f_{k+1} (resp. f_{k+1}, f_{k+2}). The functions $A_1,A_2,\dots,$ are just used to define that element in F =B which should be "subdivided" (this they do by finding the maximum of "local uncertainties", i.e. those corresponding to the 2k potential followers in B of the already selected elements; the idea being to achieve equilibration of the local errors, one expression of which is that

$$[\mu_m(\Delta_1)]^p\, d(\Delta_i) \simeq \varepsilon,\ S = \bigcup_i \Delta_i, \qquad (1.11)$$

where $\mu_m(\Delta_i)$ denotes the maximal mass that can be placed on a subinterval Δ_i - of length (area) $d(\Delta_i)$ - of S and the value of p depends on the distance chosen to measure the error of recovery. In the Nevanlinna-Pick problems (1.6) - (1.7) we shall use the radii of the inclusion discs and the values of the maximal masses - for given $z_1,\dots z_k$, $c_1,\dots,c_k$ - to compute the "local uncertainties". The role of the Christoffel functions of the arising, generalized power moment problems will be emphasized.

This procedure is a generalization of techniques of "interval subdivision" used in algorithms for adaptive, i.e. sequential node selection for the spline approximations of functions over a line segment or over multidimensional intervals (boxes) developped by the author, see [18] - [22]. In these papers we demonstrated the superiority of suitable constructed sequential N-step algorithms over the optimal passive algorithms (i.e. those corresponding to a simultaneous choice of the information functionals) even with respect to the order in N of the global, i.e. worst case error over classes of measures with total mass not larger than 1 for a number of moment problems.

In problems (1.6) - (1.7) we shall select the countable set F of nodes with a binary tree structure as centres of noneuclidean triangles constituting a regular subdivision of the noneuclidean space (unit disc or upper half plane).

Now these nodes can "tend", with uniform density, to any subinterval of the support set S of the measures μ unlike the traditional information: the values of Markov parameters (correlation functionals) which arise in the limit case, when all nodes are concentrated at one point. Illustrating our remark concerning connections of stability and accuracy of a recovery we point out that in many application the high index Markov parameters (as high order derivatives) cannot be evaluated accurately enough.

Now we have to give examples of distances which proved to be useful for defining the distance of measures. Let us denote by $m_r^{\mu}(s)$ the function which is obtained by r-fold integration of the monoton function μ defined for $S = [0, \pi]$ in (1.7) and $S \supseteq [\alpha, \beta]$ in (1.6), (1.8), thus in the latter case

$$m_r^{\mu}(s) := \int_{\alpha}^{\beta} (s-q)_r^{r-1} \, d\mu(q) \tag{1.12}$$

and define

$$||\mu_1-\mu_2||_r := \sup\{|m_r^{\mu_1}(s) - m_r^{\mu_2}(s)| \; \alpha \le s \le \beta \} .$$

A stronger norm (distance) seems to be of interest also

$$||\mu_1-\mu_2||_0 := \left(\int_{\alpha}^{\beta} |\mu_1(s)-\mu_2(s)|^2 ds\right)^{1/2} . \tag{1.13}$$

For the functions (1.6), (1.7) useful norms can be defined by (fixing a parameter $\gamma \geq 1$)

$$||h_1-h_2||_{\gamma} := \sup_{\mathrm{Im} z>0, \mathrm{Re} z \in [\alpha,\beta]} \mathrm{Im} z^{\gamma} |h_1(z) - h_2(z)| \tag{1.14}$$

$$||\Omega_1-\Omega_2||_{\gamma} := \sup_{|z|<1} (1-|z|)^{\gamma} |\Omega_1(z) \; \Omega_2(z)| \quad . \tag{1.15}$$

While these distances are natural and interesting for the mass recovery problems, it is not yet clear: what is their relation to the coefficient (impendance) recovery problems? Our final remark is that identification methods based on solving a nonlinear least square problem for finding the optimal parameters $(\sigma_1,\dots,\sigma_p)$ yielding best LS fit with finitely parametrized impendance function to the observed data usually lead to nonconvex and rather ill conditioned value functions (to be minimized by a sequential search in the parameter space). Examples of this situation are well known, see e.g. the rather instructive ones in [3], [4]. While such method are instrinsicly "sequential", they usually fail to exploit the deeper structure of the problem. What we propose is not optimizing with respect to a fixed number of free parameters, but trying to get increasingly better parametrizations of the solution with a growing number of parameters (using and equilibrating some measures of local uncertainty). Of course this general idea is (and can only be) implemented by exploiting the deeper algebraic structure of the problem, (simple) recursive expressions for the "uncertainties" :maximal masses, inclusion discs for the values of c(.)). This algebraic structure is partially illuminated by the use of an important paper [24], in which the strength operator theoretic methods for the solution of moment problems (1.6) - (1.7) have been demonstrated (for the related theories of extensions, dilations,... of operators, see [25] and [9]).

2. The analytical centre solution of moment problems

In the finite dimensional case (1.3) the polyhedron of localization for $\mu \in R^m$, $K(t^N,c^N)$, can be described (more simply) by a set of m linear inequalities for $\mu \longleftrightarrow z \in R^{m-N}$

$$K(t^N,c^N) \longleftrightarrow P(a^m,b^m) = \{z \mid b_i \geq <a_i,z>,\ i=1,\dots,m,\ z \in R^{m-N}\}. \tag{2.1}$$

The special solution characterized by (1.3) is then the solution of the following (convex, analytic) extrenum problem

$$\sup.\{(\prod_{i=1}^{m} (b_i - <a_i,z>))^{1/m} \mid z \in P(a^m,b^m)\}, \tag{2.2}$$

By "convex , analytic" it is indicated that the function $\Psi(z)$ to be maximized is strongly concave, see [20] and analytic in z over the feasible set, thus it has a unique maximum, whenever the latter set is bounded (and has a nonvoid interior). These properties together with the algebraically simple form of the gradient and Hesse matrix of $\log\Psi$

$$\operatorname{grad} \log \Psi(z) = \sum_{i=1}^{m} \frac{a_i}{b_i - <a_i,x>},\quad D^2 \log\Psi(z) = -\sum_{i=1}^{m} \frac{a_i a_i^*}{(b_i - <a_i,x>)^2}, \tag{2.3}$$

allow to construct fast numerical methods for the solution of (2.2), see [20], where the Newton method is globalized with the help of suitable homotopies and rational extrapolations (to follow the homotopy curve).

The ellipsoidal approximation (1.4) - (1.5) can be computed (from the solution $\bar{z}(a^m,b^m)$) as follows. Let the linear map $L:R^m \rightarrow R^n$ be defined by $Le_i = a_i/(b_i - <a_i,\bar{z}>)$, $i=1,\dots,m$, where e_i is the i-th unit coordinate vector in R^m,

$$B := (m-1)m\ LL^*,\quad E(a^m,b^m) := \frac{-1}{1-m}\{z \mid <Bz,z> \leq 1\}. \tag{2.4}$$

Note that B is essentially the Hesse matrix of Ψ at $z=\bar{z}(a^m,b^m)$ and (1.4), (2.4) allows to define an "analytic" condition number for a system of linear inequalities. We emphasize the specification "analytic" since there exist other,more tight ellipsoidal approximations for

$P(a^m,b^m)$ in which however the approximating ellipsoid is not an analytic, but only a piecewise smooth function of the data (a^m,b^m). For example one can prove that for the largest in volume ellipsoid, $E_{mv}(a^m,b^m)$ inside $P(a^m,b^m)$, the homotheticity constant m-1in (1.4) can be replaced by n=m-N. Now, even if the latter problem is again "convex" (since det is a concave function over the set of symmetric, positive definite matrices) the parameters of $E_{mv}(a^m,b^m)$ are more difficult to compute (they are only piecewise smooth in (a^m,b^m)).

The most remarkable property of the solution concept (1.2) is perhaps the one of its simple algebraic structure and - as consequence thereof - $O(N^2)$ computability of $\bar{\mu}(t^N,c^N)$, exhibited in the Nevanlinna-Pick type problems (1.6) - (1.7). The results describing these solutions - of course without (1.2) and our (geometric) interpretation (1.4) of it - in terms of orthogonal polinomials and Pade approximants go back presumably to Christoffel, with contributions by many authors like Stietjes Kolmogorov, Szegő, Baker, Goncar, Burg, Dewilde-Dym, Krein and others, for a survey concerning the case of (1.7) see [5] and [8], while for (1.6) further (earlier) references in [11] and [16]. Nevertheless it is interesting that for the moment problems (1.6) even in the classical case (corresponding to $z_i=\infty$, i=1,...,N)

$$c_n = \int_{-\infty}^{\infty} s^n \, d\mu(s), \quad n=0,1,\dots,2K-1=N-1 , \tag{2.5}$$

the solution $\bar{\mu}(\infty^N,c^N)$, i.e. the one, which solves

$$\sup\{ \int_{-\infty}^{\infty} \log\mu'(s)ds \mid \mu \in K(\infty^N,c^N) \text{ from } (2.5)\} , \tag{2.6}$$

-which is computable in $O(N^2)$ arithmetical operations, see Theorem 1 below - has not been identified (used) earlier! Only in the context of the problem (1.7) were the maximum entropy solutions, i.e. those determined by the problem (1.2) identified with (i.e. computed as the inverse of the squared module of) orthogonal polynomials (on the unit circle with respect to the measure $\mu_N=m(z^N,\mu)$)

$$\mu_N'(s) = \mu'(s) \left| \prod_{i=1}^{N} (s-z_j) \right|^{-2} ; \tag{2.7}$$

where - as would be natural for the case of real transfer functions - we need not to assume that the interpolation points are real symmetric, i.e. chosen in conjugate pairs or as real numbers . The specail case $z_j=0$, j=1,...,N corresponds to orthogonal polynomials with respect to μ and to the original "maximum entropy" interpretation advocated by Burg. On the other hand (and earlier), for real symmetric data in (1.6) orthogonal polynomials - with respect to square root of the inverse polynomial the weight (2.7), now $s\in R^1$ - have been identified as providing denominators for the multipoint Pade approximation problem concerning (i.e. based on the computed) values of the Stieltjes type function in (1.6), see e.g. [11], where it is shown that for real-symmetric data the approximating function is also of Stieltjes type. The roots of these polynomials provide a minimal atomic solution of the moment problem (1.6). These roots can be computed by solving an eigenvalue problem for a symmetric matrix computed from the data (z^N,c^N) which is a better way for computing them (as compared to polynomial root finding). Since this seems to be not known (see e.g. [8]) we shortly describe the algorithm,for details see [21],which is based on the operator theoretic treatment of Nevanlinna-Pick moment problems first given in [24], see also [1]. First one has to compute a factorization of the Hankel matrix H formed from the moments in (2.5) $H=C^*C$, where $C\in R^{k\times k}$ - instead of the Cholesky factorization we propose the symmetric factorization $C=C^*$, since this can be computed by a more stable, fastly convergent iteration. After this we have to recover the matrix A from the equations $\left(\text{note that A solves the "representation" } \{c_k=\langle A^k e_0,e_0\rangle\}\right)$

$$AC = (\sigma C,v), \text{ where } (Hv)^i = c_{k+i+1},\ i=1,\dots,k ,$$

here σ is the left shift on the columns of matrices. The roots are the eigenvalues of the symmetric matrix A.

However these roots are very ill conditioned functions of the data (z^N,c^N) (i.e. of the generalized "moments" - which give, in both cases (1.6) and (1.7), the classical power moments with respect to the weight (2.7)). This can be explained as the ill conditioning of extrem vertices of polyhedrons(as function of the parameters of the linear inequalities forming the polyhedron). It is to be expected that $\bar{\mu}(z^N,t^N)$ as the analytic centre is a more smooth function of the data (i.e. of the generalized moments). Indeed below we shall see that for computing $\mu(z^N,c^N)$ we have only to solve a linear equation with a Toeplitz matrix (in order to compute an orthogonal polynomial). The fact that for (1.6) - whenever it has more than one solution - $\bar{\mu}(z^N,t^N)$ is a rational function which can be computed in $O(N^2)$ arithmetical operations follows from the next theorem.

<u>Theorem</u>. The solution of the problem (2.6) is the reciprocal of a (positive) polynomial of degree not larger than 2k , which exist iff the moment problem (2.5) has at least 2 solutions and which can be computed in $O(k^2)$ operations from the data (2.5).

<u>Proof</u>. We use the classical transformation $e^{i\theta} = (it-1)/(it+1)$ from the real line to the unit circle, see e.g. [8] or [16] (where-unlike to [8]-we need not to assume that μ in (1.7) is real symmetric,but only

that

$$\int_{-\infty}^{\infty} (1+s^2)^{-1}\, d\mu(s) < \infty$$

which obviously holds for μ in (2.6) since the c_i-s are assumed to be finite, to transform the well known maximum entropy interpretation of the autoregressive" solution of the trigonometric moment problem

$$\bar{\mu}'(s) = |P_k(s^{-1})|^{-2},\ s = e^{i\Theta},$$

where $P_n(e^{i\Theta})$ is the n-th orthogonal polynomial. Now by the above transformation $\Theta \rightarrow t$ a trigonometric polynomial of support in [-k,k] is transformed into a polynomial of degree not greater than 2k, while in this transformation the first (2k+1) trigonometric moments (with indices in [-k,k]) uniquely determine the first 2k+1 power moments and vice versa, see [8]. For measures of compact support we could use alternatively the transformation $t = \frac{1}{2}(z+z^{-1})$ and the identity $\varphi_k(z)\varphi_k(z^{-1})=\Psi_k(z+z^{-1})$ for arbitrary k-th degree polynomial φ_k and suitable k-th order polynomial Ψ_k.

Notice that the fact that the solution of (2.6) is the reciprocal of a positive polynomial follows very simply also from the rule of Lagrange multipliers applied to the extremal problem (2.5)).

Recalling the form of the Christoffel function - for the classical moment problem (2.5) to which the general case can be reduced -, also the reciprocal of polynomial of degree 2k - we may expect that the rational function $\bar{\mu}(z^N, c^N)$ has similarly good properties of mass reproduction as the Christoffel function (the latter does not solve in an exact sense the partial moment problem but gives a good recovery, moreover is very useful for providing bounds, expressions for the maximal masses and the remaining uncertainties in the values of $\mu(s)$ and of the impendance functions $\Omega(z)$ and $h(z)$, see [2] and below.

3. Sequential methods of node selection for positive real functions

In [19] we already presented algorithms of node selection for the "discrete time" problem (1.7). Here we concentrate on the "continuous time" problem (1.6) (and provide also improvements of the results in [19]). First of all : the assumption about the "measurability" (i.e. accessibility) of the values c^N in (1.6) as well as in (1.7) is realistic at least more realistic than the assumption that we can measure the impulse response. The values of Imh(z), resp. ReΩ(z) may tend to infinity as z approaches a discontinuity of the measure μ (a "resonant frequency"). By the Schwarz inequality we know that - for measures of bounded total mass

$$|\mathrm{lm} z|^{\gamma} \cdot |h(z)| \leq \text{const, when } \gamma \geq 1 .$$

Therefore it seems to be natural to assume that we can measure (within fixed accuracy ε_0) the values of $|\operatorname{Im} z| h(z)$, say for $\operatorname{Im} z>0$, and for $\operatorname{Re} z \in [\alpha,\beta]$. We remark that in the case of stationary, stochastic processes the values of these "associated" functions can be approximately recovered from a finite length realization of the process by Solving the linear least square fitting of this data to an ARMA model, whose transmission zeros are fixed to be just $z_1,\dots,z_N$. Notice that for a real input output system the measure μ is symmetric with respect to zero, in which case we set $\alpha=-\beta$. The boundedness of α and β correspond to the fact that in practice we cannot generate very high frequency (energy) inputs. This assumption points to an important distinction (loss of analogy) between the discrete and continuous time case: in the latter the spectral density need not be integrable over R^1.

Now we define a countable set of "potential" measurement nodes having the structure of a binary tree (that of dyadic subintervals of $[\alpha,\beta]$ (from which $z_1,\dots,z_N$) will be selected. To this end consider a dyadic subinterval $[\gamma,\delta]\subseteq[\alpha,\beta]$ and the noneuclidean triangle formed by the three points γ, δ, $(\gamma+\delta)/2$ as vertices and the half circles (in the upper half plane) connecting them (as sides). Take an arbitrarily fixed, inner point ζ_0 of the "base" triangle corresponding to $[\alpha,\beta]$ and let $\zeta_{i_1,\dots,i_k}$ (where $i_j\in\{0,1\}$ $j=1,\dots,k$, $k=1,2,\dots$) be the points corresponding to ζ_0 in the noneuclidean transformation $T_{i_1\dots i_k}$ - corresponding to an arbitrary element $i_1,\dots,i_k$ of the binary tree B, i.e. the map transforming the base triangle to the triangle corresponding to the dyadic interval indexed by $i_1,\dots,i_k$. Each point $\zeta_{i_1},\dots,i_k$ has exactly two followers and one ancestor. Let the above system of nodes be denoted by Σ. The well known "Blaschke condition"

$$\sum_{(i_1,\dots,i_k)\in B} \operatorname{Im} \zeta_{i_1,\dots,i_k}\left(1+|\zeta_{i_1,\dots,i_k}|^2\right)^{-1} < \infty$$

being fulfilled, it follows that the values of $\{h(\zeta_{i_1,\dots,i_k}),(i_1,\dots,i_k)\in B\}$ uniquely determine (i.e. fix) the function h, see e.g. [6]. By the way this condition also shows that - for $\gamma\geq 1$ -

$$||\mu||_\gamma := \sup_{i_1,\dots,i_k\in B} |\operatorname{Im}\zeta_{i_1,\dots,i_k}|^\gamma |h(\zeta_{i_1,\dots,i_k})|$$

defines a norm in the space of the impendence functions h, see (1.14)

Now we recall the results of the Nevanlinna-Pick theory of the "interpolation" problem (1.6): for fixed values (z^N,c^N) and an arbitrary value of z the values of $\overset{\mu}{h}(z)$, when μ varies over the class $K(z^N,c^N)$ belong to (i.e. fill up) a disc whose centre $c_N(z)$ and radius $r_N(z)$ are easily-in $O(N^2)$ arithmetical operations - recursively (in N) computable functions of (z^N,c^N,z). In fact, all solutions $h(\cdot)=\overset{\mu}{h}(N)$ of (1.6) can be parametrised by an arbitrarily chosen, unimodularly bounded (Schur) function $s(.)$, i.e. an analytic function such that $|s(z)|\leq 1$, for all z with $\mathrm{Im} z\geq 0$. The special solution $\bar{h}=h^{\mu}$ corresponding to the choice (1.2) is obtained by taking the "trivial" extension: $s(z)\equiv 0$, (this follows from theorem 1 and known results for the discrete time case (1.7), see e.g. [5]).

Using the equivalence of (1.6) with a partial, polynomial moment problem (for a modified weight) we can use the well known formulas, see e.g. their exposition in [2], for the radii of inclusion discs in terms of the Christoffel functions associated to that moment problem.

Now we can propose the following sequential node selection algorithms. In them the indices of the selected nodes $z_j = \zeta_{i_1^j,\dots,i_k^j}$, $j=1,\dots,n$ constitute a regular subset T_n of the binary tree B, which means - by definition - that if $\zeta_{i_1,\dots,i_k}$ belongs to T regarded as a subset of Σ then $\zeta_{i_1,\dots,i_{k-1}}$ also belongs to T_n. The "boundary "of the set T will be defined as the set of elements in B (but not in T) which are immediate followers of an element of T.

Algorithm 1. Suppose we have already computed the values of $h(z)$ for a regular subset $z\in\{z_1,\dots,z_n\}= T_n$ of the binary tree. Compute

$$\max\{r_n(\zeta)|\mathrm{Im}\zeta|^{\gamma}|\zeta\in \text{boundary of } T_n\} \tag{3.2}$$

and select $z_{n+1} := \bar{\zeta}$, a point where the above maximum is realized. The reasons why this algorithm is expected to have essentially better performance than other (passive) ones are explained in [17] - [21]. An important feature of the above system of nodes (or linear evaluation functionals of a restricted type) is that they are maximally "separated" while maintaining the conditions of completeness (3.1). Notice that $r_n(\zeta)$ depends on all previous values (z^n,c^n). A more simple algorithm can be proposed - in which only the value of

$$\max\{\mathrm{Im}^{\gamma}\,\zeta\cdot\mathrm{Im}(h(\zeta))|\zeta\in T_n\} \tag{3.3}$$

needs to be computed, but-at each step n-two followers of the element $\bar{\zeta}$ realizing (3.2) are selected (and Imh evaluated at them) to form T_{n+1}. This second algorithm is based on the similarity of the system of Poisson kernels $\mathrm{Im}(s-z)^{-1}$, for z in Σ, as $s\in[\alpha,\beta]$, to the system of the Schauder kernels, see [17], [19]. The latter system of kernels yield an optimal linear evaluation system for recovering the measures μ in the norm $||.||_2$ defined in (1.13). Optimal order sequential algorithms for recovering μ in the norms (1.13) - for arbitrary natural $r\geq 2$ - have been constructed and proved to be essentially (globally) superior to passive ones in [18]. Recalling results from [10] showing that functions with the singularity of $(s-t)_+^r$ can be very well approximated by rational functions, we may expect that the above algorithms, analogous of those in [18] will be of optimal order error for the recovery in the norms (1.13) (and also in the norms (1.14) or (1.15) since the latter seem to be equivalent to the previous ones).

Indeed with respect to both norms in the optimal case the uncertainties of the measure over the system dyadic subintervals corresponding to a regular subset T_n must be equilibrated in the sense of (1.11) where p uniquely depends on r or on γ; for r this has been proved in [18], for γ this can be seen using the Stieltjes inversion formula

$$\mu(\delta) - \mu(\gamma) = \lim_{\tau\to 0} \int_\delta^\gamma \Im m\, h^\mu(s+i\tau)ds ,$$

see e.g. [2] or [16]. A further algorithm can be proposed remembering that the value of the Christoffel function - at an arbitrary $s \in [\alpha,\beta]$ provides the maximal mass $m_N(s)$ that can be placed at that point under the conditions (1.6). Thus, in order to achieve an equilibration of the uncertainties according to (1.11) we propose to select the element to be subdivided - at step n - by computing

$$\max(m_n(\delta) + m_n(\gamma))\ (\delta-\gamma)^p, \left(p = (r-1)^{-1} \text{ resp. } p = (\gamma-1)^{-1}\right)$$

over the "boundary" subintervals corresponding to T_n.

Finally we describe a sequential method for the recovery problem (1.8), where the norm $\|.\|_2$ in (1.13) is considered, and the Schauder functions $S_\lambda(.)$ indexed by the elements $\lambda \in B$: the binary tree of dyadic subintervals of T, say $T = [\alpha,\beta]$ are used as evaluation functionals $f_k(.)$ in (1.9), and an algorithm from [18] and [19]. In order to recover (approximately) the values of the Schauder functionals $<S_\lambda,\mu>$ we have to solve the continuous linear programming problem: find for each $\lambda \in B$ min $(\varepsilon_1+\varepsilon_2)$, with respect to the choice of $\alpha_\lambda : T\to R^1$

$$\begin{aligned} &\left|\int_T \alpha_\lambda(t)e^{its}dt - S_\lambda(s)\right| \le \varepsilon_1, \text{ for all } s \in S \\ &\varepsilon_0 \int_T |\alpha_\lambda(t)|\delta(t) \le \varepsilon_2 . \end{aligned} \qquad (3.4)$$

The value of the corresponding Schauder functional will then be approximated within error $\bar\varepsilon_1 + \bar\varepsilon_2$ - uniformly over the class of measures with bounded total mass - by the expression

$$\hat c_\lambda = \int_T \bar\alpha_\lambda(t)c(t)dt.$$

Notice that the optimal solution of (3.4) $\{\bar{\varepsilon}_1,\bar{\varepsilon}_2,\bar{\alpha}(\cdot)\}$ depends only on ε_0, $d(\cdot)$, T and S. Moreover, the fact that the functions S_λ have uniformly over $\lambda \in B$ a bounded Lipschitz constant (having the same "time" and "magnitude" of discontinuity in their derivatives) indicates that (presumably) the value $(\bar{\varepsilon}_1+\bar{\varepsilon}_2)$ can be exactly majorized in terms of the values of the values of ε_0 and $\delta(\cdot)$ alone, i.e. independently of λ. Of course, the selection of new indices λ_{j+1}, thus the algorithm should be stopped at a step j, for which $\varepsilon_1+\varepsilon_2 \geq \text{const}\cdot j^{-2}$ is first satisfied. Here we used the ("linear") stability of algorithm in [18], [19] based on sequential evaluation of the Shcauder functionals and the rule: subdivide (and compute the 2 new Schauder functionals for) that subinterval which gives the largest Schauder functional,- here p=1 in (1.11) - (which in fact is sequential method for evaluation of the succesive second order divided differences of the function $m_2^\mu(s)$ at three points δ, $(\delta+\gamma)/2$, γ of a dyadic subintervals of T, in order to measure the local uncertainty concerning m_2^μ in the uniform metric over this interval) with respect to errors $(c - \tilde{c})$. Notethat for the approximation (recovery) of convex functions m_2^μ in the uniform norm based on N evaluations of m_2^μ, any passive N-step algorithm has a global error larger than const N^{-1}, while the error of the above N-step sequential algorithm is smaller than const N^{-2}.

References

[1] T. Ando, Truncated moment problems for operators, Acta Sci. Math. (Szeged), 31 (1970), 319-334.

[2] N.I. Akhiezer, The classical moment problem, Oliver and Boyd, Edinburgh, 1965.

[3] A. Bamberger, G. Chavent, P. Lailly, About the Stability of the Inverse Problem in 1-D Wave Equations - Application to the Interpretation of Seismic Profiles, Appl. Math. Optim. 5. (1979) 1-47.

[4] A. Bultheel, On the ill conditioning of locating transmission zeroes in least squares ARMA filtering, J. Comp. Appl. Math., 11. 1. (1974) 103-109.

[5] P. Dewilde, H. Dym, Losless Inverse Scattering, Digital Filters and Estimation Theory, IEEE Trans. Inf. Th. vol. IT 30. No 4. (1984) 644-661.

[6] J.B. Garnett, Bounded Analytic Functions, Acad. Press, New-York, 1981.

[7] W. Gautschi, Orthogonal Polynomials - Constructive theory and Applications, J. Comp. Appl. Math. 12-13. (1985) 61-77.

[8] Y. Genin et al. The trigonometric moment problem and speeach analysis, Philips J. of Research, 37 (1982) 277-292.

[9] I. Gohberg, ed., Proc. Toeplitz Centennial Conf. Operator Theory: Advances and Applications, vol. 4. Birkhäuser V., Basel, Boston, 1982.

[10] A.A. Goncar, On the speed of rational approximation of continuous functions with characteristic singularities URSS Math. Sbornik, tom. (115):4, (1967), 630-638.

[11] A.A. Goncar, On the rate of convergence of rational approximations to analytic functions, Trudi Inst. V.A. Steklowa, vol.166 (1984), Nauka, Moscow.

[12] W.B. Gragg, W.I. Harrod, The Numerically Stable Reconstruction of Jacobi Matrices from Spectral Data, Numer. Math. 44 (1984) 317-335.

[13] U. Grenander, G. Szegõ, Toeplitz Forms and Their Applications, Univ. of Calif. Press, Berkeley, 1958.

[14] Th. Kailath et al., Differential Methods of Inverse Scattering, SIAM J. Appl. Math. v. 45 No 2. (1985) 312-335.

[15] P. Nevai, Géza Freud, Orthogonal Polynomials and Christoffel Functions (A Case Study) Ohio State University, (1985), 289 p.

[16] M.G. Krein, A.A. Nudelman, Markow'-s moment problem and extremal problems, Nauka, Moscow, 1973. (in russian).

[17] Gy. Sonnevend, About the superiority of sequential methods for the solution of moment problems, to appear in the proceedings of the Seminar on Approximation Theory, Intern. Banach Center, ed. by Z. Ciesielski, Warsaw, 1986.

[18] Gy. Sonnevend, Sequential algorithms of optimal order global error for the uniform recovery of function with monotone r-th derivatives, Analysis Mathema.ica, tom. 10., fasc. 4 (1984), 311-335.

[19] Gy. Sonnevend, A class of sequential algorithms for spectral approximation with rational and with Haar functions, Proc. A. Haar Memorial Conf. (1985) Budapest, to appear at North Holland.

[20] Gy. Sonnevend, An "analytic centre" for a system of convex inequalities and its application for convex programming, Proc. 12th IFIP Conf. on System Modelling, Budapest, 1985, appears in Lect. Notes in Control and Inf. Sci. Springer V.

[21] Gy. Sonnevend, Stable algorithms for solving some problems arising in inverse scattering, Proc. Conf. Numerical Methods (J. Bolyai Math. Soc.), Miskolc, Hungary, 1986.

[22] Gy. Sonnevend, An optimal sequential algorithm for the uniform approximation of convex functions on $[0,1]^2$, Appl. Math. Optim. 10 (1983), 127-142.

[23] W.W. Symes, Impendance Profile Inversion via the First Transport Equation, J. Math. Anal. Appl., vol 94 (1983), 435-453.

[24] B. Szõkefalvi-Nagy, A. Korányi, Operatorentheoretische Behandlung und Verallgemeinerung eines Problemkreises in der komplexen Funkcionentheorie, Acta Methematica, vol. 100 (1958) 171-202.

[25] B. Szõkefalvi-Nagy, C. Foias, Analyse harmonique des opérateurs de l'espace de Hilbert, Masson et C^{ie}-Akadémiai Kiadó, 1967.

[26] D.I. Thompson, Spectrum estimation and harmonic analysis, Proc. IEEE. 70 (1982) 1055-1096.

[27] V.V. Voevodin, E.E. Tirtishnikow, Computations with Toeplith matrices in "Numerical Processes and Systems", ed. by. G.I. Marchuk, vol. 1. (1983) pp. 124-266.

[28] K. Yosida, Functional Analysis, Springer Verlag, Berlin-New York, 1965.

[29] Z.D. Yuan, L. Ljung, Black Box Identification of Multivariable Transfer Functions-Asymptotic Properties and Optimal Input Design, Int. J. Control, vol. 40. No 2. (1984) 223-256.

MODELLING AND CONTROL OF TWO COORDINATED ROBOT ARMS

T.J. Tarn and X. Yun
Washington University
Department of Systems Science
and Mathematics
St. Louis, MO 63130 U.S.A.

A.K. Bejczy
California Institute of Technology
Jet Propulsion Laboratory

Pasadena, CA 91109 U.S.A.

ABSTRACT: In this paper we model two coordinated robot arms by considering the two arms (1) working on the same object simultaneously and (2) as a closed kinematic chain. In both formulations a new dynamic control method is discussed. It is based on feedback linearization and simultaneous output decoupling. In the first formulation the control method uses a dynamic coordinator acting on relative position and velocity task space errors and on relative force-torque errors between the two arms as sensed at the end effectors. This method is novel because we could superimpose the position and velocity error feedback with the force-torque error feedback in the task space simultaneously. In the second formulation the nonlinear feedback is augmented with optimal error correcting controller, which operates on the task level. This formulation has the advantage of automatically handling the coordination and load distribution between two robot arms through the dynamic equations. By choosing independent generalized coordinates, kinematic and dynamic constraints have been taken into account in the process of deriving the equations of motion.

1. INTRODUCTION

It is an easy daily routine to tie up shoelaces by two hands. How can we let two robot arms tie up shoelaces? Among these necessities such as proper hands, sensors and so forth, the coordination between two robot arms is the key to fulfill the job.

While tying-shoelace provides a good example of tasks requiring coordination, our study of coordination is mainly toward industrial applications. As application of robots on manufacturing floors and elsewhere increases, so does the use of two or more robots operating in the same work space and cooperating on the same job. The coordination among robots is essential in many industrial and other applications, such as material handling and assembly, servicing and maintenance in remote hazardous places, etc. The study of coordination problems between two robots doing a single job is in its infancy, though a two-handed human being is capable of doing almost all kinds of jobs within his capacity.

The basic research objective of the coordinated control of two arms is to design a control system which is able to command both arms in such a way that the two arms operate in a kinematically and dynamically coordinated fashion and respond to the working environment without collisions. Although the control problem of two or multiple arms is complex, some examples of applications, such as a two-arm lathe loader, a two-arm robot press loader/unloader, and two single-arm robots working together to handle stamping press loading and unloading, are given by Chimes [1]. In these applications, the problem is solved specifically. The system design is based on a solid understanding of the problem.

Coordinated control of two- and multi-robot arms has been studied by many investigators [2-8]. It appears that the existing coordinated control methods fall in lack of both systematic synthesis of the control system and full consideration of robot arm dynamics. We take two approaches to attack the coordination problem. Based on the force control method, the first approach solves the coordination problem by monitoring the interactive forces and moments between the end effectors of the two robot arms. This is a rather natural treatment to the coordination problem since the most information on which people rely to move an object by two hands is the forces felt by the skin of the fingers. Instead of considering each robot separately, our second approach treats the two robot arms as a closed chain at the very beginning. This is the case when two robot arms are holding and transferring an object from one place to another. Including the object as one of the links they form a closed chain.

For both approaches, we apply the differential geometric control theory to the dynamics of robot arms. By appropriate nonlinear feedback and diffeomorphic transformation, we are able to linearize and decouple the original nonlinear and coupled dynamic equations. The control algorithms are then designed based on the theory of linear systems. This method gives a unified approach to feedback design and extends the control theories and practices to a level where a real-time robot control system can directly absorb task space commands.

2. CONTROL COORDINATION OF TWO ROBOT ARMS VIA FORCE FEEDBACK

When two robots are located in the same work space and work on the same jobs, e.g., transfer a heavy workpiece from the convey to the working table, each robot contribute part of the force to lift and to move the workpiece. For loading and unloading tasks, the coordinated controller is to be so designed that the load is distributed between two robot arms according to their loading abilities and that no forces of the two robots are cancelled because of the opposite direction of forces, that is, we do not want two people to push or pull a door at the same time from the two sides.

Force control of robot arms has been studied by many researchers [9-24]. Recently, a dynamic hybrid control method is developed by Yoshikawa [25]. In this method, constraints on the end effectors are formulated by a set of hypersurfaces. In contrast with the previous hybrid control methods, manipulator dynamics is rigorously taken into the derivation of control law. The method can be applied to manipulators with six or more degrees of freedom. However the control law derived in the paper is task dependent.

We presented a new dynamic control strategy for force feedback in [26]. It is conceptually much clearer than those existed in the literature and appears implementable. We now use the framework described in [26] on force feedback to obtain a coordinated strategy for two robot arms working on the same object. We first incorporate the mass of the object into the dynamic projection parameters of one of the two robot arms, say robot 'a'. Let $p(q_a)$ be the position and/or orientation of the object, and let F be the force and/or torque sensed at the end effector of robot 'b'. The dynamic equation of motion for robot 'b' is

$$D_b(q_b)\, \ddot{q}_b + E_b(q_b, \dot{q}_b) + J_b'(q_b)\, F = \tau_b$$

and the dynamic equation of motion for robot 'a' is

$$D_a(q_a)\, \ddot{q}_a + E_a(q_a, \dot{q}_a) - J_a'\, (q_a)F = \tau_a.$$

Considering the enlarged output equation of the form

$$y = \begin{bmatrix} p(q_a) \\ F \end{bmatrix},$$

we apply the feedback linearization and decoupling method to the above system such that the inputs τ_a will only regulate the outputs $p(q_a)$ and the inputs τ_b will only regulate the output F. Note that the solution of this problem has application in cases where the second robot arm has to support dynamically the actions of the first robot arm which are defined in geometric terms.

In the design of coordinated control of two robot arms, as presented in our paper [27], we have used the master/slave mode to obtain an optimal coordinator (loop 1 in Figure 2, [27]). This dynamic coordinator is acting on relative position and velocity errors between the two arms. In future study we would like to investigate the use of indistinguished mode as shown by loop 3 in figure 2 of [27]. With the force feedback strategy outlined above we would like to investigate the combination of optimal coordinator with force feedback (loop 1 plus loop 2 in Figure 2, [27]) and of optimal error corrector with force feedback (loop 3 plus loop 2 in Figure 2, [27]).

In [33], Leahy, Nugent, Valavanis and Saridis pointed out the requirement of better dynamic models on real-time closed loop robot arm control. In the development of force feedback strategy as presented here the robot arm dynamics has been taken into account rigorously.

3. COORDINATED CONTROL OF TWO ROBOT ARMS: CLOSED CHAIN APPROACH

In those robot applications where two robot arms transfer an object by holding it from the two ends, a closed chain is formed by the two robot arms and the object through the ground. To describe the dynamic behavior of the whole mechanical system, we will establish equations of motion by considering the system as a closed chain from the beginning.

Now we consider two robot arms holding an object which can move gently between the tips of the end effectors. We assume that the object and the end effectors are mechanically locked and that each robot arm has six links. The closed chain has 13 links and 14 joints $(m=14)$. Those two joints connecting the object and the end effectors have no actuators. From Gruebler's equation [28], the degrees of freedom of the two-arm chain is $n = m-3 = 14-3 = 11$.

We denote the joint variables of the two-arm chain by

$$\theta = [\theta_1\ \theta_2\ \cdots\ \theta_7\ \theta_{1'}\ \theta_{2'}\ \cdots\ \theta_{7'}]'$$

where $\theta_1, \ldots \theta_6$ are the original joint variables of the first robot (or robot 'a'), θ_7 is the joint angle of the joint connecting the object and the end effector of robot 'a'. $\theta_{1'}, \ldots \theta_{7'}$ have the same meaning for the second robot (robot 'b'). The joint driving torque (or force) vector is denoted by $F = [F_1, F_2, \ldots, F_{14}]'$. In case that a joint has no actuator, the corresponding component of the force vector F is assigned to be zero. Choosing the generalized coordinates in the following way

$$\begin{aligned} q &= [q_1\ q_2\ q_3\ q_4\ q_5\ q_6 \ \big|\ q_7\ q_8\ q_9\ q_{10}\ q_{11}]' \\ &\triangleq [\theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5\ \theta_6 \ \big|\ \theta_{1'}\ \theta_{2'}\ \theta_{3'}\ \theta_{4'}\ \theta_{5'}]' \end{aligned}$$

then we can easily get $\theta = \Theta(q)$ from the geometric arrangement of the two robot arms.

Suppose that a world coordinate frame has been located in the work space and that one coordinate frame has been assigned to each link of the closed chain. In the process of expressing the energy, we will describe the energy of the object in terms of θ_7 instead of $\theta_{7'}$. Using homogeneous coordinates together with the Denavit-Hartenberg four-parameter representation of robot arm kinematics, and using the Lagrangian formulation of kinetics, the dynamic model of the closed-chain is as follows:

$$D(q)\ddot{q} + E(q,\dot{q}) + G(q) = J_\theta F \qquad (1)$$

where $D(q) = J_\theta' \tilde{D}(\theta(q))\, J_\theta = \begin{bmatrix} D_{11}(q) & \cdots & D_{1,11}(q) \\ \vdots & & \vdots \\ D_{11,1}(q) & \cdots & D_{11,11}(q) \end{bmatrix}$

$$J_\theta = \frac{\partial\, \theta}{\partial\, q'},$$

$$\tilde{D}(\theta) = \begin{bmatrix} \tilde{D}_{11} & \cdots & \tilde{D}_{17} & & & & \\ \vdots & & \vdots & & \bigcirc & & \\ \tilde{D}_{71} & & \tilde{D}_{77} & & & & \\ & & & \tilde{D}_{1'1'} & \cdots & \tilde{D}_{1'6'} & 0 \\ & \bigcirc & & \vdots & & \vdots & \vdots \\ & & & \tilde{D}_{6'1'} & \cdots & \tilde{D}_{1'6'} & 0 \\ & & & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$E(q,\dot{q}) = J_\theta' \tilde{D}(\theta(q)) \begin{bmatrix} \dot{q}' \dfrac{\partial^2\, \theta_1}{\partial\, q^2}\, \dot{q} \\ \vdots \\ \dot{q}' \dfrac{\partial^2\, \theta_{7'}}{\partial\, q^2}\, \dot{q} \end{bmatrix} + J_\theta' \begin{bmatrix} \dot{q}'\, J_\theta'\, D^1 \\ \vdots \\ \dot{q}'\, J_\theta'\, D^{7'} \end{bmatrix} J_\theta \dot{q},$$

$$D^p = (D_{pij}),\ p=1,\ \ldots,\ 7';$$

$$G(q) = -J_\theta' \begin{bmatrix} D_1 \\ \vdots \\ D_{7'} \end{bmatrix}.$$

Note that $\tilde{D}_{ij}(q)$ is the inertial load projection function to joint "i" related to acceleration at joint "j", $D_{ijk}(q)$ is the centripetal $(j=k)$ or Coriolis $(j\neq k)$ force projection function to joint "i" related to velocities at joints "j" and "k", and $D_i(q)$ is the gravity load at joint "i". The general function definitions of the D_i, $\tilde{D}_{ij}$ and D_{ijk} dynamic projection functions can be found in [29, 30].

For transferring tasks we take output equations to be the position and orientation of the object in the world coordinate frame. More specifically, the outputs can be described by a 6-dimensional vector

$$y = \begin{bmatrix} h_1(q) \\ h_2(q) \\ \vdots \\ h_6(q) \end{bmatrix} \tag{2}$$

in terms of the generalized coordinate q. The first three components h_1, h_2 and h_3 of y represent the position and the last three components h_4, h_5 and h_6 represent the orientation of the object.

To perform linearization and output block decoupling for the system (1) with output equation (2), we can now use the algorithm developed by us in [31, 32] to find the required nonlinear feedback and the required nonlinear coordinate transformation. The control problem of the two arm closed chain is then simplified to a design problem of linear systems.

Note that the obtained linear system consists of six independent subsystems. Since each subsystem is controllable, we may locate the poles of each subsystem by adding a constant feedback. As we have done for one arm control system [32], an optimal correction loop may also be designed to reduce the tracking error and to improve the robustness against model uncertainties.

This formulation has the advantage of automatically handling the coordination and load distribution between two robot arms through the dynamic equations. By choosing independent generalized coordinates, kinematic and dynamic constraints have been taken into account in the process of deriving the equations of motion.

4. CONCLUSIONS

Our approaches to the design of coordinated control of two robot arms are motivated by the desire of making rigorous use of the dynamics of two robot arms in contrast to the existing two arm control algorithms in which kinematic constraints are considered only.

Using the results from differential geometric system theory, we are able to linearize and to decouple the complicated dynamic equations of two robot arms including the object held by the two arms. Independent of the approach being taken, we eventually deal with a linear, decoupled system. Thus we can have a unified design technique for coordinated control of two robot arms.

Our presentation in this paper is for the feedback system design of two coordinated robot arms. However our approaches can easily be extended to multi-robot arms.

It should be noted that both methods used in this paper are systematic and are robot arm independent. The most important feature is that the control algorithms are task independent, that is, there is no need to change the structure of the controller or even the parameters of the controller from task to task. As natural as would be, the change of tasks only causes the adjustment of the input command which is conveniently given in the task space rather than in the joint space. The two control methods can be used in slightly different situations. For example, if the two robot arms are loosely connected through the object, the force control approach is preferable; if the two robot arms are mechanically locked while transferring the object, the closed-chain approach is more likely a solution.

The new dynamic control method proposed here also brings the feedback implementation closer to "intelligent control" of robot arms. By definition, "intelligent control" operates on the task level, and it is being manifested through robot performance in the task space relative to task space commands and task space variables. The new dynamic feedback method described in this paper transforms the robot arm control problems to the task space and performs robot serving or regulation in terms of task space variables within a linear system frame, allowing also the use of powerful techniques from optimal control of linear systems. Since the new dynamic feedback method establishes a direct control response to task space commands, it renders the control "intelligent" in the sense of the above definition of intelligent control.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation under grants DMC-8309527, DMC-8505843, ECS-8515899, and INT-8519654.

REFERENCES

[1] P.R. Chimes, "Multiple-Arm Robot Control Systems," Robotics Age, Oct. 1985, pp. 5-10.

[2] C.O. Alford, S.M. Belyeu, "Coordinated Control of Two Robot Arms," International Conference on Robotics, Atlanta, Georgia, March 13-15, 1984, pp. 468-473.

[3] Y.F. Zheng, J.Y.S. Luh, "Constrained Relations Between Two Coordinated Industrial Robots," Proc. of 1985 Conference of Intelligent Systems and Machines, Rochester, Michigan, April 23-24, 1985.

[4] Y.F. Zheng, J.Y.S. Luh, "Control of Two Coordinated Robots in Motion," Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, Florida, Dec. 11-13, 1985, pp. 1761-1765.

[5] J. Lim, D.H. Chyung, "On a Control Scheme for Two Cooperating Robot Arms," Proc. of 24th Conference on Decision and Control, Fort Lauderdale, Florida, Dec. 11-13, 1985, pp. 334-337.

[6] E. Freund, H. Hoyer, "Collision Avoidance in Multi-Robot Systems," The Second International Symposium of Robotics Research, Kyoto-Kaikan, Kyoto, Japan, Aug. 20-23, 1984, pp. 135-146.

[7] E. Freund, "On the Design of Multi-Robot Systems," International Conference on Robotics, Atlanta, Georgia, March 13-15, 1984, pp. 477-490.

[8] E. Freund, H. Hoyer, "On the On-Line Solution of the Findpath Problem in Multi-Robot Systems," The Third International Symposium of Robotics Research, Gouvieux, France, Oct. 7-11, 1985.

[9] J.K. Salisbury, "Active Stiffness Control of a Manipulator in Cartesian Coordinates," Proc. of 19th IEEE Conference on Decision and Control, Dec. 1980, pp. 95-100.

[10] Daniel E. Whitney, "Historical Perspective and State of the Art in Robot Force Control," Proc. of 1985 IEEE International Conference on Robotics and Automation, St. Louis, Missouri, March 1985, pp. 262-268.

[11] D.E. Whitney, "Force Feedback Control of Manipulator Fine Motions," Transactions of ASME, Journal of Dynamic Systems, Measurement, and Control, June 1977, pp. 91-97.

[12] M. Mason, "Compliance and Force Control for Computer Controlled Manipulators," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-11, No. 6, June 1981, pp. 418-432.

[13] N. Hogan, "Impedance Control: An Approach to Manipulation - Theory, Implementation, and Applications," Journal of Dynamic Systems, Measurement, and Control. Vol. 107, March 1985, pp. 1-24.

[14] H. Hemami and B. Wyman, "Indirect Control of the Forces of Constraint in Dynamic Systems," J. Dynamic Systems, Measurement and Control, Vol. 101, 1979, pp. 355-360.

[15] D.E. Orin and S.Y. Oh, "Control of Force Distribution in Robotic Mechanisms Containing Closed Kinematic Chains," J. Dynamic Systems, Measurement, and Control, Vol. 102, June 1981, pp. 134-141.

[16] T. Ishida, "Force Control in Coordination of Two Arms," Proceedings of the 5th International Joint Conference on Artificial Intelligence, Aug. 1977, pp. 717-722.

[17] R.P. Paul and B. Shimano, "Compliance and Control," Proceedings of the Joint Automatic Control Conference, Purdue University, July 1976, pp. 694-699.

[18] M.H. Raibert and J.J. Craig, "Hybrid Position/Force Control of Manipulators," Transactions of ASME, Journal of Dynamic Systems, Measurement, and Control,Vol. 103, No. 2, June 1981, pp. 126-133.

[19] Harry West and H. Asada, "Task Representation and Constraint Analysis for Hybrid Position/force Controller Design," IFAC Symposium on Robot Control, Barcelona, Spain, Nov. 1985, pp. 311-316.

[20] Harry West and H. Asada, "A Method for the Control of Robot Arms Constrained by Contact with the Environment," Proc. of American Control Conference, Boston, June 1985.

[21] C. Reboulet and A. Robert, "Hybrid Control of a Manipulator Equipped with an Active Compliant Wrist," Preprints of 3rd International Symposium of Robotics Research, Gouvieux, France, Oct. 1985, pp. 76-80.

[22] Oussama Khatib and Joel Burdick, "Motion and Force Control of Robot Manipulators," Proc. of 1986 IEEE International Conference on Robotics and Automation, April 1986, San Francisco, California.

[23] H. Kazerooni, T.B. Sheridan, and P.K. Houpt, "Robust Compliant Motion for Manipulators, Part I: The Fundamental Concepts of Compliant Motion; Part II: Design Method," IEEE Journal of Robotics and Automation, Vol. RA-2, No. 2, June 1986, pp. 83-105.

[24] Samad Hayati, Hybrid Position/Force Control of Multi-Arm Cooperating Robots," Proc. of 1986 IEEE International Conference on Robotics and Automation, April 1986, San Francisco, California, pp. 82-89.

[25] T. Yoshikawa, "Dynamic Hybrid Position/Force Control of Robot Manipulators - Description of hand Constraints and Calculation of Joint Driving Force," Proc. of 1986 IEEE International Conference on Robotics and Automation, San Francisco, California, April 1986, pp. 1393-1398.

[26] T.J. Tarn, A.K. Bejczy and X. Yun, "Dynamic Coordination of Two Robot Arms," Proceedings of the 25th IEEE Conference on Decision and Control, Athens, Greece, December 10-12, 1986.

[27] T.J. Tarn, A.K. Bejczy and X. Yun, "Coordinated Control of Two Robot Arms," Proceedings of the 1986 International Conference on Robotics and Automation, San Francisco, April, 7-10, 1986.

[28] George N. Sandor and Arthur G. Erdman, Advanced Mechanism Design: Analysis and Synthesis, Vol. 2, Prentice Hall, 1984.

[29] A.K. Bejczy, "Robot Arm Dynamics and Control," JPL Technical Memorandum 33-669, 1974.

[30] Richard P. Paul, "Robot Manipulators: Mathematics, Programming, and control," The MIT Press, 1981.

[31] Y. Chen, Nonlinear Feedback and Computer Control of Robot Arms, D.Sc. Dissertation, Washington University, St. Louis, Missouri, Dec. 1984.

[32] T.J. Tarn, A.K. Bejczy, A. Isidori and Y. Chen, "Nonlinear Feedback in Robot Arm Control," Proceedings of the 23rd IEEE Conference on Decision and Control, Las Vegas, December 12-14, 1984.

[33] M.B. Leahy, Jr., L.M. Nugent, K.P. Valavanis and G.N. Saridis, "Efficient Dynamics for a PUMA 600," Proceedings of the 1986 IEEE Inter. Conf. on Robotics and Automation, San Francisco, April 7-10, 1986.

ELLIPSOIDAL APPROXIMATIONS IN PROBLEMS OF CONTROL

István Vályi
International Institute for Applied Systems Analysis
A-2361 Laxenburg
Austria

Abstract

The subject of study in this paper is an adaptive control problem involving uncertainties. It is a special case of the one considered in the paper by Kurzhanski [1], in the present volume. The system is described by differential inclusions and, accordingly, its solution, a feedback control ensuring that certain feasibility constraints be fulfilled, is sought in the form of a set valued map. We apply recent results of ellipsoidal calculus to develop an easily implementable algorithm that gives approximations to the known exact formulae. The paper is therefore an attempt to carry out the program proposed in the above mentioned article.

1. Introduction

General convex sets are difficult to handle because their analytical description involves an infinite number of scalar parameters. In contrast to this, the family of ellipsoids can be identified by the coordinates of their center and a positive definite matrix representing their "shape". Ellipsoids are well suited for using as approximates of compact convex sets for the reason that many operations over convex sets can be followed in a relatively easy way by operations over their estimating ellipsoids. The idea was first used in the late sixties for estimating the propagation of numerical errors by Faddeev and Faddeeva [2] and in the study of uncertain dynamical systems by Schweppe [3]. After a decade without much activity in the field, new results have been obtained by Kurzhanski, Chernousko and others, an indication of renewed interest. Now, in addition to the known ellipsoidal approximations for the reachable sets of nonconstrained linear systems [4], [5], [6], [7], analogous results for both reachable sets and viable domains are available in the constrained case.

As indicated in the abstract, the solution of the problem that we shall consider is known, i. e. formulae are given for the computation of the support function of the control at each instant. The calculations involved are, however, very complex. (See also Kurzhanski and Nikonov [8]). Our aim here is to obtain an approximate solution in a simpler, and more constructive way. This is done through two steps. The first is to change to a surrogate problem in order to get rid of infinite operations involved in the original construction, and the second is to approximate the solution of this problem with the intersection of a finite number of ellipsoids.

Accordingly, we consider the differential inclusion

$$\dot{p}(t)=C(t)p(t)+u(t) \qquad t\in\mathbf{T}=[t_0,t_1] \tag{1.1}$$

with the initial condition

$$p(t_0)\in\mathbf{P}^{(0)} \tag{1.2}$$

and the constraint on the controls of the form

$$u(t)\in\mathbf{V}(t) \qquad t\in\mathbf{T}\ .$$

Additionally, we require first that a viability condition of the form

$$p(t)+\mathbf{Q}[t]\subset\mathbf{K}(t) \qquad t\in\mathbf{T}^* \tag{1.3}$$

is met, with $\mathbf{T}^*\subset\mathbf{T}$ being finite, i. e.

$$\mathbf{T}^* = \{\ \tau_i\in\mathbf{T} : i\in\overline{1,r}\ \}$$

and $\mathbf{Q}[t]\subset\mathbf{R}^n$, $t\in\mathbf{T}$ consisting of all the values $q(t)\in\mathbf{R}^n$ that are compatible with incoming measured information represented by the function

$$y:\mathbf{T}\ \rightarrow\ \mathbf{R}^m\ .$$

As information arrive in real time, at the instant $t\in\mathbf{T}$, only the function

$$y_t\ :\ [t_0,t]\bigcap\mathbf{T}\ \rightarrow\ \mathbf{R}^m$$

$$y_t(\tau)=y(\tau)$$

is available. The variable q is defined by:

$$\dot{q}(t)\in A(t)q(t)+\mathbf{P}(t) \qquad t\in\mathbf{T} \tag{1.4}$$

$$q(t_0)\in\mathbf{Q}^{(0)} \tag{1.5}$$

$$y(t)\in G(t)q(t)+\mathbf{R}(t) \qquad t\in\mathbf{T}. \tag{1.6}$$

The family of measurements $y(t)\in\mathbf{R}^m$, $t\in\mathbf{T}$ that are compatible with the system (1.4), (1.5) and (1.6) will be denoted by $\mathbf{Y}$.

Second, we also want that the trajectory arrives to a given set at the final instant:

$$p(t_1)\in\mathbf{M}\ . \tag{1.7}$$

We suppose that the mappings

$$C:\mathbf{T}\rightarrow\mathbf{R}^{n\times n}$$

$$\mathbf{V} : \mathbf{T} \to conv\mathbb{R}^n$$

$$\mathbf{K} : \mathbf{T}^* \to conv\mathbb{R}^n$$

$$A : \mathbf{T} \to \mathbb{R}^{n\times n}$$

$$\mathbf{P} : \mathbf{T} \to conv\mathbb{R}^n$$

$$G : \mathbf{T} \to \mathbb{R}^{n\times m}$$

$$\mathbf{R} : \mathbf{T} \to conv\mathbb{R}^m$$

are continuous. The sets $\mathbf{P}^{(0)}$, $\mathbf{M} \in conv\mathbb{R}^n$ and $\mathbf{Q}^{(0)} \in conv\mathbb{R}^m$ are nonvoid elements of the metric space of convex, compact sets defined by the Hausdorff metric h.

The next section deals with ellipsoidal calculus, that is, among others gives some ellipsoidal estimates of the Minkowski sum, the geometric difference and finally the Riemannian integral of ellipsoids. Later these results are applied to find ellipsoidal estimates for the reachable set and viable domain of constrained linear systems. Finally, we return to the above problem, giving an exact definition of the control problem, and then we use the results of preceeding sections to construct a solution. Proofs of the statements are generally omitted, because of limited space, except for the main theorem. A comprehensive publication about the subject is forthcoming.

The author would like to express his gratitude to Academician A. B. Kurzhanski for his guidance and encouragement.

2. Ellipsoidal calculus

We start with a brief overview of the most important notions and facts that we shall rely on.

We represent ellipsoids defining them through their support function, (Rockafellar [9]). The support function of the convex set $\mathbf{H} \in conv\mathbb{R}^n$ will be denoted by $\rho(\cdot|\mathbf{H})$ and the distance function by $d(.,\mathbf{H})$. We shall also use the seminorm of sets in $conv\mathbb{R}^n$ defined as

$$||\mathbf{H}|| = ||\rho(\cdot|\mathbf{H})||_\infty \, .$$

$S(0,\epsilon) \subset \mathbb{R}^n$ denotes the closed unit ball and $C(\mathbf{T}',\mathbb{R}^{n\times n})$ the family of continuous, matrix valued functions over the set $\mathbf{T}' \subset \mathbf{T}$. For other notions related to set valued functions the book by Castaing and Valadier [10] can be used as a reference.

Definition 2.1.

Let $a \in \mathbf{R}^n$ stand for the center, and a symmetric positive semidefinite matrix, $Q \in \mathbf{R}^{n \times n}$, represent the 'shape' of the ellipsoid $\mathbf{E} = \mathbf{E}(a,Q)$, i. e.

$$\rho(l\,|\mathbf{E}) = <a,l> + <Ql,l>^{\frac{1}{2}} \qquad \forall\ l \in \mathbf{R}^n$$

$$\mathbf{E}(a,Q) = \{\ x \in \mathbf{R}^n : <x,l> \ \leq \rho(l\,|\mathbf{E}), \forall\ \ l \in \mathbf{R}^n\}$$

For fixed positive definite matrices Q_1 and Q_2, λ_i, $i \in \overline{1,n}$ will denote the eigenvalues of the pencil of matrices $Q_1 - \lambda Q_2$ i. e. the set of solutions of the equation $det(Q_1 - \lambda Q_2) = 0$. λ_{min} and $\lambda_{max} \in \mathbf{R}$ will stand for the minimal and maximal eigenvalue. We refer the reader to Gantmacher |10| for these notions and the properties of pencils of matrices.

The basic operations over ellipsoids that will be considered are the following.

The first is the Minkowski-sum of sets given by the formula

$$\mathbf{H}_1 + \mathbf{H}_2 = \{\ h_1 + h_2 \in \mathbf{R}^n : h_1 \in \mathbf{H}_1,\ h_2 \in \mathbf{H}_2\ \}.$$

Besides this, we shall have to deal with the difference of sets. The family of convex sets not being closed under forming differences, we need the following definitions (Pontriagin [12], Nurminski and Uriasiev |13|) for an "internal" and an "external" operation:

Definition 2.2.

Consider the convex sets $\mathbf{H}_1$, $\mathbf{H}_2 \subset \mathbf{R}^n$ and suppose that there exists a $h \in \mathbf{R}^n$ such that

$$\{\ h\ \} + \mathbf{H}_1 \supset \mathbf{H}_2\ .$$

We define the geometric (or internal) difference $\mathbf{H}_1 \dot{-} \mathbf{H}_2 \subset \mathbf{R}^n$ as

$$\mathbf{H}_1 \dot{-} \mathbf{H}_2 = \{\ h \in \mathbf{R}^n : h + h_2 \subset \mathbf{H}_1 \quad \forall\ h_2 \subset \mathbf{H}_2\ \}.$$

The result of the other (external) operation is not unique: The family Θ consisting of sets in $\mathbf{R}^n$, is an external difference, if

$$inf\ \{\ \rho(l\,|\mathbf{H}) \in\ \mathbf{R} : \mathbf{H} \in \Theta\ \} = \rho(l,|\mathbf{H}_1)\ -\ \rho(l,|\mathbf{H}_2) \qquad \forall\ \ l \in \mathbf{R}^n$$

One example of such an external difference is the following:

$$\Theta_0 = \{\ \mathbf{H} \subset\ \mathbf{R}^n : \mathbf{H} + \mathbf{H}_2 \supset \mathbf{H}_1,\ \mathbf{H} \subset conv\,\mathbf{R}^n\ \}.$$

We want to construct internally estimating ellipsoids for the geometric difference of the ellipsoids $\mathbf{E}_1 = \mathbf{E}(a_1,Q_1)$ and $\mathbf{E}_2 = \mathbf{E}(a_2,Q_2)$, and externally estimating ellipsoids for some of their external difference. Keeping in mind the essentially different meaning of the word "difference" for internal and external estimates, we shall use the following definition and terminology for ellipsoidal estimates.

Definition 2.3.

The ellipsoid $\mathbf{E}_+$ is an external estimate of the difference of $\mathbf{E}_1=\mathbf{E}(a_1,Q_1)$ and $\mathbf{E}_2=\mathbf{E}(a_2,Q_2)$, if $\mathbf{E}_+ + \mathbf{E}_2 \supset \mathbf{E}_1$ and the ellipsoid $\mathbf{E}_-$ is an internal estimate, if $\mathbf{E}_- + \mathbf{E}_2 \subset \mathbf{E}_1$.

This definition means, in terms of general sets on one hand, that $\mathbf{E}_+$ can be a member of a family of sets that form an external difference for $\mathbf{E}_1$ and $\mathbf{E}_2$ and, on the other hand, that $\mathbf{E}_-$ is a subset of the geometric difference.

In relation to continuous systems we shall have to consider the continuous sum of a family of ellipsoids, i. e. their integral.

Definition 2.4.

Let us consider the family of ellipsoids $\mathbf{E}(a(t),Q(t))\subset\mathbf{R}^n$, $t\in\mathbf{T}$. the ellipsoid $\mathbf{E}(a_0,Q_0)\subset\mathbf{R}^n$ and suppose that the functions $a:\mathbf{T}\rightarrow\mathbf{R}^n$, $Q:\mathbf{T}\rightarrow\mathbf{R}^{n\times n}$ are Riemann-integrable. The integral $\mathbf{X}(t_1)\subset\mathbf{R}^n$ of the above ellipsoids is defined as

$$\mathbf{X}(t_1) = \{\, x\in\mathbf{R}^n : x=x(t_0)+\int_{t_0}^{t_1} u(t)dt,\ x(t_0)\in\mathbf{E}(a_0,Q_0),\ u(t)\in\mathbf{E}(a(t),Q(t)),\ t\in\mathbf{T}\,\} \ .$$

As is well known, $\mathbf{X}(t_1)\subset\mathbf{R}^n$ is the limit, in the metrics of Hausdorff-distance h, of the sets corresponding to the Riemannian approximating sums, or in other words, sums of a finite number of ellipsoids. This means the pointwise convergence of the respective support functions. That is

$$\rho(l,|\mathbf{X}(t_1))=<a_0,l>+\int_{t_0}^{t_1}<a(t),l>\,dt+<Q_0l,l>^{\frac{1}{2}}+\int_{t_0}^{t_1}<Q(t)l,l>^{\frac{1}{2}}\,dt \qquad \forall\ l\in\mathbf{R}^n.$$

Finally we recall the way ellipsoids are transformed by affine transformations:

$$x\in\mathbf{E}(a,Q) \text{ if and only if } Ax+b\in\mathbf{E}(Aa+b,AQA').$$

Now we state a simple symmetry related property of internally and externally estimating ellipsoids.

Proposition 2.1.

Suppose that for the convex set $\mathbf{H}\subset\mathbf{R}^n$ we have $\mathbf{H}=-\mathbf{H}$. Then $\mathbf{H}\subset\mathbf{E}(a,Q)$ implies $\mathbf{H}\subset\mathbf{E}(0,Q)$ and $\mathbf{H}\supset\mathbf{E}(a,Q)$ implies $\mathbf{H}\supset\mathbf{E}(0,Q)$.

Let us introduce now some simple formulae for estimating the Minkowski sum and the "difference" of two ellipsoids from both sides.

Proposition 2.2.

Let us consider the ellipsoids $\mathbf{E}_1=\mathbf{E}(a,Q_1)$ and $\mathbf{E}_2=\mathbf{E}(a,Q_2)$ and use the following notation:

$$Q(\mu)=(1+\mu^{-1})Q_1+(1+\mu)Q_2, \quad \mu\in\mathbf{R}\setminus\{0\}\ .$$

Then for any $\mu\in(0,\infty)$,

(a) the ellipsoid $\mathbf{E}=\mathbf{E}(a_1+a_2,Q(\mu))$ is an external ellipsoidal approximation of the Minkowski-sum $\mathbf{E}_1+\mathbf{E}_2$, i. e.

$$\mathbf{E}\supset\mathbf{E}_1+\mathbf{E}_2 \quad \forall\ \mu\in(0,\infty)$$

and if we select

$$\mu=\frac{Tr^{\frac{1}{2}}(Q_1)}{Tr^{\frac{1}{2}}(Q_2)}$$

then this value defines the ellipsoid containing the sum that has minimal trace, or, sum of squares of semiaxes.

(b) if we suppose that $int(\mathbf{E}(0,Q_1))\supset\mathbf{E}(0,Q_2)$ holds then $\mathbf{E}=\mathbf{E}(a_1-a_2,Q(-\mu))$ is an internal ellipsoidal approximation of the difference of $\mathbf{E}_1$ and $\mathbf{E}_2$, i. e.

$$\mathbf{E}+\mathbf{E}_2\subset\mathbf{E}_1 \quad \forall\ \mu\in[1,\lambda_{min}]$$

where the choice of

$$\mu=min\left\{\frac{Tr^{\frac{1}{2}}(Q_1)}{Tr^{\frac{1}{2}}(Q_2)},\lambda_{min}\right\}$$

produces the ellipsoid of maximal trace contained in the difference.

Proposition 2.3.

(a) The ellipsoid $\mathbf{E}(a_1+a_2,[Q_1^{\frac{1}{2}}+Q_2^{\frac{1}{2}}]^2)$ is an internal estimate of the Minkowski-sum $\mathbf{E}_1+\mathbf{E}_2$.

(b) Suppose that $int(\mathbf{E}_1)\supset\mathbf{E}_2$ holds. Then the ellipsoid $\mathbf{E}(a_1-a_2,[Q_1^{\frac{1}{2}}-Q_2^{\frac{1}{2}}]^2)$ is an external estimate of the difference of $\mathbf{E}_1$ and $\mathbf{E}_2$.

Let us consider now the problem of finding external ellipsoidal estimates for the sum of more than two ellipsoids. This will serve as the basis for handling the integral i. e. the continuous sum of an infinite number of ellipsoids.

Let the ellipsoids, $\mathbf{E}_i=\mathbf{E}(a_i,Q_i)$, $i\in\overline{1,r}$ be given and denote their Minkowski-sum by $S(r)\subset\mathbf{R}^n$. Let us, in addition, consider the family of ellipsoids $\mathbf{E}(b(t),F(t))\subset\mathbf{R}^n$, $t\in\mathbf{T}$, the ellipsoid $\mathbf{E}(b_0,F_0)\subset\mathbf{R}^n$ and their integral, denoted by $\mathbf{X}(t_1)\subset\mathbf{R}^n$. Here we suppose that the functions $b:\mathbf{T}\to\mathbf{R}^n$, $F:\mathbf{T}\to\mathbf{R}^{n\times n}$ are Riemann-integrable.

Definition 2.5.

The family $\Sigma(r)$ is called the set of linear external approximations of the set $\mathbf{S}(r)=\sum_{i=1}^{r}\mathbf{E}_i$, where

$$\Sigma(r) = \{ \mathbf{E}=\mathbf{E}(a,Q)\subset\mathbf{R}^n : a=\sum_{i=1}^{r}a_i,\ Q=\sum_{i=1}^{r}x_i\sum_{j=1}^{r}\frac{Q_j}{x_j},\ x_i\in(0,\infty),\ \forall i\in\overline{1,r}\ \}$$

The ellipsoid $\bar{\mathbf{E}}=\mathbf{E}(\bar{a},\bar{Q})\in\Sigma(r)$ is a linear trace-minimal external estimate of the sum $\mathbf{S}(r)\subset\mathbf{R}^n$, if $Tr(\bar{Q})\leq Tr(Q)$, for each $\mathbf{E}=\mathbf{E}(a,Q)\in\Sigma(r)$, and it is a linear trace-minimal external estimate of the integral $\mathbf{X}(t_1)\subset\mathbf{R}^n$ if it is a limit of linear trace-minimal external estimates, in Hausdorff distance, of the Riemannian approximating sums, if this limit exists.

Here we should like to point out that there are simple examples even for $r=3$ showing that the family $\Sigma(r)$ does not contain all the inclusion minimal external ellipsoidal estimates of the sum $\mathbf{S}(r)\subset\mathbf{R}^n$.

Proposition 2.4.

Let $\mathbf{E}(a,Q)\subset\mathbf{R}^n$ be the linear trace-minimal estimate of the Minkowski-sum $\mathbf{S}(r)\subset\mathbf{R}^n$, where $a\in\mathbf{R}^n$ and $Q\in\mathbf{R}^{n\times n}$ are defined by (2.2) and (2.3). Then for the parameters $x_i\in(0,\infty)$, $i\in\overline{1,r}$ we have, up to a constant multiplier,

$$x_k=Tr^{\frac{1}{2}}(Q_k) \qquad k\in\overline{1,r}\ .$$

Theorem 2.1.

Suppose that the functions $b:\mathbf{T}\rightarrow\mathbf{R}^n$, $F:\mathbf{T}\rightarrow\mathbf{R}^{n\times n}$ are Riemann-integrable. Let further $g_0\in(0,\infty)$ and the function $g:\mathbf{T}\rightarrow\mathbf{R}$ be defined with

$$g_0 = Tr^{\frac{1}{2}}[F_0]$$

and

$$g(t) = Tr^{\frac{1}{2}}[F(t)]$$

Then the linear Tr-minimal external estimate of the integral $\mathbf{X}(t_1)\subset\mathbf{R}^n$ is well defined and is an external estimate. It is of the form $\mathbf{E}(a(t_1),Q(t_1))\subset\mathbf{R}^n$ where

$$a(t_1)=b_0+\int_{t_0}^{t_1}b(t)\,dt\ ,$$

$$Q(t_1)=\left[g_0+\int_{t_0}^{t_1}g(s)\ ds\right]\left[\frac{F_0}{g_0}+\int_{t_0}^{t_1}\frac{F(t)}{g(t)}\ dt\right]\ .$$

Proposition 2.5.

Let

$$a_{(r)} = \sum_{i=1}^{r} a_i$$

and

$$Q_{(r)}^{\frac{1}{2}} = \sum_{i=1}^{r} Q_j^{\frac{1}{2}}$$

then

$$\mathbf{S}(r) \supset \mathbf{E}(a_{(r)}, Q_{(r)}).$$

Based on the Cauchy formula, the internal and external estimates for the Minkowski-sum and the integral can be used to obtain such estimates for the reachable sets of linear differential inclusions. These estimates are not the best, i. e. least external or greatest internal, but, as their parameters can be defined as solutions of some nonlinear differential equations, retain the semi-group property. Of these we formulate one, that relies on Proposition 2.3. together with a simple limit argument.

Theorem 2.2.

Consider the problem

$$\dot{x}(t) \in A(t)x(t) + \mathbf{E}(\bar{p}(t), \bar{P}(t)) \qquad t \in T \tag{2.1}$$

with the initial condition

$$x(t_0) \in \mathbf{E}(\bar{x}^0, \bar{X}^0)$$

where the mappings $A(.)$, $\bar{P}(.)$ and $\bar{p}(.)$ are Riemann integrable.

Then the trajectory of ellipsoids $\mathbf{E}(a(t), Q(t)) \subset \mathbf{R}^n$ is an internal estimate of the reachable set $\mathbf{X}[t] \subset \mathbf{R}^n$, for each $t \subset T$ where the parameters of the ellipsoids are defined by the initial value problem

$$\dot{a}(t) = A(t)a(t) + \bar{p}(t) \qquad t \in T$$

$$a(t_0) = \bar{x}^0$$

$$\dot{Q}(t) = A(t)Q(t) + Q(t)A(t) + Q^{\frac{1}{2}}(t)\bar{P}^{\frac{1}{2}}(t) + \bar{P}^{\frac{1}{2}}(t)Q^{\frac{1}{2}}(t) \qquad t \in T$$

$$Q(t_0) = \bar{X}^0$$

and, consequently, for each $s < t$ we also have

$$\mathbf{E}(a(t), Q(t)) \subset X(t,s)\mathbf{E}(a(s), Q(s)) + \int_s^t X(t,\xi)\mathbf{E}(\bar{p}(\xi), \bar{P}(\xi))\, d\xi \,.$$

Here $\mathbf{X}(.,.)$ denotes the fundamental system associated to (2.1).

Finally let us quote a formula for the ellipsoid containing the intersection of a finite number of ellipsoids. (Exercise 5.13. in Schweppe |2|).

Proposition 2.6.

$$\mathbf{E}(a,Q) \supset \bigcap\{ \mathbf{E}(a_i,Q_i)\subset \mathbf{R}^n : i\in \overline{1,r} \}$$

where

$$Q^{-1} = \sum_{i=1}^{r} \alpha_i Q_i^{-1}$$

$$Q^{-1}a = \sum_{i=1}^{r} \alpha_i Q_i^{-1} a_i$$

and

$$\sum_{i=1}^{r} \alpha_i = 1, \qquad \alpha_i \geq 0.$$

3. Dynamical systems with phase constraints

In this section we deal with such continuous dynamical systems where a restriction on the solution trajectories is added. We shall study the reachable sets an the viable domains for such systems. These results are different from those mentioned in the introduction.

We shall consider the differential inclusion

$$\dot{x}(t) \in A(t)x(t) + \mathbf{P}(t) \qquad t \in \mathbf{T} \tag{3.1}$$

with the initial condition

$$x(t_0) \in \mathbf{X}^0 \tag{3.2}$$

or the final condition

$$x(t_1) \in \mathbf{X}^{t_1} \tag{3.3}$$

and we shall require that, additionally, the so called viability condition is met

$$y(t) \in G(t)x(t) + \mathbf{R}(t), \qquad t \in \mathbf{T}' \tag{3.4}$$

where $\mathbf{T}'\subset\mathbf{T}=[t_0,t_1]$ is a Borel measurable set and the mappings

$$A : \mathbf{T} \to \mathbf{R}^{n\times n}$$

$$\mathbf{P} : \mathbf{T} \to conv\mathbf{R}^n$$

$$G : \mathbf{T}' \to \mathbf{R}^{n\times m}$$

$$\mathbf{R} : \mathbf{T}' \to conv\mathbf{R}^m$$

are continuous and $\mathbf{X}^0$, $\mathbf{X}^{t_1}\in conv\mathbf{R}^n$ nonvoid.

Let us fix $\epsilon\geq 0$, and consider the attainability set defined by the ϵ-viable trajectories of the form

$$\mathbf{X}_{\mathbf{T}',\epsilon}(s)=\{\ x(s)\in\ \mathbf{R}^n :$$

$$\dot{x}(t)\in A(t)x(t)+\mathbf{P}(t),\ t\in[t_0,s],$$

$$y(t)\in G(t)x(t)+\mathbf{R}(t)+\mathbf{S}(0,\epsilon),\ t\in\mathbf{T}'\cap[t_0,s],$$

$$x(t_0)\in\mathbf{X}^0\ \}\ .$$

The special case $\epsilon=0$ gives us the usual notion of attainability set, $\mathbf{X}_{\mathbf{T}',0}(s)=\mathbf{X}_{\mathbf{T}'}(s)$. To denote the reachable set in the absence of viability constraints we shall write $\mathbf{X}_{\emptyset}(s)$.

We define viable domains in a similar way:

$$\mathbf{X}^-{}_{\mathbf{T}',\epsilon}(s)=\{\ x(s)\in\ \mathbf{R}^n :$$

$$\dot{x}(t)\in A(t)x(t)+\mathbf{P}(t),\ t\in[s,t_1],$$

$$y(t)\in G(t)x(t)+\mathbf{R}(t)+\mathbf{S}(0,\epsilon),\ t\in\mathbf{T}'\cap[s,t_1],$$

$$x(t_1)\in\mathbf{X}^{t_1}\ \}\ .$$

For $\epsilon=0$ we obtain the usual notion of viable set, $\mathbf{X}^-{}_{\mathbf{T}',0}(s)=\mathbf{X}^-{}_{\mathbf{T}'}(s)$. In the absence of phase constraints we shall use the notation $\mathbf{X}^-{}_{\emptyset}(s)\subset\mathbf{R}^n$.

We start with a formula describing the reachable set $\mathbf{X}_{\mathbf{T}}(t)$ quoting a slight modification of Lemma 2.1. and Theorem 5.1. from the paper by Kurzhanski and Filippova [14]. (This means that now the case $\mathbf{T}'=\mathbf{T}$ is under study.)

Theorem 3.1.

Using the above notations, the following inclusion is true

$$\mathbf{X}_{\mathbf{T}}(t)\subset\bigcap\{\ \mathbf{H}(t,M)\subset\mathbf{R}^n : M\in\mathbf{C}(\mathbf{T},\mathbf{R}^{m\times n})\ \}.$$

where

$$\mathbf{H}(t,M)=H_M(t,t_0)\mathbf{X}^0+\int_{t_0}^{t_1}H_M(t_1,t)\mathbf{P}(t)dt+\int_{t_0}^{t_1}M(t)[y(t)-\mathbf{R}(t)]dt$$

$$H_M(t_1,t)=X(t_1,t)-\int_t^{t_1} M(s)G(s)X(s,t)\,ds$$

and $X(.,.)$ is the fundamental matrix associated to the system (3.1), (3.2). Or, alternatively,

$$\mathbf{H}(t_1,M) = \mathbf{Z}(t_1,L)$$

where $\mathbf{Z}(t_1,L)\subset\mathbf{R}^n$ is the reachable set of the unconstrained system defined by (3.2) and

$$\dot{x}(t) \in [A(t)-L(t)G(t)]x(t)+\mathbf{P}(t)+L(t)[y(t)-\mathbf{R}(t)] \qquad t\in T \tag{3.5}$$

with

$$L'(t)=H_M^{-1}(t,t_0)M(t)$$

and here it is sufficient to consider only those M-s, where the above formula is well defined.

The above theorem can be combined directly with the results related to the integral of ellipsoids to produce an estimate of the reachable set $\mathbf{X_T}(t)\subset\mathbf{R}^n$ of the form of intersection of ellipsoids.

Theorem 3.2.

Let us suppose that the sets appearing in the definition of the system (3.1), (3.2) and (3.4) are ellipsoids:

$$\mathbf{X}^0=\mathbf{E}(a^0,Q^0)$$

$$\mathbf{P}(t)=\mathbf{E}(\bar{p}(t),\bar{P}(t)) \qquad t\in T$$

and

$$\mathbf{R}(t)=\mathbf{E}(\bar{r}(t),\bar{R}(t)) \qquad t\in T$$

and introduce the notations:

$$F_0=H_M(t_1,t_0)Q^0H_M'(t_1,t_0)$$

$$F_1(t)=H_M(t_1,t)\bar{P}(t)H_M'(t_1,t) \qquad t\in T$$

and

$$F_2(t)=M(t)\bar{R}(t)M'(t) \qquad t\in T$$

Then for the linear trace-minimal external ellipsoidal estimate $\mathbf{E}(a_M(t_1),Q_M(t_1))\subset\mathbf{R}^n$ of the set $\mathbf{H}(t_1,M)\subset\mathbf{R}^n$ we have the following equations:

$$a_M(t_1)=H_M(t_1,t_0)a^0+\int_{t_0}^{t_1}[H_M(t_1,t)\bar{p}(t)+M(t)\bar{r}(t)]\,dt$$

and

$$Q_M(t_1)=\left[Tr^{\frac{1}{2}}(F_0)+\int_{t_0}^{t_1}\left[Tr^{\frac{1}{2}}|F_1(s)|+Tr^{\frac{1}{2}}|F_2(s)|\right]ds\right]\left[\frac{F_0}{Tr^{\frac{1}{2}}(F_0)}+\int_{t_0}^{t_1}\left[\frac{F_1(t)}{Tr^{\frac{1}{2}}|F_1(t)|}+\frac{F_2(t)}{Tr^{\frac{1}{2}}|F_2(t)|}\right]dt\right]$$

And so

$$\mathbf{X}_T(t) \subset \bigcap \{ \mathbf{E}(a_M(t),Q_M(t))\subset\mathbf{R}^n : M\in\mathbf{C}(T,\mathbf{R}^{m\times n}) \}.$$

The analogous results related to viable domains are based on Lemma 5.1. and Theorem 6.1. of Kurzhanski and Filippova [14] that we quote again:

Theorem 3.3.

$$\mathbf{X}\ {}_T(t_0) \subset \bigcap \{ \mathbf{H}^-(t_0,N)\subset\mathbf{R}^n : N\in\mathbf{C}(T,\mathbf{R}^{m\times n}) \}.$$

where

$$\mathbf{H}^-(t_0,N)=H^-{}_N(t_0,t_1)\mathbf{X}^{t_1}-\int_{t_0}^{t_1} H^-{}_N(t_0,t)\mathbf{P}(t)dt-\int_{t_0}^{t_1} N(t)|y(t)-\mathbf{R}(t)|dt$$

$$H^-{}_N(t_0,t)=X(t_0,t)+\int_{t_0}^{t} N(s)G(s)X(s,t)ds$$

and $X(.,.)$ is the same as above. Or, alternatively,

$$\mathbf{H}^-(t_0,N) = \mathbf{Z}(t_0,L)$$

where $\mathbf{Z}(t_0,L)\subset\mathbf{R}^n$ is the reachable set of the unconstrained system defined by inclusion (3.5) and final condition (3.3) with

$$L'(t)=(H^-{}_N)^{-1}(t,t_0)N(t)$$

and here again it is sufficient to consider only those N-s, where the above formula is well defined.

In the sequel we shall consider only the case when $\mathbf{T}'\ \ \mathbf{T}^*$, i. e. is the finite set of Section 1. By this restriction we are able to apply compactness arguments and reduce to a finite family the sets that are to be intersected in the formulae for the reachable sets or viable domains. The results of Kurzhanski and Filippova quoted in Theorems 3.1. and 3.3. clearly remain valid in this case and therefore the statements related to ellipsoidal estimates, as well. Now the integrals over the set $\mathbf{T}^*$ turn into finite sums.

Corollary 3.1.

$$\mathbf{X}_{\mathbf{T}^*}(t_1) \subset \bigcap \{ \mathbf{H}(t_1,M)\subset \mathbb{R}^n : M\in \mathbf{C}(\mathbf{T}^*,\mathbb{R}^{m\times n}) \}.$$

where

$$\mathbf{H}(t_1,M)=\Pi_M(t_1,t_0)\mathbf{X}^0+\int_{t_0}^{t_1}\Pi_M(t_1,t)\mathbf{P}(t)dt+\sum_{i=1}^{r} M(\tau_i)[y(\tau_i)-\mathbf{R}(\tau_i)] \qquad (3.6)$$

$$\Pi_M(t_1,t)=X(t_1,t)-\sum_{\tau_i\in[t,t_1]} M(\tau_i)G(\tau_i)X(\tau_i,t)$$

Corollary 3.2.

$$\mathbf{X}^-_{\mathbf{T}^*}(t_0)\subset\bigcap\{\mathbf{H}^-(t_0,N)\subset\mathbb{R}^n : N\in\mathbf{C}(\mathbf{T}^*,\mathbb{R}^{m\times n})\}.$$

where

$$\mathbf{H}^-(t_0,N)=\Pi^-_N(t_0,t_1)\mathbf{X}^{t_1}-\int_{t_0}^{t_1}\Pi^-_N(t_0,t)\mathbf{P}(t)dt-\sum_{\tau_i\in[t_0,t_1]}N(\tau_i)[y(\tau_i)-\mathbf{R}(\tau_i)]$$

$$\Pi^-_N(t_0,t)=X(t_0,t)-\sum_{\tau_i\in[t_0,t]}N(\tau_i)G(\tau_i)X(\tau_i,t)$$

For this special case we develop a formula that approximates the sets $\mathbf{X}_{\mathbf{T}^*}(t)\subset\mathbb{R}^n$ and $\mathbf{X}^-_{\mathbf{T}^*}(t)\subset\mathbb{R}^n$, in some sense, with a finite intersection of ellipsoids. Our first proposition describes a larger set, $\mathbf{X}_{\mathbf{T}^*,\epsilon}(t)\subset\mathbb{R}^n$ from the inside, with bounds imposed on the norm of the function M, or $N\in\mathbf{C}(\mathbf{T}^*,\mathbb{R}^{m\times n})$.

Proposition 3.1.

If we keep the notations of Corollary 3.1. then the following inclusion is true:

$$\mathbf{X}_{\mathbf{T}^*,\epsilon}(t)\supset\bigcap\{\mathbf{H}(t,M)\subset\mathbb{R}^n : M\subset\mathbf{C}(\mathbf{T}^*,\mathbb{R}^{m\times n}),\ ||M||_\infty\le C\}$$

where

$$C=\frac{2\cdot||\mathbf{X}_\emptyset(t)||}{\epsilon}$$

Proposition 3.2.

The mapping

$$\mathbf{H}(t_1,.) : \mathbf{C}(\mathbf{T}^*,\mathbb{R}^{m\times n})\to conv\,\mathbb{R}^n$$

defined by equality (3.6) is continuous.

In particular,

$$h(\mathbf{H}(t_1,M),\mathbf{H}(t_1,M_0))\le c\cdot||M-M_0||_\infty$$

where

$$c=||\mathbf{X}^0||\cdot\sum_{i=1}^{r}||G(\tau_i)X(\tau_i,t_0)||_\infty+\int_{T}||\mathbf{P}(t)||\sum_{i=1}^{r}||G(\tau_i)X(\tau_i,t)||_\infty dt+\sum_{i=1}^{r}||\mathbf{R}(\tau_i)|| \tag{3.7}$$

Theorem 3.4.

Let us use the notations of Theorem 3.2. and Corollary 3.1. then for each $\epsilon>0$ there exists a finite set of functions

$$M_j : \mathbf{T}^* \to \mathbf{R}^{m\times n}, \qquad j\in J=J_\epsilon$$

such that

$$\mathbf{X}_{\mathbf{T}^*,\epsilon}(t_1)+\mathbf{S}(0,\epsilon) \supset \bigcap\{\mathbf{H}(t_1,M_j)\subset\mathbf{R}^n : j\in J\}$$

where the set of functions $\{M_j : j\in J\}$ is an ϵ/c-net in the set:

$$\{\ M\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{m\times n}) :\ ||M||_\infty\leq \frac{2\cdot\ ||\mathbf{X}_\emptyset(t_1)||}{\epsilon}\ \}$$

and $c\in\mathbf{R}$ is given by (3.7).

An analogous statement is valid for the viable domain, that we shall formulate in combination with Theorem 2.2.:

Theorem 3.5.

Let us use the notations of Theorem 3.4. and Corollary 3.2., then for each $\epsilon>0$ there exists a finite set of functions

$$N_j : \mathbf{T}^* \to \mathbf{R}^{m\times n}, \qquad j\in J=J_\epsilon$$

such that

$$\mathbf{X}^-{}_{\mathbf{T}^*,\epsilon}(t_0)+\mathbf{S}(0,\epsilon) \supset \bigcap\{\mathbf{H}^-(t_0,N_j)\subset\mathbf{R}^n : j\in J\}$$

where the set of functions $\{N_j\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{m\times n}) : j\in J\}$ is an ϵ/c-net in

$$\{\ N\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{m\times n}) :\ ||N||_\infty\leq \frac{2\cdot\ ||\mathbf{X}^-{}_\emptyset(t_0)||}{\epsilon}\ \}$$

and

$$c=||\mathbf{X}^{t_1}||\cdot\sum_{i=1}^{r}||G(\tau_i)X(\tau_i,t_1)||_\infty+\int_{T}||\mathbf{P}(t)||\sum_{i=1}^{r}||G(\tau_i)X(\tau_i,t)||_\infty dt+\sum_{i=1}^{r}||\mathbf{R}(\tau_i)|| .$$

Further

$$\mathbf{X}^-{}_{\mathbf{T}^*,\epsilon}(t_0)+\mathbf{S}(0,\epsilon) \supset \bigcap\{\mathbf{E}(a_j(t_0),Q_j(t_0))\subset\mathbf{R}^n : j\in J\}$$

where the parameters of the ellipsoids are defined by the solutions to the problem

$$\dot{a}_j(t)=[A(t)-L_j(t)G(t)]a(t)+\bar{p}(t)+L_j(t)[y(t)-\bar{r}(t)] \qquad t\in T$$

$$a(t_1)=a^{t_1}$$

$$\dot{Q}_j(t)=[A(t)-L_j(t)G(t)]Q_j(t)+Q_j(t)[A(t)-L_j(t)G(t)]+$$

$$+Q_j^{\frac{1}{2}}(t)[\bar{P}^{\frac{1}{2}}(t)+(L_j(t)\bar{R}(t)L'_j(t))^{\frac{1}{2}}]+[\bar{P}^{\frac{1}{2}}(t)+(L_j(t)\bar{R}(t)L'_j(t))^{\frac{1}{2}}]Q_j^{\frac{1}{2}}(t) \qquad t\in T$$

$$Q_j(t_1)=Q^{t_1}$$

with the set of generalized functions (distributions)

$$L_j : T \to \mathbf{R}^{m\times n}, \qquad j\in J$$

defined by

$$\int_T L_j'(s)\varphi(s)ds = \sum_{i=1}^{r}(H^-{}_{N_j})^{-1}(t_0,\tau_i)N_j(\tau_i)\varphi(\tau_i) \ .$$

4. Adaptive control

The formal definition for the control can be given as follows. Given the set $\mathbf{M}\subset\mathbf{R}^n$ we look for admissible controls, i. e. functions

$$\mathbf{U} : \mathbf{T}\times\mathbf{R}^n\times\mathbf{Y} \to conv\mathbf{R}^n$$

$$\mathbf{U}(t,p,y_t)\subset\mathbf{V}(t) \qquad t\in\mathbf{T}$$

that are measurable in $t\in\mathbf{T}$ for each fixed $p\in\mathbf{R}^n$ and upper semicontinuous with respect to p. These controls have to meet for almost all $t\in\mathbf{T}$

$$\dot{p}(t)\in C(t)p(t)+\mathbf{U}(t,p(t),y_t)$$

as well as (1.3) and (1.7).

Instead of the above problem, we shall deal with a restriction, where the condition (1.3) is substituted by a stronger one. This is the blunt [1], or raw [8], case. We shall also need the following:

Condition 4.1.

There exists an admissible control $u(t)\in\mathbf{V}(t)$, $t\in\mathbf{T}$ such that for the corresponding trajectory $p(t)\in\mathbf{R}^n$, $t\in\mathbf{T}$, (1.3) and (1.7) holds simultaneously for each measurement $y\in\mathbf{Y}$.

Denote further

$$\mathbf{W}_{\mathrm{b}}[t]=\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\subset\mathbf{R}^n$$

the viable domain at $t\in\mathbf{T}$ of the system constituted by (1.1), (1.7) and the inclusion

$$p(s)\in\mathbf{K}_1(s\,|t) \tag{4.1}$$

holding for each $s\in[t,t_1]\bigcap\mathbf{T}^*$ where

$$\mathbf{K}_1(s\,|t)=\mathbf{K}(s)\dot{-}\mathbf{Q}^*(s\,|t,\mathbf{Q}(t,y_t\,|t_0,\mathbf{Q}^{(0)}))\qquad s\in[t,t_1]\ . \tag{4.2}$$

Here $\mathbf{Q}(t,y_t\,|t_0,\mathbf{Q}^{(0)}))\subset\mathbf{R}^n$ denotes the reachable set of the system (1.4), (1.5) and (1.6) at $t\in\mathbf{T}$, and $\mathbf{Q}^*(s\,|t,\mathbf{Q})\subset\mathbf{R}^n$ the reachable set of the unconstrained system (1.4), at $s\in[t,t_1]$, started from the initial condition $q(t)\in\mathbf{Q}$.

If we fix some $t\in\mathbf{T}$, then we clearly have that (1.3) holds, as obviously

$$\mathbf{K}_1(s\,|t)+\mathbf{Q}[s]\subset\mathbf{K}[s]\qquad\forall\,s\in\mathbf{T},$$

that is usually a proper inclusion. Therefore, if we are able to ensure that (4.1) holds over all $s\in\mathbf{T}^*$, then the original constraint (1.3) also holds.

According to Lemma 4. and Theorem 2. of Kurzhanski and Nikonov [8], we have:

Theorem 4.1.

If Condition 4.1. is valid, then the sets $\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\in conv\mathbf{R}^n$, $t\in\mathbf{T}$ are not void, the mapping $\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})$, $t\in\mathbf{T}$ is measureable in $t\in\mathbf{T}$ and is upper semicontinuous from below in the variable $p\in\mathbf{R}^n$.

Further, if we suppose that

$$p(t_0)\in\mathbf{W}_{\mathrm{b}}(t_0,y_{t_0},\mathbf{M})$$

and define the control $\mathbf{U}_{\mathrm{b}}[t]=\mathbf{U}_{\mathrm{b}}(t,p,y_t)$ with the relation

$$\mathbf{U}_{\mathrm{b}}(t,p,y_t)=\begin{cases}\partial\rho(l,|\mathbf{V}(t)),\text{ with } l\in\partial d(p\,|\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})), & \text{if } p\notin\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\\ \mathbf{V}(t), & \text{if } p\in\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\end{cases} \tag{4.3}$$

then it will be admissible, and for any corresponding trajectory $p(t)\in\mathbf{R}^n$, $t\in\mathbf{T}$ and any admissible measurement $y\in\mathbf{Y}$, the relation

$$p(t)\in\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\qquad t\in\mathbf{T} \tag{4.4}$$

holds, i. e. the control solves the problem.

Let us now describe the procedure, relying on our ellipsoidal calculus, that yields a set $\mathbf{W}_e(t,y_t,\mathbf{M})\subset\mathbf{R}^n$ such that for each $t\in\mathbf{T}$,

$$\mathbf{W}_e(t,y_t,\mathbf{M})\subset\mathbf{W}_b(t,y_t,\mathbf{M})+\mathbf{S}(0,\epsilon)\ . \tag{4.5}$$

The construction follows the definition of the set $\mathbf{W}_b(t,y_t,\mathbf{M})\in conv\,\mathbf{R}^n$ ensuring at each step that the appropriate type of inclusion be valid. We shall suppose that the sets defining the problem are ellipsoids, that means no restriction, as original sets can be substituted by their ellipsoidal estimates. Accordingly we use the following notations:

$$\mathbf{V}(t)=\mathbf{E}(\bar{v}(t),\bar{V}(t)) \qquad t\in\mathbf{T}$$

$$\mathbf{M}=\mathbf{E}(\bar{m},\bar{M})$$

$$\mathbf{P}(t)=\mathbf{E}(\bar{p}(t),\bar{P}(t)) \qquad t\in\mathbf{T}$$

$$\mathbf{R}(t)=\mathbf{E}(\bar{r}(t),\bar{R}(t)) \qquad t\in\mathbf{T}$$

$$\mathbf{Q}^{(0)}=\mathbf{E}(\bar{q}^{(0)},\bar{Q}^{(0)})$$

$$\mathbf{K}(t)=\mathbf{E}(\bar{k}(t),\bar{K}(t)) \qquad t\in\mathbf{T}^*$$

(i) External ellipsoidal estimate for the set $\mathbf{Q}^*(s\,|t,\mathbf{Q}(t,y_t\,|t_0,\mathbf{Q}^{(0)}))\subset\mathbf{R}^n$

By the definition of $\mathbf{Q}^*(s\,|t,\mathbf{Q}(t,y_t\,|t_0,\mathbf{Q}^{(0)}))\subset\mathbf{R}^n$ and Theorem 3.2. we have for an arbitrary set of $M_i\in\mathbf{C}([t_0,t],\mathbf{R}^{m\times n})$, $i\in J_1$ that

$$\mathbf{Q}(t,y_t\,|t_0,\mathbf{Q}^{(0)})\subset\bigcap\{\ \mathbf{E}(a_{M_i}(t),Q_{M_i}(t))\subset\mathbf{R}^n : M_i\in\mathbf{C}([t_0,t],\mathbf{R}^{m\times n}),\ i\in J_1\}$$

where

$$a_M(t)=H_M(t_0)\bar{q}^{(0)}+\int_{t_0}^{t}H_M(s)\bar{p}(s)ds+\int_{t_0}^{t}M(s)[y(s)-\bar{r}(s)]ds$$

and

$$Q_M(t)=$$

$$\left[Tr^{\frac{1}{2}}(F_0)+\int_{t_0}^{t}Tr^{\frac{1}{2}}[F_1(s)]ds+\int_{t_0}^{t}Tr^{\frac{1}{2}}[F_2(s)]ds\right]\left[\frac{F_0}{Tr^{\frac{1}{2}}(F_0)}+\int_{t_0}^{t}\frac{F_1(s)}{Tr^{\frac{1}{2}}[F_1(s)]}ds+\int_{t_0}^{t}\frac{F_2(s)}{Tr^{\frac{1}{2}}[F_2(s)]}ds\right]$$

$$H_M(s)=X_2(t,s)-\int_{s}^{t}M(\xi)G(\xi)X_2(\xi,s)d\xi$$

and

$$F_0=H_M(t_0)\bar{Q}^{(0)}H_M'(t_0)$$

$$F_1(s)=H_M(s)P(s)H_M'(s) \qquad s\in[t_0,t]$$

$$F_2(s)=M(s)\bar{R}(s)M'(s) \qquad s\in[t_0,t]\ .$$

Here the function $X_2 : \mathbf{T}\times\mathbf{T} \rightarrow \mathbf{R}^{n\times n}$ is the fundamental system associated to (1.4).

Let us follow the evolution of each ellipsoid obtained in this way, now without constraints. Then we obtain

$$\mathbf{Q}^*(s\,|\,t,\mathbf{Q}(t,y_t\,|\,t_0,\mathbf{Q}^{(0)}))\subset\bigcap\ \{\ \mathbf{E}(a_{M_i}(s\,|\,t),Q_{M_i}(s\,|\,t))\subset\mathbf{R}^n\ :\ M_i\in\mathbf{C}([t_0,t],\mathbf{R}^{m\times n}),\ i\in J_1\ \}$$

where

$$a_M(s\,|\,t)=X_2(s\,|\,t)a_M(t)+\int_t^s X_2(s,\xi)\bar{p}(\xi)d\xi$$

and

$$Q_M(s\,|\,t)=$$

$$\left[Tr^{\frac{1}{2}}[X_2(s,t)Q_M(t)X_2'(s,t)]+\int_t^s Tr^{\frac{1}{2}}[X_2(s,\xi)\bar{P}(\xi)X_2'(s,\xi)]d\xi\right]\cdot$$

$$\cdot\left[\frac{X_2(s,t)Q_M(t)X_2'(s,t)}{Tr^{\frac{1}{2}}[X_2(s,t)Q_M(t)X_2'(s,t)]}+\int_t^s\frac{X_2(s,\xi)\bar{P}(\xi)X_2'(s,\xi)}{Tr^{\frac{1}{2}}[X_2(s,\xi)\bar{P}(\xi)X_2'(s,\xi)]}d\xi\right]$$

Let us use now Proposition 2.8. to give an external ellipsoidal estimate for the intersection. According to this, if

$$\sum_{i\in J_1}\alpha_i=1$$

$$\alpha_i\geq 0$$

then the ellipsoid $\mathbf{E}(a(s\,|\,t),Q(s\,|\,t))\subset\mathbf{R}^n$ defined by

$$[Q(s\,|\,t)]^{-1}=\sum_{i\in J_1}\alpha_i[Q_{M_i}(s\,|\,t)]^{-1}$$

$$[Q(s\,|\,t)]^{-1}a(s\,|\,t)-\sum_{i\in J_1}\alpha_i[Q_{M_i}(s\,|\,t)]^{-1}a_{M_i}(s\,|\,t)$$

contains the intersection of $\mathbf{E}(a_i(s\,|\,t),Q_i(s\,|\,t))\subset\mathbf{R}^n$, $i\in J_1$. Hence using Proposition 2.2., we have

$$\mathbf{Q}^*(s\,|\,t,\mathbf{Q}(t,y_t\,|\,t_0,\mathbf{Q}^{(0)}))+\mathbf{S}(0,\epsilon)\subset\mathbf{E}(a^\epsilon(s\,|\,t),Q^\epsilon(s\,|\,t)) \qquad (4.6)$$

where

$$a^{\epsilon}(s\,|t)=a(s\,|t)$$

$$Q^{\epsilon}(s\,|t)=\left[Tr^{\frac{1}{2}}Q(s\,|t)+\epsilon s^{\frac{1}{2}}\right]\cdot\left[\frac{Q(s\,|t)}{Tr^{\frac{1}{2}}Q(s\,|t)}+\frac{E}{\epsilon s^{\frac{1}{2}}}\right].$$

(ii) Internal ellipsoidal estimate for the set $\mathbf{K}_1(s\,|t)\subset\mathbf{R}^n$

Condition 4.2.

Suppose that for each $t\in\mathbf{T}$ and $s>t$

$$int(\mathbf{E}(\bar{k}(s),\bar{K}(s))\supset\mathbf{E}(a^{\epsilon}(s\,|t),Q^{\epsilon}(s\,|t))\ .$$

By formulae (4.1), (4.2), (4.6) and Proposition 2.2. we have that under the above condition

$$\mathbf{K}_1(s\,|t)\supset\mathbf{E}(\bar{k}_1(s\,|t),\bar{K}_1(s\,|t))+\mathbf{S}(0,\epsilon)$$

where

$$\bar{k}_1(s\,|t)=\bar{k}(s)-a^{\epsilon}(s\,|t)$$

$$\bar{K}_1(s\,|t)=(1-\kappa^{-1})\bar{K}(s)-(1-\kappa)\,Q^{\epsilon}(s\,|t)$$

$$\kappa=min\left\{\frac{Tr^{\frac{1}{2}}[\bar{K}(s)]}{Tr^{\frac{1}{2}}[Q^{\epsilon}(s\,|t)]},\lambda_{min}(s\,|t)\right\}$$

and $\lambda_{min}(s\,|t)\subset(0,\infty)$ is the minimal eigenvalue of the pencil of matrices $\bar{K}(s)-\lambda Q^{\epsilon}(s\,|t)$.

(iii) The construction for the set $\mathbf{W}_{\mathrm{e}}(t,y_t,\mathbf{M})\subset\mathbf{R}^n$

Our task now is to find an internal estimate for the viable domain $\mathbf{W}_{\mathrm{b}}(t,y_t,\mathbf{M})\subset\mathbf{R}^n$ of the system (1.1), (4.1), (4.2) and (1.7). The previous construction ensures that Theorem 3.5. can be applied here. In addition to the requirements of the theorem, the finite set $\{\ N_j\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{n\times n})\ :\ j\in J_2\ \}$ can be chosen in such a way that:

(i) It does not depend on the actual value of $t\in\mathbf{T}$.

(ii) If it contains the function $N\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{n\times n})$ then it also contains all the functions of the form $N^{(i)}\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{n\times n})$, $i\in\overline{1,r}$ where

$$N^{(i)}(\tau_j)=\begin{cases}0 & \text{if } j<i\\ N(\tau_j) & \text{if } j\geq i\end{cases}\ .$$

(iii) For the functions obtained in the above procedure we have

$$det\,[H^{-}{}_{N}(t,\xi)]\neq 0\qquad\forall\ t\in\mathbf{T},\ \xi\in[t,t_1]\ .$$

The parameters of the ellipsoids at the instant $t \in \mathbf{T}$ are defined by the value at $t \in \mathbf{T}$ of the solutions to the problem

$$\dot{a}_L(s) = [C(s) - L(s)]a_L(s) + \bar{v}(s) + L(s)\bar{k}_1(s\,|t) \qquad s \in [t, t_1]$$

$$a_L(t) = p^{t_1}$$

$$\dot{Q}_L(s) = [C(s) - L(s)]Q_L(s) + Q_L(s)[C(s) - L(s)] +$$

$$+ Q_L^{\frac{1}{2}}(s)[\bar{V}^{\frac{1}{2}}(s) + (L(s)\bar{K}_1(s\,|t)L'(s))^{\frac{1}{2}}] + [\bar{V}^{\frac{1}{2}}(s) + (L(s)\bar{K}_1(s\,|t)L'(s))^{\frac{1}{2}}]Q_L^{\frac{1}{2}}(s) \qquad s \in [t, t_1]$$

$$Q_L(t_1) = Q^{t_1} .$$

Here L is one of the generalized functions (distributions)

$$L_j : \mathbf{T} \to \mathbf{R}^{n \times n}, \qquad j \in J_2 .$$

defined by

$$\int_{\mathbf{T}} L_j'(s)\varphi(s)ds = \sum_{i=1}^{r} (H^-{}_{N_j})^{-1}(t_0, \tau_i)N_j(\tau_i)\varphi(\tau_i) .$$

Completing the construction for the set $\mathbf{W}_e(t, y_t, \mathbf{M}) \subset \mathbf{R}^n$ we are able to formulate the main result of the paper, where we define a control that keeps the trajectories within this set. This means that, although the trajectories will not meet condition (4.1), but as a consequence of (4.5), the relation

$$d(p(t), \mathbf{K}[t] \dot{-} \mathbf{Q}[t]) \leq \epsilon, \qquad \forall\, t \in \mathbf{T}^*$$

will hold.

Of course, it may happen that certain trajectories are not approximated by the above construction. Our technique allows us, however, to do this, as well. Namely, its converse can be given in the special case when (1.4) is substituted with the corresponding difference equation over the set $\{t_0\} \bigcup \mathbf{T}^* \subset \mathbf{T}$. Then, instead of the ϵ-internal estimate in (4.5), we can obtain ϵ-external estimates, i. e. a mapping with the property

$$\mathbf{W}^e(t, y_t, \mathbf{M}) + \mathbf{S}(0, \epsilon) \supset \mathbf{W}_b(t, y_t, \mathbf{M})$$

In that case, an analogously defined control will keep the trajectories within the sets $\mathbf{W}^e(t, y_t, \mathbf{M}) \subset \mathbf{R}^n$, $t \in \mathbf{T}$.

Condition 4.3.

There exists an $\epsilon > 0$ and an admissible control $u(t) \in \mathbf{V}(t)$, $t \in \mathbf{T}$ such that for the corresponding trajectory $p(t) \in \mathbf{R}^n$, $t \in \mathbf{T}$, the relation

$$p(t) \in \mathbf{E}(\bar{k}_1(t\,|t), \bar{K}_1(t\,|t)) \qquad \forall\ t \in \mathbf{T}^*$$

and (1.7) holds simultaneously for each measurement $y \in \mathbf{Y}$.

Theorem 4.2.

Let us suppose that Conditions 4.2. and 4.3. hold. Then the sets $\mathbf{W}_e(t,y_t,\mathbf{M}) \subset \mathbf{R}^n$, $t \in \mathbf{T}$ are not void and the corresponding mapping is measureable in $t \in \mathbf{T}$ and is upper semicontinuous from below in the variable $p \in \mathbf{R}^n$.

If we suppose that

$$p(t_0) \in \mathbf{W}_e(t_0,y_{t_0},\mathbf{M})$$

and define the control $\mathbf{U}_e(t,y_t,\mathbf{M})$ with formula (4.3) after a substitution of $\mathbf{W}_b(t,y_t,\mathbf{M}) \subset \mathbf{R}^n$ by $\mathbf{W}_e(t,y_t,\mathbf{M}) \subset \mathbf{R}^n$. then it will be admissible, and for any corresponding trajectory $p(t) \in \mathbf{R}^n$, $t \in \mathbf{T}$ and any admissible measurement $y \in \mathbf{Y}$, the relation

$$p(t) \in \mathbf{W}_e(t,y_t,\mathbf{M}) \subset \mathbf{W}_b|t| + \mathbf{S}(0,\epsilon) \qquad t \in \mathbf{T}$$

holds, i. e. the control approximately solves the problem.

Proof.

The proof of admissibility for the control $\mathbf{U}_e$ is identical with that of Lemma 4. in Kurzhanski and Nikonov |8|.

In the sequel, we shall use the following notations:

$$\mathbf{W}_e|t| = \mathbf{W}_e(t,y_t,\mathbf{M})$$

$$\mathbf{U}_e[t] = \mathbf{U}_e(t,y_t,\mathbf{M})$$

and $X_B(.,.)$ is the solution of the matrix differential equation

$$\dot{X}_B(t,t_0) = B(t)X_B(t,t_0) \qquad t \in \mathbf{T}$$

$$X_B(t_0,t_0) = E \ .$$

If there is a $t^* \in \mathbf{T}$ such that $p|t^*| \notin \mathbf{W}_e|t^*|$, then it is easy to see that there is even an interval $(\tau_1,\tau_2) \subset \mathbf{T}$ such that

$$p[s] \notin \mathbf{W}_e(s) \qquad s \in (\tau_1,\tau_2) \ .$$

Now, again by the definition of the distance function and that of subgradient, this implies that over all this interval, we shall have

$$<l,u|s|> = \rho(l,|\mathbf{W}_e[s|) \tag{4.7}$$

and

$$u|s| \in \mathbf{V}(s) \ .$$

for the control u that produced the trajectory p.

We shall estimate the increment of the distance

$$\Delta d[t] = d[t+\Delta t]-d[t], \qquad [t,t+\Delta t]\subset(\tau_1,\tau_2)$$

with

$$d[t] = d(p[t],\mathbf{W}_\mathbf{e}[t]) \ .$$

Let us have

$$\Delta d[t] = \Delta_1+\Delta_2$$

$$\Delta_1 = d(p[t+\Delta t],\mathbf{W}_\mathbf{e}[t+\Delta t]) - d(X_C(t+\Delta t,t)p[t],X_C(t+\Delta t,t)\mathbf{W}_\mathbf{e}[t])$$

$$\Delta_2 = d(X_C(t+\Delta t,t)p[t],X_C(t+\Delta t,t)\mathbf{W}_\mathbf{e}[t]) - d(p[t],\mathbf{W}_\mathbf{e}[t])$$

Let us denote now by $l=l_{\Delta t}$ the vector that defines the distance in the first term of Δ_1. Then

$$\Delta_1 \leq <l,X_C(t+\Delta t,t)p[t]> + \int_t^{t+\Delta t} <l,X_C(t+\Delta t,\xi)u[\xi]>d\xi - \rho(l,|\mathbf{W}_\mathbf{e}[t+\Delta t]) -$$

$$-<l,X_C(t+\Delta t,t)p[t]> + \rho(l,|X_C(t+\Delta t,t)\mathbf{W}_\mathbf{e}[t+\Delta t])$$

$$\Delta_1 \leq \int_t^{t+\Delta t} <l,X_C(t+\Delta t,\xi)u[\xi]>d\xi + \Delta\rho$$

As the sets $\mathbf{W}_\mathbf{e}$ are a finite intersection of ellipsoids, their support function is the minimum of the supports of those ellipsoids. Let us select now the function L in such a way that this minimum is achieved in the second term of $\Delta\rho$, and extend it to the interval $[t,t+\Delta t]\subset\mathbf{T}$ with zero values. We selected the set of functions $\{ N_j\in\mathbf{C}(\mathbf{T}^*,\mathbf{R}^{n\times n}) : j\in J_2 \}$ in such a way that this is possible. Then:

$$\Delta\rho \leq \rho(l,|\mathbf{E}(a_L(t+\Delta t),Q_L(t+\Delta t))) - \rho(l,|X_C(t+\Delta t,t)\mathbf{E}(a_L(t+\Delta t),Q_L(t+\Delta t)))$$

By the semigroup property of the ellipsoidal estimates stated in Theorem 2.2., and the special form of L, we have:

$$\Delta\rho \leq \rho(l,|X_{C-L}(t+\Delta t,t)\mathbf{E}(a_L(t),Q_L(t))) -$$

$$- \int_t^{t+\Delta t} <l,X_{C-L}(t+\Delta t,\xi)\,\bar{V}(\xi)>d\xi - \rho(l,|X_C(t+\Delta t,t)\mathbf{E}(a_L(t),Q_L(t)))$$

Again by the special form of L, we have that the two fundamental systems coincide over the interval $[t,t+\Delta t]\subset\mathbf{T}$, and so we obtain that

$$\Delta d[t] \leq \int_t^{t+\Delta t} <-l,X_C(t+\Delta t,\xi)\,\bar{V}(\xi)>d\xi -$$

$$\int_{t}^{t+\Delta t} <-l,X_C(t+\Delta t,\xi)u[\xi]>d\xi \;+\; ||X_C(t+\Delta t,t)-E||_\infty \, d[t] \; .$$

Hence, using formula (4.7),

$$\Delta d[t] \leq C_1 \cdot d[t] \cdot \Delta t + o(\Delta t) \; .$$

This, together with the easy to see relation

$$\Delta d[t] \leq C_2 \cdot \Delta t, \qquad t \in \mathbf{T},$$

ensure that the function $d[.]$ decreases over each interval where it has positive values, and so, the proof is complete.

We should like to point out here that the definition of $\mathbf{U_e}$ involves finding the nearest point map for a finite intersection of ellipsoids that is nothing else than the wellknown optimization problem of minimizing a quadratic function under a finite number of quadratic constraints.

5. References

1. KURZHANSKI, A. B. *Set-Valued Calculus in Problems of Adaptive Control,* Present volume

2. FADDEEV, D. K., FADDEEVA V. N. *Stability in Linear Algebra Problems,* Proc. IFIP Congress 1968, Edinburgh, 1968 Vol. 1: Mathematics, Software, North-Holland, Amsterdam, pp.33-39. (1969)

3. SCHWEPPE F. C. *Uncertain Dynamic Systems,* Prentice Hall, Englewood Cliffs, N. J. (1973)

4. KURZHANSKI, A. B. *Control and Observation under Conditions of Uncertainty,* Nauka, Moscow, (1977)

5. CHERNOUSKO, F. L. *Optimal Guaranteed Estimates of Indeterminacies with the Aid of Ellipsoids, I., II., III.,* Izv. Acad. Nauk SSSR, Tekhn. Kibernetika Nos. 3,4,5 (1980)

6. OVSEEVICH, A. I. *Extremal Properties of Ellipsoids, Approximating Attainability Sets,* Problems of Control and Information Theory, Vol 12, No 1. pp. 1-11.

7. BONIN, V. A. *Guaranteed Estimations of the State of Linear Systems with the Help of Ellipsoids,* in: Evolutionary Systems in Estimation Problems, A. B. Kurzhanski, T. F. Filippova Editors, Sverdlovsk, (1980)

8. KURZHANSKI, A. B., NIKONOV O. I. *On Adaptive Processes in Guaranteed Control,* Izv. Acad. Nauk SSSR, Tekhn. Kibernetika No. 4. pp. 3-15. (1986)

9. ROCKAFELLAR, R. T. *Convex Analysis,* Princeton University Press (1970)

10. CASTAING, C., VALADIER, M. *Convex Analysis and Measurable Multifunctions,* Lecture Notes in Mathematics, Vol. 580, Springer Verlag, (1977)

11. GANTMACHER, F. R. *Matrix Theory,* Vols I-II. Chelsea Publishing Co. New York, (1960)

12. PONTRIAGIN, L. S. *Linear Differential Games of Pursuit,* Mathematicheski Sbornik, Vol 112 (154) No.3 (7), pp. 307-330. (1980)

13. NURMINSKI, E. A., URYASIEV, S. P. *The Difference of Convex Sets,* Doklady AN Ukrainian SSR, Ser. A., No. 1. (1985)

14. KURZHANSKI, A. B., FILIPPOVA, T. F. *On Viable Solutions for Uncertain Systems,* IIASA, CP-86-011, Laxenburg (1986)

This series aims to report new developments in the fields of control and information sciences – quickly, informally and at a high level. The type of material considered for publication includes:

1. Preliminary drafts of monographs and advanced textbooks
2. Lectures on a new field, or presenting a new angle on a classical field
3. Research reports
4. Reports of meetings, provided they are
   a) of exceptional interest and
   b) devoted to a specific topic.

The timeliness of a manuscript is more important than its form, which may be unfinished or tentative. If possible, a subject index should be included. Publication of Lecture Notes is intended as a service to the international scientific and engineering community, in that a commercial publisher, Springer-Verlag, can offer a wider distribution of documents which would otherwise have a restricted readership. Once published and copyrighted, they can be documented in the scientific literature.

Manuscripts

Manuscripts should be no less than 100 and preferably no more than 500 pages in length.
They are reproduced by a photographic process and therefore must be typed with extreme care. Symbols not on the typewriter should be inserted by hand in indelible black ink. Corrections to the typescript should be made by pasting in the new text or painting out errors with white correction fluid. The typescript is reduced slightly in size during reproduction; best results will not be obtained unless on each page a typing area of 18 x 26,5 cm (7 x 10½ inches) is respected. On request, the publisher can supply paper with the typing area outlined. More detailed typing instructions are also available on request.

Manuscripts generated by a word-processor or computerized typesetting are in principle acceptable. However if the quality of this output differs significantly from that of a standard typewriter, then authors should contact Springer-Verlag at an early stage.

Authors of monographs receive 50 free copies; editors of proceedings receive 75 free copies; all authors are free to use the material in other publications.

Manuscripts, preferably in English, should be sent to Professor Dr.-Ing. M. Thoma, Institut für Regelungstechnik, Technische Universität, Appelstrasse 11, D-3000 Hannover, Prof. Dr. A. Wyner, Dept. Head, Communications Analysis Research Dept., AT & T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974, or directly to Springer-Verlag Heidelberg.

Springer-Verlag, Tiergartenstraße 17, D-6900 Heidelberg 1
Springer-Verlag, 175 Fifth Avenue, New York, NY 10010/USA

ISBN 3-540-19019-8
ISBN 0-387-19019-8