پلتفرم اختصاصی
مهندسی کنترل

CONTROL ENGINEERS

Dedicated Control Engineering Platform

Website: www.controlengineers.ir

Instagram:@controlengineers.ir

Telegram:@controlengineers

Discrete-Time Control System Analysis and Design

ACADEMIC PRESS

CONTROL AND DYNAMIC SYSTEMS

Advances in Theory and Applications

Volume 71

CONTRIBUTORS TO THIS VOLUME

TONGWEN CHEN
PATRIZIO COLANERI
OSWALDO L. V. COSTA
PAOLO D'ALESSANDRO
ELENA DE SANTIS
BRUCE A. FRANCIS
JEAN-CLAUDE HENNET
SHERWOOD TIFFANY HOADLEY
O. THOMAS HOLLAND
VIVEK MUKHOPADHYAY
WENDY L. POSTON
ANTHONY S. POTOTZKY
CAREY E. PRIEBE
RICCARDO SCATTOLINI
NICOLA SCHIAVONI
HANNU T. TOIVONEN
CAROL D. WIESEMAN

CONTROL AND DYNAMIC SYSTEMS

ADVANCES IN THEORY AND APPLICATIONS

Edited by

C. T. LEONDES

School of Engineering and Applied Science
University of California, Los Angeles
Los Angeles, California

VOLUME 71: DISCRETE-TIME CONTROL SYSTEM ANALYSIS AND DESIGN

ACADEMIC PRESS

San Diego New York Boston
London Sydney Tokyo Toronto

This book is printed on acid-free paper. ♾

Academic Press, Inc.
A Division of Harcourt Brace & Company
525 B Street, Suite 1900, San Diego, California 92101-4495

United Kingdom Edition published by
Academic Press Limited
24-28 Oval Road, London NW1 7DX

International Standard Serial Number: 0090-5267

International Standard Book Number: 0-12-012771-7

PRINTED IN THE UNITED STATES OF AMERICA
95 96 97 98 99 00 QW 9 8 7 6 5 4 3 2 1

CONTENTS

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Tongwen Chen (1), *Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4*

Patrizio Colaneri (95), *Dipartimento di Elettronica e Informazione, Politecnico di Milano, 32-20133 Milano, Italy*

Oswaldo L. V. Costa (291), *Department of Electronics Engineering, Escola Politíca da Universidade de São Paulo, 05508 900 São Paulo SP, Brazil*

Paolo d'Alessandro (35), *Department of Mathematics, 3rd University of Roma, 00146 Roma, Italy*

Elena De Santis (35), *Department of Electrical Engineering, University of L'Aquila, 67040 Poggio di Roio, L'Aquila, Italy*

Bruce A. Francis (1), *Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M58 1A4*

Jean-Claude Hennet (157), *Laboratoire d'Automatique et d'Analyse des Systèmes du CNRS, F-31077 Toulouse, France*

Sherwood Tiffany Hoadley (263), *Langley Research Center, National Aeronautics and Space Administration, Hampton, Virginia 23681*

O. Thomas Holland (131), *Naval Surface Warfare Center, Dahlgren Division, Dahlgren, Virginia 22448*

Vivek Mukhopadhyay (263), *Langley Research Center, National Aeronautics and Space Administration, Hampton, Virginia 23681*

Wendy L. Poston (131), *Naval Warfare Center, Dahlgren Division, Dahlgren, Virginia 22448*

Anthony S. Pototzky (263), *Lockheed Engineering and Science Co., Hampton, Virginia 23666*

Carey E. Priebe (131), *Naval Surface Warfare Center, Dahlgren Division, Dahlgren, Virginia 22448*

Riccardo Scattolini (95), *Dipartimento di Elettronica e Informazione, Politecnico di Milano, 32-20133 Milano, Italy*

Nicola Schiavoni (95), *Dipartimento di Elettronica e Informazione, Politecnico di Milano, 32-20133 Milano, Italy*

Hannu T. Toivonen (215), *Process Control Laboratory, Department of Chemical Engineering, Abo Akademi University, 20500 Turku/Åbo, Finland*

Carol D. Wieseman (263), *Langley Research Center, National Aeronautics and Space Administration, Hampton, Virginia 23681*

PREFACE

Effective control concepts and applications date back over millennia. One very familiar example of this is the windmill. It was designed to derive maximum benefit from windflow, a simple but highly effective optimization technique. Harold Hazen's 1932 paper in the *Journal of the Franklin Institute* was one of the earlier reference points wherein an analytical framework for modern control theory was established. There were many other notable landmarks along the way, including the MIT Radiation Laboratory Series volume on servomechanisms, the Brown and Campbell book, *Principles of Servomechanisms*, and Bode's book entitled *Network Analysis and Syntheses Techniques*, all published shortly after mid-1945. However, it remained for Kalman's papers of the late 1950s (wherein a foundation for modern state space techniques was established) and the tremendous evolution of digital computer technology (which was underpinned by the continuous giant advances in integrated electronics) for truly powerful control systems techniques for increasingly complex systems to be developed. Today we can look forward to a future that is rich in possibilities in many areas of major significance, including manufacturing systems, electric power systems, robotics, and aerospace systems, as well as many other systems with significant economic, safety, cost, and reliability implications. Thus, this volume is devoted to the most timely theme of "Discrete-Time System Analysis and Design Techniques."

The first contribution to this volume is "H_2-Optimal Control of Discrete-Time and Sampled-Data Systems," by Tongwen Chen and Bruce A. Francis. This contribution presents a state space solution to the discrete H_2 control problem and also presents direct formulas for an H_2-optimal sampled-data control problem with state feedback and disturbance feedforward. The derivations presented in this contribution are new and quite self-contained, with the derived formulas being applicable to the sampled-data problem via the powerful lifting technique, which is described in the latter part of this chapter. As such, this is a most important contribution with which to begin this volume.

The next contribution is "Techniques for Reachability in Input Constrained Discrete Time Linear Systems," by Paolo d'Alessandro and Elena

De Santis. Constraints on the input of a discrete-time system result in constraints (reachability) on the system state. Therefore, this issue is of essential importance in the analysis and design of discrete-time systems. This contribution is an in-depth treatment of the many aspects involved in this essential issue.

The next contribution is "Stabilization, Regulation, and Optimization of Multirate Sampled-Data Systems," by Patrizio Colaneri, Riccardo Scattolini, and Nicola Schiavoni. There are two primary reasons for the importance of multirate digital control in practice. One of these is the fact that, in practice in many diverse applications, sensors and actuators distributed throughout a complex system involve different sampling rates, i.e., multirate sampling. The second reason rests on the fact that the use of multirate and periodically time-varying controllers can significantly improve the closed-loop performance of a sampled-data system in terms of model matching, sensitivity reduction, disturbance rejection, and pole and zero assignment with state feedback. This contribution is an in-depth treatment of these issues, and, as such, is also an essential element of this volume.

The next contribution is "Maximizing the Fisher Information Matrix in Discrete-Time Systems," by Wendy L. Poston, Carey E. Priebe, and O. Thomas Holland. One of the most important aspects of the design and analysis problem for discrete-time systems is that of developing and verifying a material model of the system to which discrete-time control is being applied. One of the most important methods for model verification is the Fisher Information Matrix Technique. This contribution is an in-depth treatment of this technique, including illustrative examples for model development and verification.

The next contribution is "Discrete-Time Constrained Linear Systems," by Jean-Claude Hennet. The existence of hard constraints on state and control variables often generate problems in the practical implementation of control laws. Methods for generating control techniques which avoid state or input (control) saturations and including these aspects in the system design are presented in this contribution. Numerous examples are presented throughout this contribution which illustrate the effectiveness of the techniques presented.

The next contribution is "Digital Control with H_∞ Optimality Criteria," by Hannu T. Toivonen. The limitations of standard continuous and discrete design methods in the treatment of sampled-data control systems have recently led to the development of a robust control theory for sampled-data control systems. This contribution presents the various approaches to the development of robust control systems by means of solving the sampled-data H_∞ control problem. Several major new issues and techniques are also presented in this contribution.

The next contribution is "Techniques in On-Line Performance Evaluation of Multiloop Digital Control Systems and Their Application," by Carol D. Wieseman, Vivek Mukhopadhyay, Sherwood Tiffany Hoadley, and Anthony S. Pototzky. This contribution develops a controller performance eval-

uation (CPE) methodology to evaluate the performance of multivariable digital control systems. The power and utility of the method is exemplified in this contribution through its utilization and validation during the wind-tunnel testing of an aeroelastic model equipped with a digital flutter suppression controller. Through the CPE technique a wide range of sophisticated real-time analysis tools are available for rather complex discrete-time system problems.

The final contribution to this volume is "Impulse Control of Piecewise Deterministic Systems," by Oswaldo L. V. Costa. There is a wide and diverse variety of discrete-time systems where control is taken by intervention; that is, the decision to act or apply control is taken at discrete times. In this contribution the impulse control problem of piecewise deterministic processes (PDPs) is addressed. Powerful computational techniques are presented and illustrated.

The contributors to this volume are all to be highly commended for their contribution to this rather comprehensive treatment of discrete-time system analysis and design techniques. The contributors to this volume have produced a modern treatment of the subject which should provide a unique reference on the international scene for individuals working in many diverse areas for years to come.

This Page Intentionally Left Blank

$\mathcal{H}_2$-Optimal Control of Discrete-Time and Sampled-Data Systems

Tongwen Chen
Dept. of Electrical and Computer Engineering
University of Calgary
Calgary, Alberta
Canada T2N 1N4

Bruce A. Francis
Dept. of Electrical Engineering
University of Toronto
Toronto, Ontario
Canada M5S 1A4

Abstract

This paper gives a complete state-space derivation of the discrete-time $\mathcal{H}_2$-optimal controller. This derivation can be extended to treat a sampled-data $\mathcal{H}_2$ control problem, resulting in a new direct solution to the sampled-data problem. A design example for a two-motor systems is included for illustration.

I. Introduction

A recent trend in synthesizing sampled-data systems is to use the more natural continuous-time performance measures. This brought solutions to several new $\mathcal{H}_2$-optimal sampled-data control problems [1, 2, 3], each reducing to an $\mathcal{H}_2$ problem in discrete time.

Discrete-time $\mathcal{H}_2$ (LQG) theory was developed in the 1970's, see, e.g., [4, 5, 6, 7, 8, 9]. As in the continuous-time case, the discrete optimal controller is closely related to the solutions of two Riccati equations. In [10], the solution to a continuous-time $\mathcal{H}_2$-optimal control problem was rederived using the state-space approach. This

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

gives a clean treatment of the problem and provides compact formulas for the optimal controller. Since complete, general formulas for the discrete optimal controller are not readily available in the literature, we ask the question here, can a state-space treatment be accomplished for discrete-time $\mathcal{H}_2$ problems?

The goal in this paper is twofold: to present a state-space solution to the discrete $\mathcal{H}_2$ control problem and to give direct formulas for an $\mathcal{H}_2$-optimal sampled-data control problem with state feedback and disturbance feedforward. Though the results in the discrete-time case are known in various forms, we believe the derivation is new and quite self-contained, and therefore has some pedagogical value. Moreover, the formulas derived can be applied to the sampled-data problem via the powerful lifting technique [11, 12, 13, 14].

The organization of the paper is as follows. In the next section we collect and prove some preliminary results on Riccati equations; the presentation follows closely that in [10] in continuous time. Section III gives a complete state-space derivation of the discrete-time $\mathcal{H}_2$-optimal control, first via state feedback and disturbance feedforward and then via dynamic output feedback. Section IV presents new direct formulas for a sampled-data $\mathcal{H}_2$ problem using state measurement. In Section V we apply the optimal sampled-data control in Section IV to a two-motor system and compare with the optimal analog control. Finally, concluding remarks are contained in Section VI.

The notation in this paper is quite standard: $\mathbb{C}$ is the complex plane, $\mathbb{D} \subset \mathbb{C}$ is the open unit disk, and $\partial\mathbb{D}$ is the boundary of $\mathbb{D}$, namely, the unit circle. Also, $\mathbb{Z}$ is the set of all integers and $\mathbb{Z}_+$ ($\mathbb{Z}_-$) is the nonnegative (negative) subset of $\mathbb{Z}$. The space $\ell_2(\mathbb{Z}_+)$, or simply ℓ_2, consists of all square-summable sequences, perhaps vector-valued, defined on $\mathbb{Z}_+$. Similarly for $\ell_2(\mathbb{Z})$ and $\ell_2(\mathbb{Z}_-)$. The discrete-time frequency-domain space $\mathcal{H}_2(\mathbb{D})$, or simply $\mathcal{H}_2$, is the Hardy space defined on $\mathbb{D}$. We use $\mathcal{RH}_2$ for the real-rational subspace of $\mathcal{H}_2$. In discrete time, we use λ-transforms instead of z-transforms, where $\lambda = z^{-1}$. If a linear discrete system G has a state-space realization (A, B, C, D), then we denote the transfer matrix $D + \lambda C(I - \lambda A)^{-1}B$

by

$$\hat{g}(\lambda) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

Finally, $\hat{g}^{\sim}(\lambda)$ stands for the transposed matrix $\hat{g}(1/\lambda)'$.

II. Riccati Equation

It is well-known that Riccati equations play an important role in the $\mathcal{H}_2$ optimization problem. The solution of a Riccati equation can be obtained via the stable eigenspace of the associated symplectic matrix if the state transition matrix of the plant is nonsingular. If this matrix is singular, as is the case when the plant has a time delay, then the symplectic matrix is not defined; but we can use the stable generalized eigenspace of a certain matrix pair [9].

Let A, Q, R be real $n \times n$ matrices with Q and R symmetric. Define the ordered pair of $2n \times 2n$ matrices

$$H = (H_1, H_2) := \left(\begin{bmatrix} A & 0 \\ -Q & I \end{bmatrix}, \begin{bmatrix} I & R \\ 0 & A' \end{bmatrix} \right).$$

A pair of matrices of this form is called a *symplectic pair*. (This definition is not the most general one.) Note that if A is nonsingular, then $H_2^{-1}H_1$ is a symplectic matrix.

Introduce the $2n \times 2n$ matrix

$$J := \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

It is easily verified that $H_1 J H_1' = H_2 J H_2'$. Thus the generalized eigenvalues (including those at infinity) for the matrix pair H (i.e., those numbers λ satisfying $H_1 x = \lambda H_2 x$ for some nonzero x) are symmetric about the unit circle, i.e., λ is a generalized eigenvalue iff $1/\lambda$ is [9].

Now we *assume* H has no generalized eigenvalues on $\partial\mathbb{D}$. Then it must have n inside and n outside. Thus the two generalized eigenspaces $\mathcal{X}_i(H)$ and $\mathcal{X}_o(H)$, corresponding to generalized eigenvalues inside and outside the unit circle respectively, both have dimension n. Let us focus on the stable subspace $\mathcal{X}_i(H)$. There exist

$n \times n$ matrices X_1 and X_2 such that

$$\mathcal{X}_i(H) = \mathrm{Im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Then for some stable $n \times n$ matrix H_i,

$$H_1 \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = H_2 \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H_i. \tag{1}$$

Some properties of the matrix $X_1'X_2$ are useful.

Lemma 1 *Suppose H has no eigenvalues on $\partial\mathbf{D}$. Then*

(i) $X_1'X_2$ is symmetric;

(ii) $X_1'X_2 \geq 0$ if $R \geq 0$ and $Q \geq 0$.

Proof Rewrite (1) as two equations:

$$AX_1 = X_1H_i + RX_2H_i \tag{2}$$

$$-QX_1 + X_2 = A'X_2H_i. \tag{3}$$

Part (i) can be derived easily from these two equations (see, e.g., [15]). For part (ii), we define $M := X_1'X_2 = X_2'X_1$ and pre-multiply (2) by $H_i'X_2'$ to get

$$H_i'X_2'AX_1 = H_i'MH_i + H_i'X_2'RX_2H_i. \tag{4}$$

Take transpose of (3) and then post-multiply by X_1 to get

$$-X_1'QX_1 + M = H_i'X_2'AX_1. \tag{5}$$

Thus equations (4) and (5) give

$$H_i'MH_i - M + H_i'X_2'RX_2H_i + X_1'QX_1 = 0.$$

This is a Lyapunov equation in M. Since H_i is stable, the unique solution is

$$M = \sum_{k=0}^{\infty} H_i'^k(H_i'X_2'RX_2H_i + X_1'QX_1)H_i^k,$$

which is ≥ 0 since R and Q are ≥ 0.

■

Now *assume* further that X_1 is nonsingular, i.e., the two subspaces

$$\mathcal{X}_i(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary. Set $X := X_2 X_1^{-1}$. Then

$$\mathcal{X}_i(H) = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix}. \tag{6}$$

Note that the $n \times n$ matrix X is uniquely determined by the pair H (though X_1 and X_2 are not), that is, $H \mapsto X$ is a function. We shall denote this function by *Ric* and write $X = Ric(H)$.

To recap, *Ric* is a function $\mathcal{R}^{2n\times 2n} \to \mathcal{R}^{n\times n}$ that maps H to X, where X is defined by equation (6). The domain of *Ric*, denoted *dom Ric*, consists of all symplectic pairs H with two properties, namely, H has no generalized eigenvalues on $\partial\mathbb{D}$ and the two subspaces

$$\mathcal{X}_i(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary.

Some properties of X are given next.

Lemma 2 *Suppose $H \in$ dom Ric and $X = Ric(H)$. Then*

(i) X is symmetric;

(ii) X satisfies the algebraic Riccati equation

$$A'X(I + RX)^{-1}A - X + Q = 0;$$

(iii) $(I + RX)^{-1}A$ is stable.

Proof Setting $X_1 = I$ and $X_2 = X$ gives (i) from Lemma 1; moreover, (2) and (3) simplify to the following two equations

$$A = (I + RX)H_i \tag{7}$$

$$-Q + X = A'XH_i. \tag{8}$$

If A is nonsingular, so is H_i and then $I + RX$ by (7); if A is singular, by [16] (Lemma 1.5) $I + RX$ is still nonsingular. Hence

$$H_i = (I + RX)^{-1}A. \tag{9}$$

This proves (iii) since H_i is stable. Substitute (9) into (8) to get the Riccati equation. ■

Lemma 2 is quite standard, see, e.g., [9, 15]. The following result gives verifiable conditions under which H belongs to *dom Ric*.

Theorem 1 *Suppose H has the form*

$$H = (H_1, H_2) = \left(\begin{bmatrix} A & 0 \\ -C'C & I \end{bmatrix}, \begin{bmatrix} I & BB' \\ 0 & A' \end{bmatrix} \right)$$

with (A, B) stabilizable and (C, A) having no unobservable modes on $\partial\mathbf{D}$. Then $H \in$ dom Ric and Ric$(H) \geq 0$.

Proof We first show that H has no generalized eigenvalues on the unit circle. Suppose, on the contrary, that $e^{j\theta}$ is a generalized eigenvalue and $\begin{bmatrix} x \\ z \end{bmatrix}$ a corresponding eigenvector; that is,

$$\begin{bmatrix} A & 0 \\ -C'C & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = e^{j\theta} \begin{bmatrix} I & BB' \\ 0 & A' \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}.$$

Write as two equations and re-arrange:

$$(A - e^{j\theta})x = e^{j\theta}BB'z \tag{10}$$

$$e^{j\theta}(A' - e^{-j\theta})z = -C'Cx. \tag{11}$$

Pre-multiply (10) and (11) by $e^{-j\theta}z^*$ and x^* respectively to get

$$\begin{aligned} e^{-j\theta}z^*(A - e^{j\theta})x &= \|B'z\|^2 \\ e^{j\theta}x^*(A' - e^{-j\theta})z &= -\|Cx\|^2. \end{aligned}$$

Take complex-conjugate of the latter equation to get

$$\begin{aligned} -\|Cx\|^2 &= e^{-j\theta} z^*(A - e^{j\theta})x \\ &= \|B'z\|^2. \end{aligned}$$

Therefore $B'z = 0$ and $Cx = 0$. So from (10) and (11)

$$\begin{aligned} (A - e^{j\theta})x &= 0 \\ (A - e^{j\theta})^* z &= 0. \end{aligned}$$

We arrive at the equations

$$\begin{aligned} z^*[A - e^{j\theta} \quad B] &= 0 \\ \begin{bmatrix} A - e^{j\theta} \\ C \end{bmatrix} x &= 0. \end{aligned}$$

By controllability and observability of modes on $\partial\mathbb{D}$ it follows that $x = z = 0$, a contradiction.

Next, we will show that the two subspaces

$$\mathcal{X}_i(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary. As in the proof of Lemma 1 bring in matrices X_1, X_2, H_i to get equations (2) and (3), re-written as below ($R = BB', Q = C'C$):

$$AX_1 = X_1 H_i + BB'X_2 H_i \tag{12}$$

$$-C'CX_1 + X_2 = A'X_2 H_i. \tag{13}$$

We want to show that X_1 is nonsingular, i.e., $\text{Ker}\, X_1 = 0$. First, it is claimed that $\text{Ker}\, X_1$ is H_i-invariant. To prove this, let $x \in \text{Ker}\, X_1$. Pre-multiply (12) by $x'H_i'X_2'$ and post-multiply by x to get

$$x'H_i'X_2'X_1 H_i x + x'H_i'X_2'BB'X_2 H_i x = 0.$$

Note that since $X_2'X_1 \geq 0$ (Lemma 1), both terms on the left are ≥ 0. Thus $B'X_2 H_i x = 0$. Now post-multiply (12) by x to get $X_1 H_i x = 0$, i.e., $H_i x \in \text{Ker}\, X_1$. This proves the claim.

Now to prove that X_1 is nonsingular, suppose on the contrary that $\operatorname{Ker} X_1 \neq 0$. Then $H_i|\operatorname{Ker} X_1$ has an eigenvalue, μ, and a corresponding eigenvector, x:

$$\begin{aligned} H_i x &= \mu x, \\ |\mu| &< 1, \quad 0 \neq x \in \operatorname{Ker} X_1. \end{aligned} \tag{14}$$

Post-multiply (13) by x and use (14):

$$(\mu A' - 1)X_2 x = 0.$$

If $\mu = 0$, then $X_2 x = 0$. Otherwise, since $B'X_2 x = 0$ from $B'X_2 H_i x = 0$ and (14), we have

$$x^* X_2' \begin{bmatrix} A - \frac{1}{\bar{\mu}} & B \end{bmatrix} = 0.$$

Then stabilizability implies $X_2 x = 0$ as well. But if $X_1 x = 0$ and $X_2 x = 0$, then $x = 0$, a contradiction. This concludes the proof of complementarity.

Now set $X := Ric(H)$. By Lemma 1 ($R = BB', Q = C'C, X_1 = I, X_2 = X$), $X \geq 0$.

■

This theorem has various forms in the literature; for example, in [6] similar results were given when the matrix A is nonsingular and in [9] an indirect proof was given that X_1 is nonsingular. Our proof here is along the lines of a continuous-time proof in [17].

III. Discrete-Time Case

This section rederives in a state-space approach the perhaps-known results for a discrete-time $\mathcal{H}_2$-optimal control problem.

We begin with the standard setup shown in Figure 1. We have used dotted lines for discrete signals and will reserve continuous lines for continuous signals. The input ω is standard white noise — zero mean, unit covariance matrix. The problem is to design a K that stabilizes G and minimizes the root-mean-square value of ζ; it can be shown that this is equivalent to minimizing the norm on $\mathcal{H}_2$ of the transfer matrix from ω to ζ.

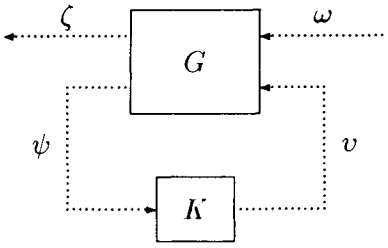


Figure 1: The standard discrete-time setup.

A. State Feedback and Disturbance Feedforward

First we allow the controller to have full information. In this case, as we will see, the optimal controller is a constant state feedback with a disturbance feedforward. With the exogenous input being some pulse function, say, $\omega = \omega_0 \delta_d$ (ω_0 is a constant vector and δ_d the discrete unit pulse), we can even think of v as unconstrained. The precise problem is as follows:

- Given the system equations

$$\begin{aligned} G: \quad \xi(k+1) &= A\xi(k) + B_1\omega(k) + B_2 v(k), \quad \omega = \omega_0 \delta_d \\ \zeta(k) &= C_1\xi(k) + D_{11}\omega(k) + D_{12}v(k) \end{aligned}$$

  with the assumptions

  (i) (A, B_2) is stabilizable;

  (ii) $D_{12}'D_{12} = I$ and D_{12};

  (iii) the matrix

$$\begin{bmatrix} A - \lambda & B_2 \\ C_1 & D_{12} \end{bmatrix}$$

  has full rank $\forall \lambda \in \partial\mathbf{D}$.

- Solve the optimization problem

$$\min_{v \in \ell_{2e}} \|\zeta\|_2^2.$$

Note that for ease of presentation we initially allow v to be in ℓ_{2e}, the *extended* space for ℓ_2; however, the optimal v, to be seen later, will actually lie in ℓ_2. Assumptions (i) and (iii) are mild restrictions and (ii) basically means that the number of outputs to be controlled is no less than the number of control inputs and the control weighting is nonsingular. If $D_{12}'D_{12}$ is nonsingular but not identity, we can normalize it by defining the new v to be $(D_{12}'D_{12})^{1/2}v$.

The setup can be depicted as in Figure 2, where the transfer

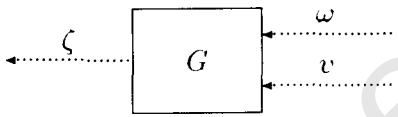


Figure 2: The full-information discrete-time setup.

matrix for G is

$$\hat{g}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \end{array}\right].$$

We will first derive the solution for a special case and then come back to the general one.

1. Orthogonal Case

An additional assumption is now made:

(iv) $D_{12}'C_1 = 0$.

This assumption is an orthogonality condition: It amounts to orthogonality of $C_1\xi$ and $D_{12}v$ in the output ζ.

It follows readily from assumptions (iii) and (iv) that (C_1, A) has no unobservable modes on $\partial\mathbb{D}$. Thus by Theorem 1 the symplectic pair

$$H := \left(\left[\begin{array}{cc} A & 0 \\ -C_1'C_1 & I \end{array}\right], \left[\begin{array}{cc} I & B_2B_2' \\ 0 & A' \end{array}\right]\right)$$

belongs to *dom Ric* and, moreover, $X := Ric(H) \geq 0$. Define the matrices

$$F = -(I + B_2'XB_2)^{-1}B_2'XA$$

$$F_1 = -(I + B_2' X B_2)^{-1}(B_2' X B_1 + D_{12}' D_{11})$$

and the transfer matrix

$$\hat{g}_c(\lambda) = \left[\begin{array}{c|c} A + B_2 F & B_1 + B_2 F_1 \\ \hline C_1 + D_{12} F & D_{11} + D_{12} F_1 \end{array}\right].$$

By Lemma 2, $A + B_2 F$ is stable and so $\hat{g}_c \in \mathcal{RH}_2$.

Theorem 2 *The unique optimal control is* $v_{opt} = F\xi + F_1\omega$. *Moreover,*

$$\min_v \|\zeta\|_2 = \|\hat{g}_c \omega_0\|_2.$$

In contrast with the full-information continuous-time case where the optimal control is a constant state feedback, the discrete-time optimal control law involves a disturbance feedforward term, and this is true even when $D_{11} = 0$.

A useful trick is to change variable [10]. Start with the system equations

$$\begin{aligned} \xi(k+1) &= A\xi(k) + B_1\omega(k) + B_2 v(k) \\ \zeta(k) &= C_1\xi(k) + D_{11}\omega(k) + D_{12} v(k) \end{aligned}$$

and define a new control variable

$$\nu := v - F\xi - F_1\omega.$$

The equations become

$$\begin{aligned} \xi(k+1) &= (A + B_2 F)\xi(k) + (B_1 + B_2 F_1)\omega(k) + B_2\nu(k) \\ \zeta(k) &= (C_1 + D_{12}F)\xi(k) + (D_{11} + D_{12}F_1)\omega(k) + D_{12}\nu(k). \end{aligned}$$

So in the frequency domain

$$\hat{\zeta} = \hat{g}_c\omega_0 + \hat{g}_i\hat{\nu},$$

where $\hat{g}_c$ is as above and $\hat{g}_i$ is seen to be

$$\hat{g}_i(\lambda) = \left[\begin{array}{c|c} A + B_2 F & B_2 \\ \hline C_1 + D_{12}F & D_{12} \end{array}\right].$$

The matrices $\hat{g}_c$ and $\hat{g}_i$ have the following two useful properties:

Lemma 3 *The matrix $\hat{g}_i^{\sim}\hat{g}_c$ belongs to $\mathcal{RH}_2^{\perp}$ and*

$$\hat{g}_i^{\sim}\hat{g}_i = I + B_2' X B_2.$$

Proof To simplify notation, define

$$\begin{aligned} A_F &= A + B_2 F \\ C_{1F} &= C_1 + D_{12} F. \end{aligned}$$

Then we have the power series representations

$$\begin{aligned} \hat{g}_i(\lambda) &= D_{12} + \lambda C_{1F} B_2 + \lambda^2 C_{1F} A_F B_2 + \cdots \\ \hat{g}_c(\lambda) &= (D_{11} + D_{12} F_1) + (\lambda C_{1F} + \lambda^2 C_{1F} A_F + \cdots)(B_1 + B_2 F_1). \end{aligned}$$

Thus

$$\hat{g}_i^{\sim}(\lambda) = D_{12}' + \frac{1}{\lambda} B_2' C_{1F}' + \cdots.$$

Using these formulas, write $\hat{g}_i^{\sim}\hat{g}_c$ as a series in λ, with both positive and negative powers. It remains to check that the coefficients of $\lambda^0, \lambda, \lambda^2, \ldots$ are all zero; this can be proved using the Riccati equation and the definitions of F and F_1.

The proof of the second statement is similar. ■

Proof of Theorem 2 Since v is free in ℓ_{2e}, so is ν. Thus we can formally write in the time domain

$$\begin{aligned} \langle G_c \omega_0 \delta_d, G_i \nu \rangle &= \langle G_i^* G_c \omega_0 \delta_d, \nu \rangle \\ &= 0, \end{aligned}$$

since $\nu \in \ell_{2e}$ and by Lemma 3 $G_i^* G_c \omega_0 \delta_d \in \ell_2(\mathcal{Z}_-)$. Then in the frequency domain we can write

$$\|\hat{\zeta}\|_2^2 = \|\hat{g}_c \omega_0\|_2^2 + \|\hat{g}_i \hat{\nu}\|_2^2.$$

Now note that $\hat{g}_i^{\sim}\hat{g}_i = I + B_2' X B_2$ by Lemma 3 to get

$$\|\hat{\zeta}\|_2^2 = \|\hat{g}_c \omega_0\|_2^2 + \|(I + B_2' X B_2)^{1/2} \hat{\nu}\|_2^2.$$

This equation gives the desired conclusion: The optimal $\hat{\nu}$ is $\hat{\nu} = 0$ (i.e., $v = F\xi + F_1 \omega$) and the minimum norm of ζ equals $\|\hat{g}_c \omega_0\|_2$. ■

With v_{opt} applied, the resultant system is stable since $A + B_2 F$ is stable; thus v_{opt} indeed lies in ℓ_2, as commented before.

2. General Case

Now we return to the situation at the start of Section III-A, without assumption (iv). Our approach is to reduce the problem via a change of variable to one where the orthogonality condition holds.

Define a new control signal

$$v_{new} = v + D_{12}'C_1\xi. \tag{15}$$

Note that v_{new} is a free sequence in ℓ_{2e} if v is. The equivalent system, having ξ as its state vector too, is then shown in Figure 3, where

Figure 3: The equivalent full-information setup.

$$\hat{g}_{new}(\lambda) = \left[\begin{array}{c|cc} A - B_2D_{12}'C_1 & B_1 & B_2 \\ \hline (I - D_{12}D_{12}')C_1 & D_{11} & D_{12} \end{array}\right].$$

The three assumptions made on G at the beginning of Section III-A are also satisfied by G_{new}; for example, assumption (iii) is verified by the following matrix identity:

$$\begin{bmatrix} A - \lambda & B_2 \\ C_1 & D_{12} \end{bmatrix} \begin{bmatrix} I & 0 \\ -D_{12}'C_1 & I \end{bmatrix} = \begin{bmatrix} A - B_2D_{12}'C_1 - \lambda & B_2 \\ (I - D_{12}D_{12}')C_1 & D_{12} \end{bmatrix}.$$

Moreover, G_{new} satisfies the orthogonality condition

$$D_{12}'[(I - D_{12}D_{12}')C_1] = 0.$$

Now invoke Theorem 2 to get the optimal v_{new}, and then the optimal v via (15).

Let us summarize. The given system is

$$\begin{aligned} G: \quad \xi(k+1) &= A\xi(k) + B_1\omega(k) + B_2v(k), \quad \omega = \omega_0\delta_d \\ \zeta(k) &= C_1\xi(k) + D_{11}\omega(k) + D_{12}v(k) \end{aligned}$$

and the problem is $\min_v \|\zeta\|_2$. Under assumptions (i)-(iii), define

$$\begin{aligned} H &= \left(\begin{bmatrix} A - B_2 D_{12}' C_1 & 0 \\ -C_1'(I - D_{12} D_{12}') C_1 & I \end{bmatrix}, \begin{bmatrix} I & B_2 B_2' \\ 0 & (A - B_2 D_{12}' C_1)' \end{bmatrix} \right) \\ X &= Ric(H) \\ F &= -(I + B_2' X B_2)^{-1} B_2' X (A - B_2 D_{12}' C_1) - D_{12}' C_1 \\ &= -(I + B_2' X B_2)^{-1} (B_2' X A + D_{12}' C_1) \\ F_1 &= -(I + B_2' X B_2)^{-1} (B_2' X B_1 + D_{12}' D_{11}) \\ \hat{g}_c(\lambda) &= \left[\begin{array}{c|c} A + B_2 F & B_1 + B_2 F_1 \\ \hline C_1 + D_{12} F & D_{11} + D_{12} F_1 \end{array} \right]. \end{aligned}$$

Theorem 3 *The unique optimal control is* $v_{opt} = F\xi + F_1\omega$. *Moreover,*

$$\min_v \|\zeta\|_2 = \|\hat{g}_c \omega_0\|_2.$$

B. Output Feedback

Now we study the $\mathcal{H}_2$-optimal control problem posed at the start of Section III, where the measured output ψ does not have full information and therefore dynamic feedback is necessary. All discussion pertains to the standard discrete-time setup. Let $T_{\zeta\omega}$ denote the closed-loop system from ω to ζ. We say a causal, finite-dimensional, linear time-invariant controller K is *admissible* if it achieves internal stability. Our goal is to find an admissible K to minimize $\|\hat{t}_{\zeta\omega}\|_2$.

Again, we will first do the orthogonal case in detail and then present the solution for the general case.

1. Two Special Problems

For later benefit, we begin with two special $\mathcal{H}_2$-optimal control problems.

The *first special problem* has a G of the form

$$\hat{g}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & I & 0 \end{array} \right]$$

with the assumptions

(i) (A, B_2) is stabilizable;

(ii) $D_{12}' \begin{bmatrix} C_1 & D_{12} \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}$;

(iii) the matrix

$$\begin{bmatrix} A - \lambda & B_2 \\ C_1 & D_{12} \end{bmatrix}$$

has full rank $\forall \lambda \in \partial \mathbf{D}$;

(iv) $A - B_1 C_2$ is stable.

Since $D_{21} = I$, the disturbance, ω, enters the measurement directly. Define

$$\begin{aligned} H &= \left(\begin{bmatrix} A & 0 \\ -C_1' C_1 & I \end{bmatrix}, \begin{bmatrix} I & B_2 B_2' \\ 0 & A' \end{bmatrix} \right) \\ X &= Ric(H) \\ F &= -(I + B_2' X B_2)^{-1} B_2' X A \\ F_1 &= -(I + B_2' X B_2)^{-1} (B_2' X B_1 + D_{12}' D_{11}) \\ \hat{g}_c(\lambda) &= \left[\begin{array}{c|c} A + B_2 F & B_1 + B_2 F_1 \\ \hline C_1 + D_{12} F & D_{11} + D_{12} F_1 \end{array} \right]. \end{aligned}$$

The next result says that the optimal controller achieves the same performance as the optimal state feedback and disturbance feedforward were the state and the disturbance directly measured.

Theorem 4 *The unique optimal controller is*

$$\hat{k}_{opt}(\lambda) := \left[\begin{array}{c|c} A + B_2 F - B_2 F_1 C_2 - B_1 C_2 & B_1 + B_2 F_1 \\ \hline F - F_1 C_2 & F_1 \end{array} \right].$$

Moreover,

$$\min_K \|\hat{t}_{\zeta\omega}\|_2 = \|\hat{g}_c\|_2.$$

Proof Apply the controller K_{opt} and let η denote its state. Then the system equations are

$$\begin{aligned} \xi(k+1) &= A\xi(k) + B_1 \omega(k) + B_2 v(k) \\ \zeta(k) &= C_1 \xi(k) + D_{11} \omega(k) + D_{12} v(k) \end{aligned}$$

$$\begin{aligned}\psi(k) &= C_2\xi(k) + \omega(k)\\ \eta(k+1) &= (A + B_2F - B_2F_1C_2 - B_1C_2)\eta(k) + (B_1 + B_2F_1)\psi(k)\\ v(k) &= (F - F_1C_2)\eta(k) + F_1\psi(k),\end{aligned}$$

so

$$\eta(k+1) = A\eta(k) + B_1\omega(k) + B_2v(k) + B_1[\psi(k) - C_2\eta(k) - \omega(k)].$$

Defining $\varepsilon := \xi - \eta$, we get

$$\varepsilon(k+1) = (A - B_1C_2)\varepsilon(k).$$

It is now easy to infer internal stability from stability of $A + B_2F$ and $A - B_1C_2$. For zero initial conditions on ξ and η, we have $\varepsilon(k) \equiv 0$, i.e., $\eta(k) \equiv \xi(k)$. Hence

$$\begin{aligned}v(k) &= (F - F_1C_2)\eta(k) + F_1\psi(k)\\ &= F\xi(k) + F_1[\psi(k) - C_2\xi(k)]\\ &= F\xi(k) + F_1\omega(k).\end{aligned}$$

This means that K_{opt} has the same action as the optimal state feedback and disturbance feedforward. Thus by Theorem 2 K_{opt} is optimal and in this case

$$\|\hat{t}_{\zeta\omega}\|_2 = \|\hat{g}_c\|_2.$$

The proof that K_{opt} is unique can be obtained from the proof of Theorem 6 below: For every admissible controller the equation

$$\|\hat{t}_{\zeta\omega}\|_2^2 = \|\hat{g}_c\|_2^2 + \|(I + B_2'XB_2)^{1/2}\hat{t}_{v\omega}\|_2^2$$

is valid; then show that the unique solution of $\hat{t}_{v\omega} = 0$ is the controller given. The detail is omitted.

■

The *second special problem* is the dual of the first; so G has the form

$$\hat{g}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & I \\ C_2 & D_{21} & 0 \end{array}\right]$$

with the assumptions

(i) (C_2, A) is detectable;

(ii) $\begin{bmatrix} B_1 \\ D_{21} \end{bmatrix} D_{21}' = \begin{bmatrix} 0 \\ I \end{bmatrix}$;

(iii) the matrix

$$\begin{bmatrix} A - \lambda & B_1 \\ C_2 & D_{21} \end{bmatrix}$$

has full rank $\forall \lambda \in \partial \mathbf{D}$;

(iv) $A_{\cdot} - B_2 C_1$ is stable.

Define

$$\begin{aligned} J &= \left(\begin{bmatrix} A' & 0 \\ -B_1 B_1' & I \end{bmatrix}, \begin{bmatrix} I & C_2' C_2 \\ 0 & A \end{bmatrix} \right) \\ Y &= Ric(J) \\ L &= -AYC_2'(I + C_2 Y C_2')^{-1} \\ L_1 &= -(D_{11} D_{21}' + C_1 Y C_2')(I + C_2 Y C_2')^{-1} \\ \hat{g}_f(\lambda) &= \left[\begin{array}{c|c} A + LC_2 & B_1 + LD_{21} \\ \hline C_1 + L_1 C_2 & D_{11} + L_1 D_{21} \end{array} \right]. \end{aligned}$$

Theorem 5 *The unique optimal controller is*

$$\hat{k}_{opt}(\lambda) := \left[\begin{array}{c|c} A + LC_2 - B_2 L_1 C_2 - B_2 C_1 & L - B_2 L_1 \\ \hline C_1 + L_1 C_2 & L_1 \end{array} \right].$$

Moreover,

$$\min_K \|\hat{t}_{\zeta\omega}\|_2 = \|\hat{g}_f\|_2.$$

Proof Notice the duality: $\hat{k}_{opt}$ is the unique optimal controller for $\hat{g}$ iff $\hat{k}_{opt}'$ is the unique optimal controller for $\hat{g}'$. Then the results follow from Theorem 4.

■

2. Orthogonal Case

Now consider the case with

$$\hat{g}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right]$$

and with the following assumptions:

(i) (A, B_2) is stabilizable and (C_2, A) is detectable;

(ii) $D_{12}' \begin{bmatrix} C_1 & D_{12} \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}$ and $\begin{bmatrix} B_1 \\ D_{21} \end{bmatrix} D_{21}' = \begin{bmatrix} 0 \\ I \end{bmatrix}$;

(iii) the matrices

$$\begin{bmatrix} A - \lambda & B_2 \\ C_1 & D_{12} \end{bmatrix}, \begin{bmatrix} A - \lambda & B_1 \\ C_2 & D_{21} \end{bmatrix}$$

have full rank $\forall \lambda \in \partial \mathbb{D}$;

The first parts of assumptions (i)-(iii) were seen in Section III-A-1. The second parts of assumptions (i)-(iii) are dual to their first parts: Together they guarantee that the symplectic pair J introduced above belongs to *dom Ric*. Finally, the second part of assumption (ii) concerns how the exogenous signal ω enters G: The plant disturbance and the sensor noise are orthogonal, and the sensor noise weighting is normalized and nonsingular.

Define

$$\begin{aligned} H &= \left(\begin{bmatrix} A & 0 \\ -C_1'C_1 & I \end{bmatrix}, \begin{bmatrix} I & B_2B_2' \\ 0 & A' \end{bmatrix} \right) \\ X &= Ric(H) \\ F &= -(I + B_2'X B_2)^{-1} B_2'X A \\ F_1 &= -(I + B_2'X B_2)^{-1}(B_2'X B_1 + D_{12}'D_{11}) \\ \hat{g}_c(\lambda) &= \left[\begin{array}{c|c} A + B_2F & B_1 + B_2F_1 \\ \hline C_1 + D_{12}F & D_{11} + D_{12}F_1 \end{array}\right] \\ J &= \left(\begin{bmatrix} A' & 0 \\ -B_1B_1' & I \end{bmatrix}, \begin{bmatrix} I & C_2'C_2 \\ 0 & A \end{bmatrix} \right) \end{aligned}$$

$$\begin{aligned} Y &= Ric(J) \\ L &= -AYC_2'(I + C_2YC_2')^{-1} \\ L_1 &= (F_1D_{21}' + FYC_2')(I + C_2YC_2')^{-1} \\ R &= (I + B_2'XB_2)^{1/2} \\ \hat{g}_f(\lambda) &= \left[\begin{array}{c|c} A + LC_2 & B_1 + LD_{21} \\ \hline R(L_1C_2 - F) & R(L_1D_{21} - F_1) \end{array}\right]. \end{aligned}$$

Theorem 6 *The unique optimal controller is*

$$\hat{k}_{opt}(\lambda) := \left[\begin{array}{c|c} A + B_2F + LC_2 - B_2L_1C_2 & L - B_2L_1 \\ \hline L_1C_2 - F & L_1 \end{array}\right].$$

Moreover,

$$\min_K \|\hat{t}_{\zeta\omega}\|_2^2 = \|\hat{g}_c\|_2^2 + \|\hat{g}_f\|_2^2.$$

The first term in the minimum cost, $\|\hat{g}_c\|_2^2$, is associated with optimal control with state feedback and disturbance feedforward and the second, $\|\hat{g}_f\|_2^2$, with optimal filtering. These two norms can easily be computed as follows:

$$\begin{aligned} \|\hat{g}_c\|_2^2 &= \text{trace}\,[(D_{11} + D_{12}F_1)'(D_{11} + D_{12}F_1) \\ &\quad +(B_1 + B_2F_1)'X(B_1 + B_2F_1)] \\ \|\hat{g}_f\|_2^2 &= \text{trace}\,\{R[(L_1D_{21} - F_1)(L_1D_{21} - F_1)' \\ &\quad +(L_1C_2 - F)Y(L_1C_2 - F)']R'\}. \end{aligned}$$

Here X and Y also satisfy respectively the two Lyapunov equations

$$\begin{aligned} (A + B_2F)'X(A + B_2F) - X + (C_1 + D_{12}F)'(C_1 + D_{12}F) &= 0 \\ (A + LC_2)Y(A + LC_2)' - X + (B_1 + LD_{21})(B_1 + LD_{21})' &= 0. \end{aligned}$$

Proof of Theorem 6 Let K be any admissible controller. Start with the system equations

$$\begin{aligned} \xi(k+1) &= A\xi(k) + B_1\omega(k) + B_2v(k) \\ \zeta(k) &= C_1\xi(k) + D_{11}\omega(k) + D_{12}v(k), \end{aligned}$$

and define a new control variable, $\nu := v - F\xi - F_1\omega$, as in Section III-A-1. The equations become

$$\begin{aligned}\xi(k+1) &= (A + B_2F)\xi(k) + (B_1 + B_2F_1)\omega(k) + B_2\nu(k)\\ \zeta(k) &= (C_1 + D_{12}F)\xi(k) + (D_{11} + D_{12}F_1)\omega(k) + D_{12}\nu(k),\end{aligned}$$

or in the frequency domain

$$\hat{\zeta} = \hat{g}_c\hat{\omega} + \hat{g}_i\hat{\nu},$$

where

$$\hat{g}_i(\lambda) = \left[\begin{array}{c|c} A + B_2F & B_2 \\ \hline C_1 + D_{12}F & D_{12}\end{array}\right].$$

This implies that

$$\hat{t}_{\zeta\omega} = \hat{g}_c + \hat{g}_i\hat{t}_{\nu\omega},$$

where $\hat{t}_{\nu\omega}$ is the transfer matrix from ν to ω. So it follows from Lemma 3 that

$$\|\hat{t}_{\zeta\omega}\|_2^2 = \|\hat{g}_c\|_2^2 + \|R\hat{t}_{\nu\omega}\|_2^2.$$

Now ν is generated as in Figure 4, where

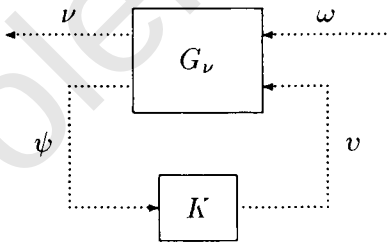


Figure 4: The system to generate ν.

$$\hat{g}_\nu(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline -F & -F_1 & I \\ C_2 & D_{21} & 0\end{array}\right].$$

Note that K stabilizes G iff K stabilizes G_ν (the two closed-loop systems have identical A-matrices). So

$$\min_K \|\hat{t}_{\zeta\omega}\|_2^2 = \|\hat{g}_c\|_2^2 + \min_K \|R\hat{t}_{\nu\omega}\|_2^2.$$

Define

$$\begin{aligned} \nu_{new} &= R\nu \\ K_{new} &= RK \\ G_{\nu new} &= \begin{bmatrix} R & 0 \\ 0 & I \end{bmatrix} G_\nu \begin{bmatrix} I & 0 \\ 0 & R^{-1} \end{bmatrix}. \end{aligned}$$

Then minimizing $\|R\hat{t}_{\nu\omega}\|_2$ is exactly minimizing the norm on $\mathcal{H}_2$ of the transfer matrix $\omega \mapsto \nu_{new}$ in Figure 5, where

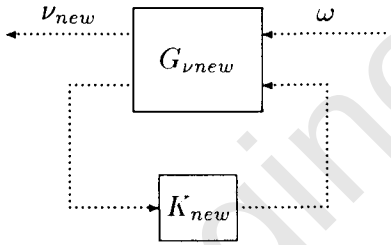


Figure 5: The system to generate ν_{new}.

$$\hat{g}_{\nu new}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2R^{-1} \\ \hline -RF & -RF_1 & I \\ C_2 & D_{21} & 0 \end{array}\right].$$

Now this is the second special problem. So by Theorem 5 the unique optimal controller is

$$\hat{k}_{new,opt}(\lambda) := \left[\begin{array}{c|c} A + LC_2 - B_2L_1C_2 + B_2F & L - B_2L_1 \\ \hline R(L_1C_2 - F) & RL_1 \end{array}\right]$$

and the minimum cost is

$$\min \|R\hat{t}_{\nu\omega}\|_2 = \|\hat{g}_f\|_2.$$

Therefore for the original problem we have

$$\begin{aligned} \hat{k}_{opt}(\lambda) &= R^{-1}\hat{k}_{new,opt}(\lambda) \\ &= \left[\begin{array}{c|c} A + LC_2 - B_2L_1C_2 + B_2F & L - B_2L_1 \\ \hline L_1C_2 - F & L_1 \end{array}\right] \\ \min \|\hat{t}_{\zeta\omega}\|_2^2 &= \|\hat{g}_c\|_2^2 + \|\hat{g}_f\|_2^2. \end{aligned}$$

■

3. General Case

Again, we start with a system of the form

$$\hat{g}(\lambda) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right]$$

The following assumptions are made:

(i) (A, B_2) is stabilizable and (C_2, A) is detectable;

(ii) $D_{12}'D_{12} = I$ and $D_{21}D_{21}' = I$;

(iii) the matrices

$$\begin{bmatrix} A - \lambda & B_2 \\ C_1 & D_{12} \end{bmatrix}, \quad \begin{bmatrix} A - \lambda & B_1 \\ C_2 & D_{21} \end{bmatrix}$$

have full rank $\forall \lambda \in \partial\mathbb{D}$;

Note that the two orthogonality conditions are not assumed. In assumption (ii) it is essential that the two matrices $D_{12}'D_{12}$ and $D_{21}D_{21}'$ be just nonsingular; for they can be normalized via changing coordinates in v and ψ:

$$\begin{aligned} v_{new} &:= (D_{12}'D_{12})^{1/2} v \\ \psi_{new} &:= (D_{21}D_{21}')^{-1/2} \psi. \end{aligned}$$

Using the optimal control for the general case in Section III-A-2, one can derive the results following the procedure in the orthogonal case used in Section III-B-2. The formulas are summarized as follows:

$$\begin{aligned} H &= \left(\begin{bmatrix} A - B_2 D_{12}' C_1 & 0 \\ -C_1'(I - D_{12}D_{12}')C_1 & I \end{bmatrix}, \begin{bmatrix} I & B_2 B_2' \\ 0 & (A - B_2 D_{12}' C_1)' \end{bmatrix} \right) \\ X &= Ric(H) \\ F &= -(I + B_2' X B_2)^{-1}(B_2' X A + D_{12}' C_1) \end{aligned}$$

$$
\begin{aligned}
F_1 &= -(I + B_2' X B_2)^{-1}(B_2' X B_1 + D_{12}' D_{11}) \\
\hat{g}_c(\lambda) &= \left[\begin{array}{c|c} A + B_2 F & B_1 + B_2 F_1 \\ \hline C_1 + D_{12} F & D_{11} + D_{12} F_1 \end{array}\right] \\
J &= \left(\begin{bmatrix} (A - B_1 D_{21}' C_2)' & 0 \\ -B_1 (I - D_{21}' D_{21}) B_1' & I \end{bmatrix}, \begin{bmatrix} I & C_2' C_2 \\ 0 & A - B_1 D_{21}' C_2 \end{bmatrix} \right) \\
Y &= Ric(J) \\
L &= -(AYC_2' + B_1 D_{21}')(I + C_2 Y C_2')^{-1} \\
L_1 &= (FYC_2' + F_1 D_{21}')(I + C_2 Y C_2')^{-1} \\
R &= (I + B_2' X B_2)^{1/2} \\
\hat{g}_f(\lambda) &= \left[\begin{array}{c|c} A + LC_2 & B_1 + LD_{21} \\ \hline R(L_1 C_2 - F) & R(L_1 D_{21} - F_1) \end{array}\right].
\end{aligned}
$$

Theorem 7 *The unique optimal controller is*

$$
\hat{k}_{opt}(\lambda) = \left[\begin{array}{c|c} A + B_2 F + LC_2 - B_2 L_1 C_2 & L - B_2 L_1 \\ \hline L_1 C_2 - F & L_1 \end{array}\right].
$$

Moreover,

$$
\min_K \|\hat{t}_{\zeta\omega}\|_2^2 = \|\hat{g}_c\|_2^2 + \|\hat{g}_f\|_2^2.
$$

IV. Sampled-Data Case

The formulas in the preceding section have direct application in sampled-data control problems. We will look at the case when the control signal is the output of a D/A device, but is otherwise unconstrained. Then the optimal control law is a sampled state feedback with a suitable disturbance feedforward.

Consider the sampled-data setup in Figure 6. Here the continuous-time system G is described by the state equations

$$
\begin{aligned}
\dot{x}(t) &= Ax(t) + B_1 w(t) + B_2 u(t) \\
z(t) &= C_1 x(t) + D_{11} w(t) + D_{12} u(t).
\end{aligned}
$$

The control input u is obtained through a zero-order hold H_0 with sampling period h, processing a discrete signal v, the control sequence. Thus u and v satisfy

$$
u(t) = v(k), \quad kh \le t < (k+1)h.
$$

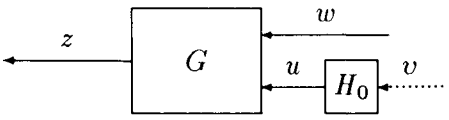


Figure 6: The sampled-data system.

The exogenous input w is assumed to be fixed and affects the system only through the first sampling period. So w has support in $[0, h)$. For example, w could be the impulse $w(t) = w_0\delta(t - t_0)$, where w_0 is a constant vector and $0 \le t_0 < h$.

Our sampled-data problem is

$$\min_{v \in \ell_{2e}} \|z\|_2,$$

the norm being on $\mathcal{L}_2[0, \infty)$. We shall assume that G has zero initial state.

Now we use the lifting technique in [14] to set the problem in the lifted space. Following the notation in [14], let $\mathcal{E}$ denote any finite-dimensional Euclidean space (its dimension will be irrelevant) and $\mathcal{K}$ denote $\mathcal{L}_2[0, h)$. The sequence space $\ell_2(\mathbf{Z}_+, \mathcal{K})$, or simply $\ell_2(\mathcal{K})$, is defined to be

$$\ell_2(\mathcal{K}) := \{\psi : \psi_k \in \mathcal{K}, \quad \sum_{k=0}^{\infty} \|\psi_k\|^2 < \infty\}.$$

The norm for ψ_k is the one on $\mathcal{K}$ and the norm for $\ell_2(\mathcal{K})$ is given by

$$\|\psi\|_2 := \left(\sum_{k=0}^{\infty} \|\psi_k\|^2\right)^{1/2}.$$

The lifting operator W, mapping $\mathcal{L}_2[0, \infty)$ to $\ell_2(\mathcal{K})$, is defined by

$$\psi = Wy \iff \psi_k(t) = y(t + kh), \quad 0 \le t < h.$$

We denote the lifted signal Wy by $\tilde{y}$.

Now we lift the system in the preceding figure to get

$$\begin{aligned} \tilde{z} &= \tilde{G}\begin{bmatrix} \tilde{w} \\ v \end{bmatrix} \\ \tilde{G} &= WG\begin{bmatrix} W^{-1} & 0 \\ 0 & H_0 \end{bmatrix}. \end{aligned}$$

Here the lifted system $\tilde{G}$ satisfies the discrete-time equations [14] [since w has support in $[0,h)$, $\tilde{w}$ is a pulse sequence in $\ell_2(\mathcal{K})$]

$$\xi(k+1) = A_d\xi(k) + \tilde{B}_1\tilde{w}_k + B_{2d}v(k), \quad \tilde{w}_k = \tilde{w}_0\delta_d(k) \tag{16}$$

$$\tilde{z}_k = \tilde{C}_1\xi(k) + \tilde{D}_{11}\tilde{w}_k + \tilde{D}_{12}v(k), \tag{17}$$

where $\xi(k) := x(kh)$ and the operators are given by

$$A_d: \ \mathcal{E} \to \mathcal{E}, \quad A_d = \mathrm{e}^{hA}$$

$$B_{2d}: \ \mathcal{E} \to \mathcal{E}, \quad B_{2d} = \int_0^h \mathrm{e}^{tA}\,dt\,B_2$$

$$\tilde{B}_1: \ \mathcal{K} \to \mathcal{E}, \quad \tilde{B}_1 w = \int_0^h \mathrm{e}^{(h-\tau)A}B_1 w(\tau)\,d\tau$$

$$\tilde{C}_1: \ \mathcal{E} \to \mathcal{K}, \quad (\tilde{C}_1 x)(t) = C_1\mathrm{e}^{tA}x$$

$$\tilde{D}_{11}: \ \mathcal{K} \to \mathcal{K}, \quad (\tilde{D}_{11}w)(t) = D_{11}w(t) + C_1\int_0^t \mathrm{e}^{(t-\tau)A}B_1 w(\tau)\,d\tau$$

$$\tilde{D}_{12}: \ \mathcal{E} \to \mathcal{K}, \quad (\tilde{D}_{12}v)(t) = D_{12}v + C_1\int_0^t \mathrm{e}^{(t-\tau)A}\,d\tau\,B_2 v.$$

The system $\tilde{G}$ can be regarded as a linear time-invariant system in discrete time, with $\tilde{z}_k$ and $\tilde{w}_k$ being infinite-dimensional (functions in $\mathcal{K}$). Since the lifting operator is norm-preserving, the equivalent discrete $\mathcal{H}_2$ problem is

$$\min_v \|\tilde{z}\|_2$$

subject to equations (16-17), the norm being on $\ell_2(\mathcal{K})$.

This problem looks almost like the one we studied in Section III-A, the difference being that now we are treating operators instead of matrices. So the derivation for the optimal control in Section III-A carries over except for a few changes such as using operator adjoints, denoted by *, instead of transposes.

In view of the assumptions in Section III-A, we assume here that

(i) (A_d, B_{2d}) is stabilizable;

(ii) the matrix $\tilde{D}_{12}^* \tilde{D}_{12}$ is invertible;

(iii) the matrix operator

$$\begin{bmatrix} A_d - \lambda & B_{2d} \\ \tilde{C}_1 & \tilde{D}_{12} \end{bmatrix}$$

is injective $\forall \lambda \in \partial \mathbf{D}$.

To write down the formulas, we need to normalize $\tilde{D}_{12}$ first. So define the matrix

$$Q = (\tilde{D}_{12}^* \tilde{D}_{12})^{-1/2} \tag{18}$$

and $v_{new} := Q^{-1} v$ to get the normalized equations

$$\begin{aligned} \xi(k+1) &= A_d \xi(k) + \tilde{B}_1 \tilde{w}_k + B_{2d} Q v_{new}(k) \\ \tilde{z}_k &= \tilde{C}_1 \xi(k) + \tilde{D}_{11} \tilde{w}_k + \tilde{D}_{12} Q v_{new}(k). \end{aligned}$$

We can now give the optimal control. Define

$$\begin{aligned} H &= \left(\begin{bmatrix} A_d - B_{2d} Q^2 \tilde{D}_{12}^* \tilde{C}_1 & 0 \\ -\tilde{C}_1^* (I - \tilde{D}_{12} Q^2 \tilde{D}_{12}^*) \tilde{C}_1 & I \end{bmatrix}, \right. \\ &\qquad \left. \begin{bmatrix} I & B_{2d} Q^2 B_{2d}' \\ 0 & A_d' - \tilde{C}_1^* \tilde{D}_{12} Q^2 B_{2d}' \end{bmatrix} \right) \qquad (19) \\ X &= Ric(H) \qquad (20) \\ F &= -(Q^{-2} + B_{2d}' X B_{2d})^{-1} (B_{2d}' X A_d + \tilde{D}_{12}^* \tilde{C}_1) \qquad (21) \\ \tilde{F}_1 &= -(Q^{-2} + B_{2d}' X B_{2d})^{-1} (B_{2d}' X \tilde{B}_1 + \tilde{D}_{12}^* \tilde{D}_{11}) \qquad (22) \end{aligned}$$

and the system $\tilde{G}_c$ (with input $\tilde{w}$ and output $\tilde{z}$) via

$$\begin{aligned} \eta(k+1) &= (A_d + B_{2d} F) \eta(k) + (\tilde{B}_1 + B_{2d} \tilde{F}_1) \tilde{w}_k \\ \tilde{z}_k &= (\tilde{C}_1 + \tilde{D}_{12} F) \eta(k) + (\tilde{D}_{11} + \tilde{D}_{12} \tilde{F}_1) \tilde{w}_k. \end{aligned}$$

Theorem 8 *The unique optimal control is* $v_{opt}(k) = F\xi(k) + \tilde{F}_1 \tilde{w}_k$. *Moreover,*

$$\min_v \|\tilde{z}\|_2 = \|\tilde{G}_c \tilde{w}\|_2.$$

The following remarks are in order:

1. H defined in (19) is a constant matrix pair because $\tilde{D}_{12}^{*}\tilde{C}_1$ as a whole is a matrix (operator on $\mathcal{E}$). The formulas for $\tilde{D}_{12}^{*}\tilde{C}_1$ can be derived easily (see Section V). Similarly, F, though involving operators, is also a matrix. However, the feedforward gain $\tilde{F}_1$ is an operator mapping $\mathcal{K}$ to $\mathcal{E}$; its action on a fixed $\tilde{w}$ can be determined a priori.

2. The optimal control can be written as

$$v_{opt}(k) = \begin{cases} \tilde{F}_1 \tilde{w}_0, & k = 0 \\ Fx(kh), & k \geq 1. \end{cases}$$

The optimal state feedback $[Fx(kh)]$ is independent of the exogenous input w and can be realized by sampling $x(t)$ at the same rate as the hold operator. In particular, if the rate of the D/A device is chosen, one does not gain any advantage by sampling $x(t)$ faster, or even by measuring $x(t)$ continuously.

3. Assumption (i) is satisfied if (A, B_2) is stabilizable in continuous time and if the sampling is non-pathological in a certain sense, see, e.g., [18].

4. It is not hard to show that assumption (iii) is satisfied if

$$\begin{bmatrix} A_d - \lambda & B_{2d} \\ C_1 & D_{12} \end{bmatrix}$$

is injective $\forall \lambda \in \partial\mathbb{D}$, which can be checked easily since it is a matrix expression.

V. Example

The theory of the preceding section is now applied to a simple setup consisting of two motors controlled by one PC. The block diagram for the system is given in Figure 7. Shown there are two identical motors, with shaft angles θ_1 and θ_2. The left-hand motor is forced by an external torque w. The controller, K, inputs the two shaft positions and their velocities, and outputs two voltages, u_1 and u_2, to the motors. The goal is that the system should act like a telerobot:

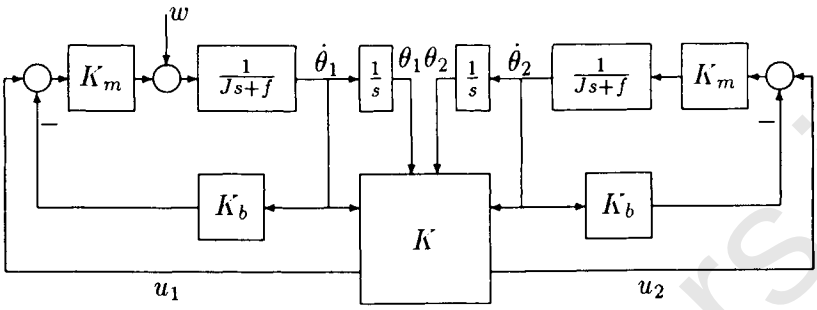


Figure 7: A two-motor control system.

When a human applies a torque w, the "master" (left-hand) motor should turn appropriately and the "slave" (right-hand) motor should follow it.

The state vector is taken to be

$$x = [\theta_1 \ \dot{\theta}_1 \ \theta_2 \ \dot{\theta}_2]'.$$

For certain values of the physical parameters, the state matrices are

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -24.51 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -24.51 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 0 \\ 2.1513 \times 10^5 \\ 0 \\ 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ 179.4 & 0 \\ 0 & 0 \\ 0 & 179.4 \end{bmatrix}.$$

The vector z to be regulated is taken to be

$$z = [10(\theta_1 - \theta_2) \ \ \dot{\theta}_1 \ \ 0.2\dot{\theta}_2 \ \ u_1 \ \ 0.1u_2]'.$$

The first component guarantees that the slave will follow the master; the second and third components are included to get the motors finally to stop moving after a finite-duration torque is applied; and the fourth and fifth components are included to make the problem

nonsingular. The constant weights were obtained by trial-and-error. With this choice for z we get

$$C_1 = \begin{bmatrix} 10 & 0 & -10 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad D_{11} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad D_{12} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

Note that (C_1, A) is not detectable, reflecting our desire to have θ_1 and θ_2 settle to nonzero final values after a finite-duration torque is applied; this will necessitate some massaging below.

First, the optimal analog state feedback controller is computed. The MATLAB command is

$$F = -\text{LQR}\ (A, B_2, C_1'C_1 + 10^{-4}I, D_{12}'D_{12}),$$

where the third matrix on the right-hand side has been perturbed to make it nonsingular. The controlled analog system was simulated for the finite-duration input

$$w(t) = \begin{cases} 0.005, & 0 \le t \le 0.1 \\ 0, & t > 0.1 \end{cases}$$

and the result is shown in Figure 8 (θ_1 solid, θ_2 dash, in degrees versus time in seconds).

Turning to the optimal sampled-data control, for the state-feedback gain F one must compute the matrices

$$\begin{aligned} \tilde{D}_{12}^*\tilde{D}_{12} &= hD_{12}'D_{12} + D_{12}'C_1 \int_0^h \int_0^t e^{\tau A} d\tau dt B_2 \\ &\quad + \left[D_{12}'C_1 \int_0^h \int_0^t e^{\tau A} d\tau dt B_2 \right]' \\ &\quad + B_2' \int_0^h \left[\int_0^t e^{\tau A'} d\tau \right] C_1'C_1 \left[\int_0^t e^{\tau A} d\tau \right] dt B_2 \\ Q &= (\tilde{D}_{12}^*\tilde{D}_{12})^{-1/2} \\ \tilde{C}_1^*\tilde{C}_1 &= \int_0^h e^{tA'} C_1'C_1 e^{tA} dt \\ \tilde{D}_{12}^*\tilde{C}_1 &= D_{12}'C_1 \int_0^h e^{tA} dt + B_2' \int_0^h \int_0^t e^{\tau A'} d\tau C_1'C_1 e^{tA} dt. \end{aligned}$$

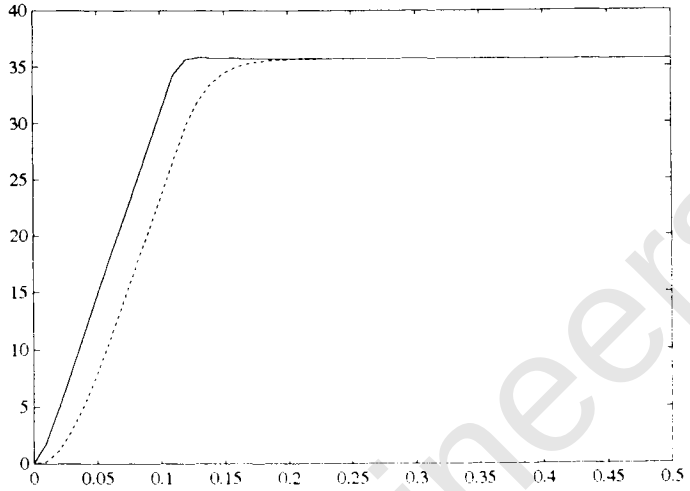


Figure 8: Optimal analog controller.

Then the MATLAB commands for F are

$$\begin{aligned}[F_{tmp}, X] &= \text{DLQR}\left[A_d - B_{2d}Q^2(\tilde{D}_{12}^{*}C_1), B_{2d},\right.\\ &\qquad \left.\tilde{C}_1^{*}\tilde{C}_1 - (\tilde{D}_{12}^{*}\tilde{C}_1)'Q^2(\tilde{D}_{12}^{*}\tilde{C}_1) + 10^{-4}I, Q^2\right]\\ F &= -(Q^{-2} + B_{2d}'XB_{2d})^{-1}(B_{2d}'XA_d + \tilde{D}_{12}^{*}\tilde{C}_1).\end{aligned}$$

The disturbance-feedforward gain $\tilde{F}_1$ is an operator $\mathcal{K} \to \mathcal{E}$, but since $w(t)$ here is constant over $[0, h)$, the action of $\tilde{F}_1$ is to multiply by a matrix, denoted, say, F_1. This matrix is computed as follows:

$$\begin{aligned}\tilde{D}_{12}^{*}\tilde{D}_{11} &= hD_{12}'D_{11} + D_{12}'C_1\int_0^h\int_0^t e^{\tau A}d\tau dtB_1\\ &\quad + \left[D_{12}'C_1\int_0^h\int_0^t e^{\tau A}d\tau dtB_1\right]'\\ &\quad + B_2'\int_0^h\left[\int_0^t e^{\tau A'}d\tau\right]C_1'C_1\left[\int_0^t e^{\tau A}d\tau\right]dtB_1\\ B_{1d} &= \int_0^h e^{\tau A}d\tau B_1\end{aligned}$$

$$F_1 = -(Q^{-2} + B_{2d}' X B_{2d})^{-1}(B_{2d}' X B_{1d} + \tilde{D}_{12}^* \tilde{D}_{11}).$$

The optimal sampled-data control is then

$$v_{opt}(k) = \begin{cases} F_1 w(0), & k = 0 \\ F x(kh). & k \geq 1. \end{cases}$$

These matrices, F and F_1, were computed for $h = 0.1$ (quite large, for illustration) and the resulting sampled-data system was simulated with the same $w(t)$ as above. The responses are shown in Figure 9. The response of the sampled-data system is comparable to that of the

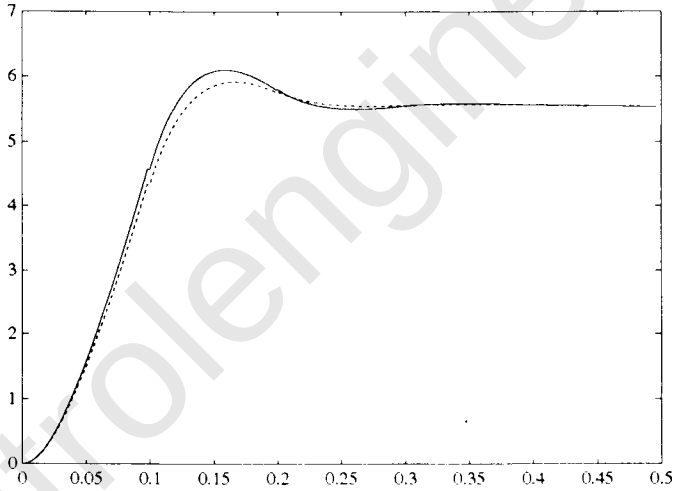


Figure 9: Optimal sampled-data controller.

analog system, except for the DC gains. [The same weights (i.e., C_1 and D_{12}) are used for both the analog and sampled-data controller, but in general weights that are good for the analog controller will not necessarily be good for the sampled-data controller, and vice versa.] However, the analog F does not even stabilize the sampled-data system for this large h. The point is, therefore, that when h is given and is appreciably large, the optimal sampled-data controller is much superior to the discretized optimal analog controller. Finally,

for interest the sampled-data system with only state feedback and not disturbance feedforward, that is,

$$v_{opt}(k) = \begin{cases} 0, & k = 0 \\ Fx(kh), & k \geq 1, \end{cases}$$

was simulated and the responses are shown in Figure 10. Not surprisingly, the response is very poor: The slave motor does not begin to move until the start of the second sampling period, by which time the tracking error is very large.

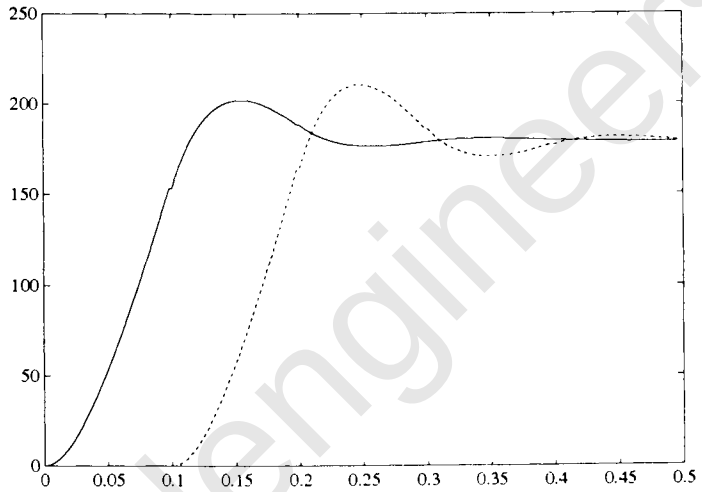


Figure 10: Optimal sampled-data controller without disturbance feedforward.

VI. Conclusion

Direct formulas for the sampled-data output-feedback case are not available because the lifted problem is inherently singular ($\tilde{D}_{21} = 0$). This obstacle does not arise in the operator-theoretic approach of [1].

Acknowledgement The authors wish to thank P. P. Khargonekar and P. A. Iglesias for helpful discussions.

References

[1] T. Chen and B. A. Francis, "$\mathcal{H}_2$-optimal sampled-data control," *IEEE Trans. Automat. Control*, vol. 36, No. 4, pp. 387-397, 1991.

[2] B. Bamieh and J. B. Pearson, "The $\mathcal{H}_2$ problem for sampled-data systems," *Systems and Control Letters*, vol. 19, pp. 1-12, 1992.

[3] P. P. Khargonekar and N. Sivashankar, "$\mathcal{H}_2$ optimal control for sampled-data systems," *Systems and Control Letters*, vol. 18, pp. 627-631, 1992.

[4] M. Athans, "The role and use of the stochastic linear-quadratic-Gaussian problem in control system design," *IEEE Trans. Automat. Control*, vol. 16, pp. 529-552, 1971.

[5] P. Dorato and A.H. Levis, "Optimal linear regulators: the discrete-time case," *IEEE Trans. Automat. Control*, vol. 16, pp. 613-620, 1971.

[6] V. Kučera, "The discrete Riccati equation of optimal control," *Kybernetica*, vol. 8, No. 5, pp. 430-447, 1972.

[7] B. P. Molinari, "The stabilizing solution of the discrete algebraic Riccati equation," *IEEE Trans. Automat. Control*, vol. 20, No. 3, pp. 396-399, 1975.

[8] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[9] T. Pappas, A. J. Laub, and N. R. Sandell, Jr., "On the numerical solution of the discrete-time algebraic Riccati equation," *IEEE Trans. Automat. Control*, vol. 25, No. 4, pp. 631-641, 1980.

[10] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ problems," *IEEE Trans. Automat. Control*, vol. 34, No. 8, pp. 831-847, 1989.

[11] H. T. Toivonen, "Sampled-data control of continuous-time systems with an $\mathcal{H}_\infty$ optimality criterion," *Automatica*, vol. 28, No. 1, pp. 45-54, 1992.

[12] Y. Yamamoto, "A new approach to sampled-data control systems — a function space approach method," *Proc. CDC*, 1990.

[13] B. Bamieh and J. B. Pearson, "A general framework for linear periodic systems with application to $\mathcal{H}_\infty$ sampled-data control," *IEEE Trans. Automat. Control*, vol. 37, pp. 418-435, 1992.

[14] B. Bamieh, J. B. Pearson, B. A. Francis, and A. Tannenbaum, "A lifting technique for linear periodic systems with applications to sampled-data control," *Systems and Control Letters*, vol. 17, pp. 79-88, 1991.

[15] P. A. Iglesias and K. Glover, "State space approach to discrete time $\mathcal{H}_\infty$ control," *Int. J. Control*, vol. 54, pp. 1031-1073, 1991.

[16] H. K. Wimmer, "Normal forms of symplectic pencils and the discrete-time algebraic Riccati equation," *Linear Algebra and Its Applications*, vol. 147, pp. 411-440, 1991.

[17] B. A. Francis, *A Course in $\mathcal{H}_\infty$ Control Theory*, Springer-Verlag, New York, 1987.

[18] B. A. Francis and T. T. Georgiou, "Stability theory for linear time-invariant plants with periodic digital controllers," *IEEE Trans. Automat. Control*, vol. 33, No. 9, pp. 820-832, 1988.

TECHNIQUES FOR REACHABILITY IN INPUT CONSTRAINED DISCRETE TIME LINEAR SYSTEMS

Paolo d'Alessandro *

Elena De Santis **

* Dep. of Mathematics, 3rd University of Roma
Via C. Segre no. 2, 00146 ROMA, Italy

** Dep. of Electrical Eng., University of L'Aquila
67040 Poggio di Roio (L'Aquila), Italy

1 - INTRODUCTION

The theory of constrained systems and their feasibility is recently attracting a growing interest, even though a seminal contribution can be traced back as far as in 1940, in a paper by Liapunov [1], as illustrated in Conti [2]. There is a rather complex articulation of the theory, not only in view of the class of systems considered, but also according to the variables that are bounded by constraints and the characteristics of the constraints themselves. In addition to input constraints, one may in fact consider constraints for the state and/or for the output of the system. The study of positive systems is a case in point.

As we shall show the constrained input case is already a complex one. A clarification of the very concept of input constrained system will therefore be useful. In fact the properties of an unconstrained systems are widely altered

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

according to the features of the constraints. The most trivial case one may think of is that of a time invariant system, which becomes a time varying constrained system. In Section 3 we illustrate these mechanisms and give precise definitions.

A major distinction between classes of input constraints is between pointwise in time sort and the opposite case. The bulk of the literature is for the first, but the second is by no means less interesting. In the first place, a constraint on the state is equivalent, as we shall see, to a non-pointwise constraint for the input. Besides this example, the classical theory of bounded norm reachability considers non pointwise constraints for the input.

Reachability is per se a special kind of constraint on the state, and hence constrained system theory is the appropriate general framework for studying reachability problems.

In addition to the reachability one can consider further constraints, e.g. on the input, giving rise to constrained reachability theory proper. We will exploit this point of view to give a complete solution of the polyhedral case.

In consequence of the difference between a constrained system and its unconstrained counterpart, we have adopted a more refined definition of reachability than the one usually given for unconstrained systems. Moreover we stress from the outset the importance of the study of problems of reachability in presence of noise. A possible assumption on noise will be that of complete lack of information, except for the presence of constraints on noise too, a setting that naturally leads to the study of robust reachability. This problem certainly deserves more attention in the research arena and more space here than we can afford.

Next we pass on to deal with the problem of characterization of reachable sets. In this respect, for the case of pointwise constraints, we will analyze how the properties of the reachable set are connected to the properties of the constraining set.

It will be then in order to study a number of special cases of major interest. First and foremost the case of polyhedral constraints. This will be handled by means of two different approaches. The first approach is based on a decomposition technique applied to the constraining sets viewed as the sum of a

linear subspace plus a cone plus a polytope. Dealing with the special case of polytopes we establish the discrete time version of the bang-bang principle. The second is based on viewing the problem as a mixed input-state constraints problem, in keeping within our initial remark on reachability. This is handled in Section 7 and, to the best of our knowledge, we give the first and only exact computation of the reachable set under no restrictive hypothesis, and, in addition, the proposed solution is explicitly parameterized in terms of the bound vectors of the constraints. This particular technique of handling reachability problems first appeared in [3], which is in turn based on [4].

The next important special case is that of conical constraints. Incidentally, if we look to constraints for the other system variables too, then the study of positive systems is encompassed in this topic. In this respect an interesting extension of the concept of positive systems is in [5].

Another important theory we survey is that of bounded norm input reachability. Here there is a bifurcation of the possible approaches. Following the classical theory we account for smooth norm boundedness (more specifically the euclidean norm in our finite dimensional territory) including robust reachability. But one might also look, e.g., at norms generating polytopic spheres connecting back this case to that of polytopic constraints.

Feasibility and optimization are very close to each other, as is well known. Even though optimization theory is beyond he present purposes, in a few instances, in which the extension is immediate and interesting, we have outlined the links of the results in question to optimization theory. In this respect notice that a general dynamic and closed loop solution for the polyhedral case is given in [4].

1.1 - Brief survey of literature

In this subsection we briefly survey the literature. For space reasons we can only confine ourselves to cite some main contributions. Thus the reader will forgive us for any relevant omission. For the same reason we cannot afford to enter into the details.

A. Marzollo [6] has studied the case of bounded norm input functions for continuous time linear time variant systems. Essentially he rielaborated former results of H.A. Antosiewiez, published in 1963 [7]. The problem here is that of controllability rather than reachability i.e., that of steering the state from one to another given point in the state space in finite time. A noteworthy contribution of Marzollo work is the study of disturbed controllability problems. More precisely, he studied the possibility of reaching a point in a sphere about the desired final state, for any disturbance of a given class.

All the further contribution cited below assume pointwise in time constraints.

R. Conti [2] surveys the case of linear continuous time systems with controls constrained by convex sets, encompassing contributions by A.A. Liapunov [1], D. Blackwell [8], J.P La Salle [9], L.L. Markus [10], R.M. Bianchini [11] and many others. He covers also minimum time optimization problems. Surprisingly enough, his excellent book is only published in italian.

M. E. Evans [12] dealts with controllability of discrete time linear systems with control values constrained to a bounded convex set. The upshot of Evans' work is to connect the controllability properties of a constrained system to those of state eigen-subspaces. He states a decomposition of the given system in subsystems, in the sense that each of them accounts for part of the spectrum of the dynamical matrix, and the controllability properties of the system can be deduced from those of the subsystems. Notice that in his work he assumes the matrix **B** equal to the identity, as it happens in many contributions coming from mathematicians rather than control theorists.

Some authors have approached the problem of finding approximations of the reachable set. We cite for example [13], [14] and [15].

In the first two papers, J. E. Gayek and M.E. Fisher developed a technique for approximating the reachable set for discrete time linear systems subject to bounded control, with the assumption that the state matrix is stable and diagonalizable. They decompose the given system into one and two dimensional subsystems and for each

one they compute a polyhedron, which is an over estimate of the reachable set. These polyhedra are finally used to define a polyhedron which contains the reachable set of the original system. Besides the limiting assumption of this paper it would be more interesting to approximate the reachable set from the inside. A similar comment applies to the paper of M.E. Fisher and W.J. Grantam, in which the reachable set of a discrete time linear system with bounded control is over-estimated by an ellipsoid, computed applying results of Liapunov stability theory.

Some other authors have been involved with the exact calculation of the reachable set. However all the approaches we know of and cited below are based on the very restrictive assumption that the state matrix of the system is nonsingular.

In [16], [17] and [18], a recursive computation of the reachable set at some time k is made, in the case of input and/or state polytopic constraints. This approach requires, at each step, the transformation of the description of the involved polytopes in terms of vertices to the description in terms of boundary hyperplanes and viceversa.

V. G. Rumchev [19] develops a method based on Farkas' lemma for positive linear discrete time systems, with polytopical constraints. This method also extends to non positive systems.

S.S. Keerthy and G. Gilbert [20] look to a different problem, i.e. that of steering the state to the origin in minimum time. They do not make any essential restriction and assume a mixed input-state constraint. Their solution requires at each step the invocation of a modified Fourier Motzkin method.

A general analysis of systems with polyhedral constraints on input - state - output variables is in [21]. In [22] the reachable set of systems with conical constraints is characterized, introducing the theory of minimal invariant cones and generalizing well known results of unconstrained reachability theory. A reachability study of input constrained systems is in [23], where, in particular, was addressed the problem of how some properties of input constraining sets reflects on reachable set. In [3] is developed a technique to compute the reachable set in the polyhedral case, without restrictive assumptions.

2 - NOTATIONS AND TERMINOLOGY

We shall consider, unless otherwise stated, points and sets in the Euclidean real space $\mathbb{R}^n$, even though many of the involved concept are valid for more general linear spaces.

A subset C of a linear space is said to be convex if $(1-\lambda)\ x + \lambda y \in C$ whenever $x, y \in C$, with $0 \leq \lambda \leq 1$.

The convex hull of a set A is the minimum convex set that contains A, and is denoted by $C(A)$.

A convex cone is a nonvoid subset C of a linear space such that $\alpha\ C \subset C$ for any real $\alpha \geq 0$ and $C + C \subset C$.

The convex conical hull of a set A is the minimum convex cone that contains A, and is denoted by $Co(A)$.

The minimal convex cone, which is invariant with respect to a linear operator A and contains a convex cone BC, where C is a convex cone and B is linear operator, is well defined and unique, and is given by $Co\ (\cup\{A^iBC:\ i = 0,1,2\ldots\})$ [22].

The lineality space L of a convex cone C is given by $L = C \cap (-C)$ and is the largest linear subspace contained in C. If the lineality space of C is the origin then C is called a pointed cone.

A closed half-space is a set of the form $\{x:\ (x,y) \leq \alpha\}$ where y is a vector and α is a real constant.

A polyhedral convex set (polyhedron) is a set which can be expressed as the intersection of a finite collection of closed half-spaces. Thus it is also the set of solutions of a finite system of linear inequalities.

The closed segment joining x and y is the set $C(\{x,y\})$, and is denoted by [x:y]. The open segment (x:y) joining x and y is given by [x:y] - {x,y}.

Let K be a convex set. A convex subset W of K is called an

extreme subset if none of its points are included in an open segment joining two points of K, which are not both in W. An extreme subset consisting of one point is an extreme point.

A face of K is a convex subset K' of K such that every closed segment in K with a relative interior point in K' has both endpoints in K'.

The recession cone of K is the set $\{y: x + \lambda y \in K, \forall \lambda \geq 0, \forall x \in K\}$.

A polyhedral convex cone is a polyhedron in which the boundary hyperplanes of the half-spaces pass through the origin. Thus a polyhedral convex cone is the set of solutions of an homogeneous finite system of linear inequalities.

An extreme ray of a polyhedral convex cone is a face which is a half-line emanating from the origin. We call generator a nonzero vector belonging to an extreme ray. A polyhedral pointed convex cone is the conical convex hull of its generators.

Polytopes are bounded polyhedra. A polytope is the convex hull of its extreme points.

3 - A GENERAL PROBLEM SETTING

The reachability problem is that of steering the state of a dynamical system from the zero vector at a certain initial instant of time to a prefixed vector of the state space at a certain final instant of time. One may consider either both instants fixed or fix only one of the two (either the initial time or the final time).

The problem has a particularly simple solution for the unconstrained case, but becomes considerably more complex in the, practically more interesting, constrained case. Constrained means that system variables are bounded to satisfy certain given relations. Such relations may involve systems variables (input, state and output) in any combination, but in this chapter we are mostly interested to the case in which the only input is involved.

Let us now state the problem more precisely starting from recalling the concept of unconstrained system.

3.1 - Unconstrained systems

As usual, a discrete time, linear, time-invariant unconstrained systems is described by means of the input - state - output equations:

$$\begin{aligned} x(t+1) &= A\,x(t) + B\,u(t) + D\,d(t) \\ y(t) &= C\,x(t) \end{aligned} \qquad (3.1.1)$$

where t is an integer variable (representing time) and the functions x, u and y and m assume values in finite dimensional linear spaces, that, without restriction of generality, will be taken as $\mathbb{R}^n$, $\mathbb{R}^p$, $\mathbb{R}^q$ and $\mathbb{R}^s$, respectively. Such functions represent the evolution in time of the state, the input, the output and the disturbance of the system. We have assumed that the unconstrained system is time invariant for the sake of simplicity, so that A, B, D and C are linear operators and hence are represented by matrices of dimensions consistent with that of the involved linear spaces. We shall not distinguish between the operators and the matrices neither terminologically nor notationally.

By these equations we represent a system in the sense that for any initial time t_0 and initial state $x(t_0)$ they define a function associating to any pair of functions u(.) and m(.) defined in the interval $[t_0, +\infty)$, a pair of functions x(.) and y(.) on the same interval, that constitute the unique corresponding solution of the equations. The computation of these functions is straightforward because the equations are easily solved by recursion and substitution.

Let us now put the solution in a form particularly convenient for our purposes. We introduce the notation $\mathbf{u}(t_0,t)$, $\mathbf{d}(t_0,t)$, $\mathbf{x}(t_0,t)$ and $\mathbf{y}(t_0,t)$ to indicate, for any t_0 and $t > t_0$, the restrictions of the functions u(.), d(.), x(.) and y(.) to the interval $[t_0,t)$ for u(.) and d(.) and to the interval $(t_0,t]$ for x(.) and y(.). Clearly these restricted functions can be

represented by block vectors in the following manner:

$$\mathbf{u}(t_0,t) = \begin{bmatrix} u(t_0) \\ \vdots \\ u(t-1) \end{bmatrix} \qquad \mathbf{d}(t_0,t) = \begin{bmatrix} d(t_0) \\ \vdots \\ d(t-1) \end{bmatrix}$$

$$\mathbf{x}(t_0,t) = \begin{bmatrix} x(t_0+1) \\ \vdots \\ x(t) \end{bmatrix} \qquad \mathbf{y}(t_0,t) = \begin{bmatrix} y(t_0+1) \\ \vdots \\ y(t) \end{bmatrix}$$

Using these notations we can write, for the state and the output of the system:

$$\mathbf{x}(t) = L(t_0,t)\, x(t_0) + C(t_0,t)\, \mathbf{u}(t_0,t) + G(t_0,t)\, \mathbf{d}(t_0,t)$$
$$\mathbf{x}(t_0,t) = L(t_0,t)\, x(t_0) + M(t_0,t)\, \mathbf{u}(t_0,t) + N(t_0,t)\, \mathbf{d}(t_0,t)$$

$$y(t) = C\, L(t_0,t)\, x(t)C + C(t_0,t)\, \mathbf{u}(t_0,t) + C\, G(t_0,t)\, \mathbf{d}(t_0,t)$$
$$Y(t_0,t) = CL(t_0,t)x(t_0) + CM(t_0,t)\mathbf{u}(t_0,t) + CN(t_0,t)\, \mathbf{d}(t_0,t)$$

where:

$$C(t_0,t) = (A^{t-t_0-1}B \ \ldots \ AB \ B)$$

$$G(t_0,t) = (A^{t-t_0-1}D \ \ldots \ AD \ D)$$

$$M(t_0,t) = \begin{bmatrix} B & 0 \ \ldots \ 0 & 0 \\ AB & B \ \ldots \ 0 & 0 \\ \vdots & & \\ A^{t-t_0-1} & B \ \ldots \ AB & B \end{bmatrix}$$

$$N(t_0,t) = \begin{bmatrix} D & 0 \ldots 0 & 0 \\ AD & D \ldots 0 & 0 \\ \vdots & & \\ A^{t-t_0-1} & D \ldots AD & D \end{bmatrix}$$

and finally

$$L(t_0,t) = A^{(t-t_0)}$$

$$L(t_0,t) = \begin{bmatrix} A \\ \vdots \\ A^{(t-t_0)} \end{bmatrix}$$

It will be convenient in the sequel to denote by $U(t_0,t)$ the set of all functions $\mathbf{u}(t_0,t)$.

3.2 - Constrained systems

We now turn to the definition of constrained system. At a superficial level such definition is simply obtained associating an unconstrained system with a set of constraints for the system variables. We should consider, however, the ensuing constrained system as a whole and completely distinct from the corresponding unconstrained system. And it turns out that, as we shall illustrate later on, the properties of this new system may be profoundly different from those of the original one. To begin with, the constrained system cannot be considered, in general, a linear one.

A rather general form for the constraints is the following:

$$\mathbf{f}(t_0,t_f,\mathbf{u}(t_0,t_f),\mathbf{d}(t_0,t_f),\mathbf{x}(t_0,t_f),\mathbf{y}(t_0,t_f)) \in Q(t_0,t_f)$$
$$\forall\ t_0,\ t_f \geq t_0 \qquad (3.2.1)$$

where f is a function assuming values in some finite dimensional linear space and $Q(t_0,t_f)$ is a given set in such a space. Time t_f may well be $+\infty$.

We may consider the same constraints for some, instead of for all, pairs t_0, t_f. However this is easy obtained letting $Q(t_0,t_f)$ be the whole space for those pairs t_0, t_f, for which there are no constraints. Thus a distinction between the two cases is not required. The set of these relations is called constraint system. Of course it is assumed that the constraint system defines non void sets of time functions for all variables and time intervals. Any function satisfying the constraint system is called admissible.

In this chapter we are interested (with a few exceptions) to constraints bounding the only input, and, possibly, the disturbance of the system. Thus the above relation takes on the form:

$$f(t_0,t_f,\mathbf{u}(t_0,t_f),\mathbf{d}(t_0,t_f)) \in Q(t_0,t_f) \qquad \forall\ t_0,\ t_f \geq t_0 \qquad (3.2.2)$$

We assume that such constraint system defines non void subsets of U_{t_0} and D_{t_0} $\forall\ t_0$, which are respectively the set of admissible input and noise functions.

For simplicity and space reasons, we shall mostly consider the case in which no noise is present and give our general definitions accordingly. Problems involving disturbances are however of paramount importance. Later on, we shall devote Section 9 to one such problem.

To give an example we introduce right away an interesting special case. Assume that the function g takes values in $\mathbb{R}^m$ and that $v(t_0,t_f)$ is a vector in such space. Then consider the following relation:

$$g(t_0,t_f,\mathbf{u}(t_0,t_f)) \leq v(t_0,t_f) \qquad \forall\ t_0,\ t_f > t_0 \qquad (3.2.3)$$

It is trivial to verify that this latter form can be reduced to the previous one. In fact it suffices to to define the set $Q(t_0,t_f)$ as the cartesian product set $X\{(-\infty,v_i(t_0,t_f)): i=1,..m\}$.

Notice that this latter form of the constraint system in the linear case becomes:

$$W(t_0,t_f)\ \mathbf{u}(t_0,t_f)) \leq v(t_0,t_f) \qquad \forall\ t_0,\ t_f \geq t_0 \qquad (3.2.4)$$

where $W(t_0,t_f)$ is a matrix with dimensions matching that of $\mathbf{U}$ and v. Despite the name we are far away from linear theory. The real nature of the problem is instead polyhedral. Anyway, it is still terminologically usual to call such constraint system a linear constraint system. We shall soon go back to this case in our first formal definition.

With reference to this form of constraints, there is no harm in selecting a certain finite interval $[t_0,t_f]$ and confining the study to the response of a system in this single interval. This is precisely what many papers and books dealing with constrained systems do. In this case there would not be much to add at this point. However, the unconstrained system has a number of interesting properties, that depend substantially on the variability of parameters t_0 and t_f. This suggest that our investigation be carried on in more depth.

Some system properties are necessarily lost when we add the constraints, others may or may not be lost according to the features of the constraints, others may be altered and, finally, new properties may arise. In the basic properties of dynamicity, causality and stationarity [24] as well as in reachability itself the two time parameters t_0 and t_f play a fundamental role. At a very basic level this role depends on the fact that for the unconstrained case the function spaces relative to intervals of the form $[t_0,t_f)$ are strictly connected, since restricting e.g. an input function to a subinterval, a legitimate input function is obtained.

For these reasons we are lead to pose some conditions to be satisfied by the constraint system. The role of such conditions

will become more and more apparent as our analysis develops. More precisely we consider the following conditions:

(i) - Assume an admissible input $\mathbf{u}$ is defined on the interval $[t_0,t_f)$, assume $t_f > t_0+1$ and consider t_i such that $t_f > t_i > t_0$. Then both the restrictions of $\mathbf{u}$ on $[t_0,t_i)$ and on $[t_i,t_f)$ are admissible inputs.

(ii) - Assume that two admissible inputs $\mathbf{u}_1$ and $\mathbf{u}_2$ are defined, respectively on the intervals $[t_0,t_i)$ and $[t_i,t_f)$ for some $t_f > t_i > t_0$. Then the function $\mathbf{u}$ on $[t_0,t_f)$ defined by $\mathbf{u}(t) = \mathbf{u}_1(t)$ if $t \in [t_0,t_i)$ and $\mathbf{u}(t) = \mathbf{u}_2(t)$ if $t \in [t_i,t_f)$ is an admissible input.

(iii) For any interval $[t_0,t_f)$ the identically zero function is admissible.

The first remark in order at this point is that assumption (iii) will be at time (and for a special case) weakened in such a way that, for the purposes of the question under study, the effect of the milder assumption is the same as that of the original one. We do not introduce any terminological distinction and stipulate that (iii) is in force whenever a substitute assumption is not explicitly stated.

A constraints system satisfying the above assumption will be called a dynamical constraint system. We shall come back to this property in the definition below.

Next we need to introduce the concept of stationary constraint system. For this purpose, for any given integer T, consider the shift operator $S(t_0,T)$ defined on $U(t_0,t_f)$ by:

$$(S(t_0,T)\mathbf{u})(t) = \mathbf{u}(t-T) \qquad \forall\, \mathbf{u} \in U(t_0,t_f) \qquad t = t_0+T,\ldots,t_f+T \qquad (3.2.5)$$

The operator $S(t_0,T)$ is linear and invertible and maps

$U(t_0,t_f)$ onto $U(t_0+T,t_f+T)$.

At this point we collect in the following definition a number of important concepts relating constraint systems and constrained systems.

DEFINITION 1: A constraint system is stationary (or time invariant) if an input function $\mathbf{u}(t_0,t_f)$ is admissible when and only when the input function $S(t_0,T)\mathbf{u}(t_0,t_f)$ is admissible for all integer T. A constrained dynamical system is called stationary if both the corresponding dynamical system and the constraint system is stationary. A constrained system is called linear if both the dynamical system and the constraint system are linear. A dynamical system associated with a dynamical system of constraints is called a dynamical constrained system.

In this definition we refer to properties of both the components of a constrained system for the sake of generality, but of course, we have already assumed, for the sake of simplicity, that the dynamical system is both linear and stationary. Notice, however, that we may still consider non stationary and/or non linear and /or non dynamical constrained systems. This is obtained by associating to our linear dynamical system a constraint system which is not stationary and/or linear and/or dynamical.

Another interesting remark is that, to define a stationary constraint system it is not required that the constraint relation should be independent of t_0 and t_f. For example consider the constraint system:

$$\sum_{t_0}^{t_f} u(t) \leq (t_f-t_0)\, v \qquad (3.2.6)$$

As the reader will immediately verify this is a special case of stationary linear constraint system.

At this point, before entering more in depth in the theory of constrained systems it is convenient to briefly recall some basic ideas underlying the classical unconstrained reachability theory.

3.3 - Review of unconstrained reachability theory

We initiate recalling the classical definition of reachability. Even though our dynamic system is stationary we refer reachability to a given instant of time in preparation of the constrained case where, as explained above, time variant systems may well occur.

DEFINITION 2 - With reference to a linear dynamic system, we call a point (state) z of $\mathbb{R}^n$ reachable at time t_s if there exists a instant of time t_r, $t_r < t_s$ and an input function u defined in $[t_r,t_s)$ such that the solution x(.) of the system corresponding to initial condition $x(t_r) = 0$ and to the input u assumes the value z at time t_s, i.e. $x(t_s) = z$. The set of all states reachable at t_s is called the reachable space at t_s. If all states are reachable at t_s the system is called reachable at t_s too.

Notice that the definition of reachability is characterized by two main facts. The first is that the initial condition is fixed to be the origin of the vector state space. The second is that reachability is a feature of possible evolutions of the system, occurring in the past, relative to the instant of reachability.

An important observation, which provides a key argument in the analysis of reachability, is that, if a state z can be reached at t_s starting from zero at time t_r, then it can also be reached starting from zero at any instant of time, say t_p, prior to t_r. In fact it suffices to apply a zero input in the interval $[t_p,t_r)$ to the system with initial condition $x(t_p) = 0$, thereby obtaining $x(t_r) = 0$, and then to concatenate this zero input with the input that steers the state from zero at t_r to z at t_s. In other words the set of states reachable at t_s starting (from zero state) at

$t_r < t_s$ is contained in the set of states reachable at the same time starting (from zero state) at earlier times.

The set of states reachable from zero state starting at one step of time ahead of t_s, starting at two steps of time ahead etc. form an increasing family of linear subspaces. Their union is the reachable space at time t_s. Because we obtain linear subspaces of a finite dimensional linear space, such sequence of linear subspaces can only increase up to a certain point and then it will become constant. In view of stationarity the set of reachable states does not depend on time and we can equate the first subspace to the space of states reachable at time one starting from state zero at time zero, the second to the space of states reachable at time two starting from the zero state at time zero and so on. Clearly the first linear subspace is the range of the matrix B, the second is the range of the block matrix (B AB) and so on.

Finally combining the above observations with the Cayley Hamilton theorem we arrive to the conclusion that the set of reachable states is the linear subspace of the state space given by the range of the matrix $\mathcal{C}(0,n)$, which can be more simply denoted by $\mathcal{C}(n)$.

At times we are interested to the possibility of reaching a state starting from another state, different from the origin. There is a simple link between reachability from the origin and reachability from a state, say x. In fact we can reach the state z at t_s from state x at t_r, if and only if we can reach from the origin that state $z\text{-}A^{(t_s - t_r)}$.

3.4 - Reachability concepts for constrained systems

Of course we might think to adopt the same definition of reachable state and reachable set at a certain time as in the unconstrained case. However, this would not be enough for constrained systems for a number of reasons.

We face now a radically different state of affairs. To mention a few novelties, the finite time reachability property does not hold anymore. What is reachable in finite time may be quite different from what is reachable in infinite time. Moreover the constrained system may not be stationary even though the unconstrained system is. Thus, for example, reachability ahead in time may be different from reachability from beforehand. Even the property that if we decrease the first extreme of the time interval the reachable set grows is missing if we allow for non dynamical constraint systems.

Consequently a more refined definition is advisable in order to capture that greater complexity. The following definition formalizes a concept that already came to the fore in the arguments underlying unconstrained reachability theory.

DEFINITION 3 - A state z of a system (either constrained or unconstrained) is reachable at time t_s from time t_r (or from time t_r in $t_s - t_r$ steps) if there exists an input (an admissible input in the constrained case) such that the solution corresponding to initial condition $x(t_r) = 0$ and to such an input assumes the value $x(t_s) = z$.

The set of reachable states at time t_s is the union of the set of reachable states from $t_s - 1$, $t_s - 2$ etc.

Note that such family of state space subsets is not in general an increasing one. Moreover if the state trajectory starting from zero state zero at t_r assumes a certain value w at an intermediate time t_i ($t_r < t_i < t_s$) it is not necessarily true that the state w is reachable at time t_i from time t_r. Both these properties do instead hold in the case of dynamical constrained systems. The proof of this facts becomes trivial if one bears in mind the arguments on which were based our analysis of the unconstrained case. These observations enlighten the role of the dynamicity assumptions.

In presence of constraints for the only input (and, possibly, disturbance), the previously illustrated link between reachability and reachability from a state holds good. This would not be true in general if the state were involved in the constraint system too.

This review of basic reachability properties does not exhaust all the interesting fields of investigation. Many other properties could be considered than there is space to cover here. However, we do treat the case of approximate robust reachability in presence of noise. Appropriate definition and fundamental results will be given in an apposite section below.

3.5 - Pointwise in time constraints

A simple but practically interesting particular form of the constraints is that of pointwise (in time) constraints for system's variables, where, at each instant of time, the value of the variable is forced to belong to a (nonvoid) set, that may be fixed or vary in time. Such constraints, for the case of the only input, are expressed by:

$$u(t) \in W(t) \quad \forall\ t \tag{3.5.1}$$

where W(t) is a nonvoid set (called the constraining set, while W(.) will be called the constraining function). The function W(.) may well be a constant function. To represent absence of constraints at a certain instant t, it suffices to put W(t) equal to the whole space of input values.

Clearly, for the constraints system to be dynamical it suffices that the origin belong to W(t) for any t. We shall weaken at times this assumption, in such a way, though, to surrogate the effects of the dynamicity hypotheses. Note also that the constrained system will be stationary if and only if W(.) is a constant function (whose value will be denoted by W).

In the stationary case the reachable set will be denoted by R_W whereas in the time varying case it will be denoted by $R_W(t)$. The symbol R without subscript will denote the unconstrained

system's reachable set. Obviously, whatever is W(.) (or W in the stationary case), $R_W(t)$ (or R_W) is contained in R.

4 - REACHABILITY UNDER GENERAL TIME-POINTWISE CONSTRAINTS

Most of the results of this and the next section were stated in [22] and [23][1] .

Our first concern is to study how set operations on the constraining sets reflect on the reachable set. We can define any operation on the functions of the form W(.) by the corresponding operations on values of the functions. That is for example:

$$(W_1(.) \cap W_2(.))(t) = W_1(t) \cap W_2(t) \quad \text{for any } t$$

With this premise we can state the following theorem:

THEOREM 1.

(i) - If $BW_1(t) \subset BW_2(t)$ for any t then $R_{W_1}(t) \subset R_{W_2}(t)$ for any t.

(ii) - Let A be a nonvoid set and $\{W_\alpha : \alpha \in A\}$ (briefly $\{W_\alpha\}$) be an arbitrary family of constraining functions. Then for any t:

$$R_{\cap\{W_\alpha\}}(t) \subset \cap \{R_{W_\alpha}(t)\}$$
$$R_{\cup\{W_\alpha\}}(t) \supset \cup \{R_{W_\alpha}(t)\}.$$

(iii) - For any constraining function W(.) and real a, for any finite family $\{W_1(.),\ldots,W_k(.)\}$ of constraining functions and $\{a_1,\ldots,a_k\}$ of reals, and for any t:

---

[1] More specifically, Theorems 1, 2, 3, 4, 5 and Lemma 1 are adapted from [22], with kind permission from Pergamon Press Ltd, Headington Hill Hall, Oxford OX3 OBW, UK.

$$R_{aW}(t) = aR_W(t).$$

$$R_{\sum a_i W_i}(t) \subset \sum R_{a_i W_i}(t)$$

In the latter relation equality prevails if $0 \in ABW_i(t)\ \forall\ t,i$.

PROOF -The first two statements have a straightforward proof, and so does the first statement in (iii). Therefore, if we prove that for any W_1 and W_2

$$R_{W_1+W_2}(t) \subset R_{W_1}(t) + R_{W_2}(t)$$

with equality prevailing if 0 belongs to both $ABW_1(t)$ and $ABW_2(t)$ for any t, then the rest of the proof will follow rather directly.

Actually if $z \in R_{W_1+W_2}(t)$ then there exists an input u starting at some time $\bar{t} < t$ having the form $u(t) = u_1(t) + u_2(t)$ with $u_1(t) \in W_1(t)$ and $u_2(t) \in W_2(t)$, such that the corresponding solution x(.), with $x(\bar{t})=0$, satisfies $x(t) = z$. At this point, that z belongs to the r.h.s. set immediately follows from the fact that the forced solution is linear with respect to the input function.

Conversely if $z \in R_{W_1}(t) + R_{W_2}(t)$, that is $z = z_1 + z_2$ with $z_1 \in R_{W_1}(t)$ and $z_2 \in R_{W_2}(t)$, then let u_1 be the input (compatible with the constraint defined by W_1) that steers the state from 0 to z_1 starting at time $t_1 < t$, and similarly let u_2 be the input corresponding to z_2, which will start at time $t_2 < t$. If $t_1=t_2$ then the control u_1+u_2, compatible with the constraint defined by W_1+W_2, will steer the system from the zero state at time t_1 to the state z at time t in view of the linearity of the solution. Otherwise,

assume, without restriction of generality, that $t_1 < t_2$. Consider an input u' that coincides with u_1 in the interval $[t_2, t)$, and is such that $u'(t) \in W_1(t)$ and $ABu'(t)=0$ for all t in the interval $[t_1,t_2)$, which is possible in view of the hypothesis. Then $u'+u_2$ will steer the system from state zero at time t_1 to the state z at time t, and the proof is therefore concluded. ∎

Notice that $W_1 \subset W_2$ implies $BW_1 \subset BW_2$. We also remark that the inclusion relations in the statement ii of Theorem 1 may occur in proper sense, as will be shown by the following example:

EXAMPLE 1

Consider a discrete time linear system described by the equations:

$$x_1(t+1) = -mx_2(t) + u_1(t)$$
$$x_2(t+1) = mx_1(t) + u_2(t) \qquad 0<m<1$$

Let W_1 and W_2 be the following sets in input space

$$W_1 = \{ (u_1,u_2): -1 \le u_1 \le 1, \quad u_2 = 0 \}$$
$$W_2 = \{ (u_1,u_2): u_1 = 0, \quad -1 \le u_2 \le 1 \}$$

It can be verified that

$$R_{W_1} = \{(x_1,x_2): -1/(1-m^2) < x_1 < 1/(1-m^2), -m/(1-m^2) < x_2 < m/(1-m^2)\}$$

$$R_{W_2} = \{(x_1,x_2): -m/(1-m^2) < x_1 < m/(1-m^2), -1/(1-m^2) < x_2 < 1/(1-m^2)\}$$

$$R_{W_1} \cap R_{W_2} = \{(x_1,x_2): -m/(1-m^2) < x_1 < m/(1-m^2), \; -m/(1-m^2) < x_2 < m/(1-m^2)\}$$

$$R_{W_1 \cap W_2} = \{ (0,0) \}$$

It is also easy to verify that the set $R_{W_1} \cup R_{W_2}$ is properly contained in the set $R_{W_1 \cup W_2}$.■

The condition $0 \in A\ B\ W_i(t)$ is essentially a condition of dynamicity, and the theorem confirms the importance of this concept. A similar comment will apply to most of the results that follow. Of course such condition is weaker than the condition that $0 \in W(t)$. An intermediate possibility, which will also be used, consists in assuming that $0 \in B\ W(t)$. In this respect one can imagine cases, in which, even though $0 \notin A\ B\ W(t)$, there is no harm in directly adding the origin to the constraining set (or to its image under the operator A B or B), so to keep the dynamical nature of the system.

We now pass on to consider the stationary case. Let us, in the first place, deal with the problem of computing the reachable set. The unconstrained reachability formula:

$$R = \text{Range}(B\ AB\ \dots\ A^{n-1}B) \tag{4.1}$$

can be rewritten in the form

$$R = \sum_{i=0}^{n-1} A^i B\ U = \lim_{k \to n-1} \{ \sum_{i=0}^{k} A^i B\ U \} \tag{4.2}$$

This result has been generalized for the case of cone constrained inputs in [22]. For general time-pointwise constraints we can state the following theorem, in which we deal also with finite time reachability. To this effect we denote by R_{W_k} the set of states reachable from the origin in at most k steps of time.

THEOREM 2.

The set R_{W_k} can be expressed as:

$$R_{Wk} = \sum_{i=0}^{k-1} A^i B\ W \tag{4.3}$$

Moreover, if $0 \in A\ B\ W$ then

$$R_W = \lim_{k\to\infty} \{ \sum_{i=0}^{k-1} A^i B\ W \} \tag{4.4}$$

The sequence of sets being actually an increasing sequence.

Finally, if $0 \in B\ W$, then $0 \in R_W$, $A^i\ B\ W \subset R_W\ \forall i$, R_W is invariant with respect to A and the set of all states reachable from states belonging to R_W is contained in $R_W + R_W$.

In the statement of this theorem the usual mathematical definition of the limit of an increasing sequence as union of the sequence itself is adopted.

PROOF - The first statement that requires a non trivial proof is that of invariance. If $x \in R_W$, then for some k

$$x \in \sum_{i=0}^{k} A^i BW \quad \text{and hence}$$

$$Ax \in A \sum_{i=0}^{k} A^i BW = \sum_{i=1}^{k+1} A^i BW$$

but since $0 \in BW$

$$\sum_{i=1}^{k+1} A^i BW \subset \sum_{i=0}^{k+1} A^i BW \subset R_W$$

To prove the last statement decompose the response in the sum of the free and forced response. The first, in view of the just

proved invariance, remains in R_W. The same is true for the second by definition of R_W and the assumption of stationarity. Thus the desired conclusion follows.

The statement of this theorem highlights some differences with the unconstrained case. The most noteworthy of these is that an unconstrained system has a finite time reachability property (that is, if $W=\mathbb{R}^n$ then the sequence increases at most up to the nth term), which does not hold in general.

The next natural question to ask regards how do the properties of R_W depend on the properties of W. In this respect recall that an operator A is a contraction if $\|A\| \leq 1$. Let us stipulate that A is a proper contraction if $\|A\| < 1$. With these premises we can state the following:

THEOREM 3.

Assume still that $0 \in A\ B\ W$ and that k is any positive integer then

(i) - If BW is convex then both R_{Wk} and R_W are convex.

(ii) - If B W is bounded and A is a proper contraction then R_W is bounded. If BW is unbounded then R_{Wk} is unbounded and hence such is R_W.

(iii) - If W has interior then both R_{Wn} and R_W have interior relative to the subspace R.

(iv) - If W is open then both R_{Wn} and R_W are open relative to R.

(v) - If B W is a subspace then both R_{Wk} and R_W are subspaces and $R_{Wk} = R_W$ for any $k \geq n$. Moreover R_W is the minimal subspace, that is invariant under A and contains the subspace BW.

(vi) - If B W is a (convex) cone then both R_{Wk} and R_W are cones. Moreover R_W is the minimal cone, that is invariant under A and contains the cone BW.

(vii) - If B W is a group under addition then both R_{Wk} and R_W are groups under addition. Moreover R_W is the minimal subgroup of $\mathbb{R}^n$, that is invariant under A and contains the group BW.

PROOF - For the sake of brevity we outline only a few crucial arguments for the proof. From these, from the previous results and from standard arguments used in linear reachability theory it is not difficult to build complete proofs.

The proof of (i) follows immediately from the expression of the reachable sets given in Theorem 2 and elementary computation rules for convex sets (see e.g. [25]).

As to (ii) note that, because BW is bounded, for some positive real r, $BW \subset S^r$, where S^r denotes the closed sphere about the origin with radius r. Thus $W \subset B^{-1}S^r$ (where B^{-1} is the inverse image function), so that, by Theorem 1, $R_W \subset R_{B^{-1}S^r}$. On the other hand:

$$\sup_{k-1}\{\|x\|: x \in R_{(B^{-1}S^r)}\} =$$

$$= \sup\{\|\sum_{k-1}^{i=0} A^i Bu(k-i-1)\| : u(i) \in B^{-1}S^r\} =$$

$$= \sup\{\|\sum_{k-1}^{i=0} A^i z(k-i-1)\| : z(i) \in S^r\} \leq$$

$$\leq \sup\{\sum_{k-1}^{i=0} \|A^i\| \, \|z(k-i-1)\| : z(i) \in S^r\} \leq$$

$$\leq r \sum_{i=0}^{k-1} \|A^i\| \leq r \frac{1}{1 - \|A\|}$$

But because the latter inequality holds for any k, the desired conclusion is immediate.

Finally if B W is unbounded then R_{Wk} and R_W must be unbounded too, since, under the present hypothesis that $0 \in A\ B\ W$, $B\ W \subset R_{Wk} \subset R_W$.

As to (iii), notice that if R_{Wn} has interior then certainly so does R_W, because $R_{Wn} \subset R_W$. Next, because $x(k) = C(k)\ u(k)$, it is apparent that R_{Wn} is the image under $C(n)$ of the set $W^n \subset \mathbb{R}^{pxn}$, which by hypothesis and well known elementary facts of vector topology has interior. But, because of the further topological fact that any linear map with finite dimensional domain is relatively open and because the range of $C(n)$ is R, the desired conclusion follows.

We omit the proof of (iv). It is largely similar to the previous one and is anyway based on elementary topological arguments.

The proof of (v) leans on standard arguments of reachability for linear systems.

As to statement (vi), it is clear from Theorem 2 that if B W is a cone then R_{Wk} and R_W are cones too. By the same theorem R_W is invariant under A. On the other hand if a cone is invariant under A and contains B W, then it must contain all the sets in the series that sum up to R_W, whose expression is given in Theorem 2. Thus it must also contain the sum of the series i.e. R_W. This shows that R_W is actually the minimal cone invariant under A and containing B W.

Finally the proof of (vii) is rather straightforward application of by now usual arguments and can therefore be safely be omitted. ∎

Observe that if W is convex then also B W is convex. Similar remarks apply to statements (ii), (v), (vi) and (vii).

Later on we shall consider a few further properties of W. For the moment, because all the statements in this theorem are sufficiencies, one may wonder about necessity. Unfortunately none of these conditions is necessary. If any of them is negated then,

as is shown by the following examples, the corresponding property of R_W may or may not hold according to the cases.

EXAMPLE 2

Consider a discrete time linear system described by the equations:

$$\begin{aligned} x_1(t+1) &= m\, x_1(t) + m\, x_2(t) + u_1(t) \\ x_2(t+1) &= m\, x_1(t) + m\, x_2(t) + u_2(t) \end{aligned} \qquad m>0$$

For any t, the input is constrained to belong to the non convex set W, described as follows:

$$W = \{ (u_1,u_2): u_1=-1,\ 0\le u_2\le 1 \} \cup \{ (u_1,u_2): -1\le u_1\le 0,\ u_2=1 \}$$

It is easy to verify that both R_{Wk}, for every $k>1$, and R_W have interior; moreover, if $0.25 < m < 0.5$, R_{Wk} may be not convex for some k, while R_W is convex and bounded. In fact, if $0.25 < m < 0.5$:

$$R_W=\{(x_1,x_2):1\le -x_1+x_2\le 2,\ x_1 > -1-(m/(1-2m)),\ x_2 < 1+(m/(1-2m))\}$$

whereas, if $m\ge 0.5$:

$$R_W =\{(x_1,x_2):\ 1 \le -x_1+ x_2 \le 2\}$$

This example shows that conditions (i) and (iii) of the statement of Theorem 3 are not necessary. ∎

EXAMPLE 3

Consider a discrete time linear system described by the equations:

$$\begin{aligned} x_1(t+1) &= -m\, x_2(t) + u(t) \\ x_2(t+1) &= m\, x_1(t) + u(t) \end{aligned} \qquad 0<m<1$$

with the constraint

$$-1 \le u(t) \le 1 \quad \forall t$$

The reachable set R_w is the rectangle:

$$R_w = \{(x_1,x_2): -2\, m/(1-m^2) < -x_1 + x_2 < 2m/(1-m^2),$$
$$-2/(1-m^2) < x_1 + x_2 < 2/(1-m^2)\}$$

It is easy to verify that the set R_w is open (counter example to necessity of statement (iv) of Theorem 3). ∎

EXAMPLE 4

Consider the discrete time linear system described by the equations:

$$x_1(t+1) = -m\, x_1(t) + u(t)$$
$$x_2(t+1) = -m\, x_2(t) + u(t) \qquad m > 1$$

with the constraint

$$0 \le u(t) \le 1 \quad \forall t$$

$$R_w = \{ (x_1,x_2) : x_1 = x_2 \}$$

This example shows that the condition (vi) in Theorem 3 is not necessary (the set W is a polytope while the set R_w is a subspace). Moreover, because a subspace is a convex cone, it also follows that the condition (vi) in the same theorem is not necessary. ∎

EXAMPLE 5

Consider the system described by the equation:

$$x(t+1) = x(t) + u(t)$$

with input constraining set $W = \{0, 1\}$ for every t.

The reachable set R_W is the set of all integers, and hence a group under addition. This fact shows that the condition (vii) of Theorem 3 is not necessary.

A few further important remarks on the theorem are in order. As a special case note that if the system is reachable and W has interior then R_W has interior, whereas if the system is not reachable R_W cannot have interior even if W does, since R_W is contained in R.

Moreover if W is convex and $0 \in B$ W and R_W is unbounded then R_W contains a convex cone. Actually in this case R_W is an unbounded convex set containing the origin and hence the recession cone of R_W is a nontrivial cone and is also the maximal convex cone contained in R_W (see [25]).

Note also that if a power of A is zero (that is, A is nilpotent) then any power with a greater exponent will be zero (exploiting the Jordan form of a matrix, it is not difficult to prove that the minimum power of A which is zero is at most the n-1th). Hence if a property of W is not inherited by the finite time horizon reachable sets, this also excludes that in general it is inherited by R_W. Conversely it may happen that a property is not in general inherited by R_W, but it is inherited by the finite time reachable sets.

It is convenient to mention some obvious negative cases for which a property of W is not inherited by R_W. This is the case when W is a closed set or when W is a sphere of a given norm or when W is a nonlinear manifold (actually in general neither the image under a linear map of a manifold nor the sum of two manifolds is a manifold).

5 - REACHABILITY UNDER POLYHEDRAL CONSTRAINTS

In this section, for simplicity, we make reference to properties of W, but it is clear that, as in the previous section, some generalization can be achieved making instead reference to B W.

On the base of the first statement of Theorem 2, and the fact that a sum of polyhedra is a polyhedron, it is clear that in the present case R_{Wk} is a polyhedron.

As is well known [26], if W is a polyhedron then

$$W = P + L + C \tag{5.1}$$

where P is a polytope, L is a linear subspace and C is a pointed polyhedral cone. Regarding this decomposition we can state the following

LEMMA 1.
The polytope P can be chosen to contain the origin if and only if W contains the origin.

PROOF - In fact assume that W does not contain the origin. Then if P contains the origin it would follow that so does P+L+C, which is a contradiction. Conversely suppose that W contains the origin but P does not. Then because both {0} and P are contained in W, which is convex, it is possible to consider the decomposition:

$$W = C(\{0\} \cup P\} + L + C$$

where the polytope $C(\{0\} \cup P)$ contains the origin.

At this point if $0 \in W$, we can choose P according to LEMMA 1, and write, in view of Theorem 1

$$R_{Wk} = R_{Pk} + R_{Lk} + R_{Ck} \tag{5.2}$$

$$R_{W} = R_{P} + R_{L} + R_{C} \tag{5.3}$$

Therefore we can look separately at the cases where W is a linear subspace or a polyhedral cone or a polytope. The significant cases are those of a polytope and of a pointed polyhedral cone.

Because a polytope is a finitely generated structure, it is not preserved by system's dynamics. However if we consider the special case of finite time horizon reachability, which has foremost practical importance (in particular in optimization problems), then the polytopic structure is preserved:

THEOREM 4.

If W is a polytope then R_{Wk} is a polytope for any k.

PROOF - We know from Theorem 2 that

$$R_{Wk} = \sum_{i=0}^{k-1} A^i BW$$

Each set in the sum is the image under a linear map of a polytope and hence is a polytope [25]. Moreover a finite sum of polytopes is a polytope (see again [25]), and therefore the desired conclusion has been achieved. ∎

An upshot of the theory of systems over polytopes is the result below, which can be considered as the generalized bang-bang principle for discrete time systems. For a continuous time version of the bang-bang principle see e.g. [2].

Let us denote by $\{e_j: j=1,..,m\}$ (briefly $\{e_j\}$) the set of extreme points of W and by E the set of all functions on $\{0,1,...\}$ to $\{e_j\}$ These functions play the role of controls with bang-bang values. Then we can state the following:

THEOREM 5.

For any k the extreme points of R_{Wk} have the form:

$$\sum_{i=0}^{k-1} A^{k-i-1} Bu(i) \qquad \text{with } u \in E$$

or, in other words, the set of extreme points of R_{Wk} is contained in the reachable set $R_{\{e_j\}k}$.

PROOF - Since $\{e_j\}$ is the set of extreme points of W, so that $W = C(\{e_j\})$, it follows that $A^iB\ W = C(A^iB\{e_j\})$. From this fact and the expression of R_{Wk} it follows:

$$R_{Wk} = \sum_{i=0}^{k-1} A^iBW = \sum_{i=0}^{k-1} C(A^iB\{e_j\}) = C\left(\sum_{i=0}^{k-1} (A^iB\{e_j\})\right)$$

where in the last passage we have exploited the elementary result which ensures that, if A and B are arbitrary sets, then $C(A+B) = C(A)+C(B)$. The desired conclusion is now an immediate consequence of the very definition of sum of sets. ∎

5.1 - An optimization example with an illustration of the bang bang principle

Optimization is not part of our concern here, but in this case an illustration of how the bang bang principles applies to optimization problems is just a few lines away.

We consider here a simple functional and constraints that lead to an immediate solution by inspection. The structure of the solution will demonstrate the bang bang principle.

Suppose that we want to maximize the functional (f, x(T). Substituting the solution of the dynamic system we obtain:

$$(f,\ x(T)) = (f, C(T)\ u(T)) + (f,\ L(T)x(0)) \qquad (5.1.1)$$

The second term in the r.h.s. is constant and thus does not intervene in the optimization. The first term can be rewritten as:

$$(f,\ C(T)\ \mathbf{u}(T)) = (C^{*}(T)f,\ \mathbf{u}(T)) \qquad (5.1.2)$$

Thus if we partition the vector $C^{*}(T)\ f$ in T blocks g_i $(i = 0,..T\text{-}1)$ corresponding to those of the vector $\mathbf{u}(T)$ and bear in mind that the blocks of $\mathbf{u}(T)$ are independently constrained by $u(i) \in W$ it is clear that the problem diagonalizes into the T optimization problems:

$$\begin{aligned} &\max\ (g_i,\ u(i)) \\ &\text{subject to } u(i) \in W \qquad\qquad i = 1,..,T\text{-}1 \end{aligned} \qquad (5.1.3)$$

Next suppose that the constraints be polytopical, e.g. of the form (box constraints):

$$m_i \leq u(i) \leq M_i \qquad i = 0,..,\ T\text{-}1 \qquad (5.1.4)$$

Then the optimum solution is clearly given by:

$$u(i)_j = \begin{cases} m_{ij} & \text{if } g_{ij} < 0 \\ \text{any value} & \text{if } g_{ij} = 0 \\ M_{ij} & \text{if } g_{ij} > 0 \end{cases} \qquad (5.1.5)$$

Because the maximum of the functional is surely attained on an extreme point of the reachable set at T and because the solution has the form contemplated by Theorem 5, this example confirms the bang bang principle.

Finally we notice that the same arguments apply to the computation of the solution in the more general case in which the functional has the form:

$$\sum_{i=1}^{T} (f_i,\ x(i)) = (f,\ \mathbf{x}(T)) \qquad (5.1.6)$$

where, of course

$$f = \begin{bmatrix} f_1 \\ \vdots \\ f_T \end{bmatrix}$$

A practically verbatim repetition of the above steps leads to the diagonalization of the problem and to the solution. We leave the details to the reader for the sake of brevity.

6 - SYSTEMS OVER CONES

We have already touched upon the case in which the input values are constrained to belong to a cone in Theorem 3, where we have introduced the theory of minimal invariant cones.

If we introduce conical constraints for the other system variables too, then the theory extends in various directions. In [22], besides the above basic reachability result, conditioned and controlled invariant cones are introduced and their application to state constrained reachability theory is illustrated. Another interesting direction of investigation is that of positive systems (see e.g. [27] for the continuous time case). Some recent developments for the same case are in [28]. A generalization of the concept of positive system for discrete time systems (but the same concepts - if not the results - apply immediately to the continuous time case) is in [5].

Here we wish to complete the case of input conical constraints along lines that parallel the case of polyhedral constraints.

First of all we observe that any convex cone C in $\mathbb{R}^n$ is the sum of its lineality subspace L plus a pointed cone given by $L^{\perp} \cap C$. Because both of these two sets contain the origin, if we constraint the input to belong to a fixed cone at any time, then both the finite time reachable set and the reachable set decompose in the sum of the reachable sets corresponding to the subspace (which is a subspace) and that corresponding to the cone (which is a cone), according to Theorems 1 and 3.

Next suppose that the input constraining cone C is polyhedral (that is, both a cone and a polyhedron at the same time). Notice that a linear subspace is a polyhedral cone and the intersection of two polyhedral cones is a polyhedral cone. Thus in the above decomposition the pointed cone $L^{-} \cap C$ is polyhedral too. Hence it is natural to complete our treatment examining the case of pointed polyhedral conical constraints. Incidentally notice that the nonnegative orthant of the space (i.e., the set of all vectors with non negative components) is a pointed polyhedral cone. Thus any theory of positive systems is a special case of the theory of systems over pointed polyhedral cones.

A major fact regarding pointed polyhedral cones is that they are in a way the unbounded counterpart of polytopes. In fact in the same way as we can say that a polytope is the convex extension of the set of its extreme points, we can affirm that a pointed polyhedral cone is the convex extension of the union of its extreme rays. An extreme ray of a cone is a ray which is also a face of the cone. This result is, e.g., in [25]. For example the non-negative orthant is the convex extension of the coordinate axes. It may be more natural to use conical instead of convex extension. Thus let a minimal generating set be a set obtained taking a non-zero vector from each extreme ray of the cone. Then the cone is the conical extension of any minimal generating set.

This similarity carries on, to same extent, to finite time reachable set. The following theorem is the counterpart of Theorem 5.

THEOREM 6.

If W is a pointed polyhedral cone then, for any k, R_{Wk} is a polyhedral (not necessarily pointed) cone and, if $\{e_j\}$ is a minimum generating set of W the cone R_{Wk} has the form:

$$Co\left(\sum_{i=0}^{k-1} A^{k-i-1}Bu(i):\ u(i) \in \{e_j\}\right)$$

PROOF - Since $W = Co(\{e_j\})$, it follows that $A^iB\ W = Co(A^iB\{e_j\})$. From this fact and the expression of R_{Wk} it follows:

$$R_{Wk} = \sum_{i=0}^{k-1} A^iBW = \sum_{i=0}^{k-1} Co(A^iB\{e_j\}) = Co(\sum_{i=0}^{k-1} (A^iB\{e_j\}))$$

where in the last passage we have exploited the elementary result which ensures that, if A and B are arbitrary sets, then $Co(A+B) = Co(A)+Co(B)$. The desired conclusion is now an immediate consequence of the very definition of sum of sets. ∎

We do not get involved here in optimization concepts though, because this would take us too far away.

7 - CONSTRAINED STATE APPROACH TO THE CONSTRAINED INPUT REACHABILITY THEORY

In this section we develop a technique to compute the finite time reachable set of a given system, when the input constraining set is a polyhedron, in general time-varying. Here, to fix the ideas, we consider the set of states reachable at time $T>0$, starting from time 0. This set will be denoted, for simplicity, by $\mathcal{R}_T$. An argument similar to that at the beginning of Sec. 5 will immediately show that $\mathcal{R}_T$ is a polyhedron.

The theory is based on a dual conical condition of nonvoidness of a polyhedron. This condition is parameterized with respect to the bound vector of the inequalities, which describe the polyhedron in question. (see [29] and [30]).

The idea is that of considering the unknown reachable set $\mathcal{R}_T$ as a constraining set for the state at the same time T. By means of a backward recursion ([4],[5]), we find the description, at each step, of the set of the states admissible (that is, for which a

solution exists) with respect to this fictitious state constraint and with respect to the constraints on the input. By imposing that the zero vector in the state space belongs to the admissible polyhedron at time t=0, we arrive at giving the expression of $\mathcal{R}_T$.

It is important to stress that in our approach no assumption is required: neither on properties of the matrices of the linear system, nor on particular structures of the constraining sets.

Before illustrating this method, some preliminaries are needed.

We shortly describe the solution of the feasibility problem (i.e. the problem of existence of solutions), where both the input and the state are constrained to belong to given sets, and then we particularize the results, to solve the problem of finding the reachable set from the origin, when the only input is constrained.

Consider the system (3.1.1) with D=0. For reasons that will be soon apparent, it is also convenient to consider an initial time t_i, an initial state $x(t_i) = \bar{x}$, and a final time t_f, with $0 \le t_i \le t_f \le T$. We assume that the state of the system is constrained in a polyhedral set C_t for all t in the interval [0, T]:

$$x(t) \in C_t \qquad \forall\, t \in [0,\ T] \tag{7.1}$$

or also, equivalently

$$G(t)x(t) \le M(t) \qquad 0 \le t \le T \tag{7.2}$$

On the other hand the input of the system is constrained in a polyhedron W_t for all t in the interval [0, T-1]:

$$u(t) \in W_t \qquad \forall\, t \in [0,\ T\text{-}1] \tag{7.3}$$

or

$$F(t)u(t) \leq V(t) \qquad 0 \leq t \leq T-1 \qquad (7.4)$$

DEFINITION 4: The problem is feasible, relative to an initial state $\bar{x}$ at t_i and to t_f, if there exists an input sequence u defined on $[t_i, t_f-1]$, $u(t) \in W_t$, such that the above state constraints are satisfied by the solution of the system.

We shall call any state, with respect to which the system is feasible, an admissible state, relative to the pair of times (t_i, t_f), and the given constraints.

The computation of the set of admissible states is based on the following general backward recursion for the set of admissible states relative to initial times T-1,..,0 and to the final time T. Let D_{T-1} be the set of admissible states relative to (T-1,T) with respect to the constraints $x(T) \in C_T$, $x(T-1) \in \mathbb{R}^n$ (that is, x(T-1) unconstrained) and $u(T-1) \in W_{T-1}$. Then it is clear that the set of states admissible relative to (T-2,T) with respect to the constraints $x(T) \in C_T$, $x(T-1) \in C_{T-1}$, x(T-2) unconstrained, $u(T-1) \in W_{T-1}$ and $u(T-2) \in W_{T-2}$ is nothing but the set of states admissible relative to (T-2,T-1) with respect to the constraints $x(T-1) \in E_{T-1}=D_{T-1} \cap C_{T-1}$, x(T-2) unconstrained and $u(T-2) \in W_{T-2}$.

Generalizing to arbitrary $t<T$, denote by D_t the set of admissible states relative to (t,T), with respect to the constraints $x(\tau) \in C_\tau$ and $u(\tau) \in W_\tau$, $t \leq \tau < T$, being $D_T=C_T$, and let $E_t = D_t \cap C_t$, with $E_T=C_T$.

At each instant t of the above backward recursion we must solve the following problem: find the set of admissible states relative to (t-1,t) with respect to the constraints $x(t) \in E_t$, x(t-1) unconstrained and $u(t-1) \in W_{t-1}$. This admissible set D_{t-1} is given by:

$$D_{t-1} = \{x: \exists\, u \in W_{t-1} \text{ such that } Ax+Bu \in E_t\} \qquad (7.5)$$

To this purpose we recall now the dual conical nonvoidness condition [29], [30].

THEOREM 7.

The convex polyhedron $\{ x: Gx \leq v , G \in \mathbb{R}^{sxn}\}$ is nonvoid if and only if

$$Qv \geq 0 \tag{7.6}$$

where

{rows of Q} = {generators of the cone: $R(G)^{\perp} \cap P$}, $R(G)$ denotes the range of the matrix G and P denotes the nonnegative orthant of $\mathbb{R}^s$.

For the sake of simplicity, the same symbol P, which appears in the statement of Theorem 7 will be used in the sequel to denote the nonnegative orthant of any Euclidean space, leaving to the context the determination of the space itself.

At this point we can give an explicit expression for E_t.

THEOREM 8.

The set $E_t = D_t \cap C_t$ is described by the inequality:

$$E_t = \{ x: \hat{G}(t)\, x \leq \hat{M}(t) \} \tag{7.7}$$

where

$$\hat{G}(t) = \begin{bmatrix} Q'_{t+1}\hat{G}(t+1)A \\ G(t) \end{bmatrix}$$

$$\hat{M}(t) = \begin{bmatrix} Q'_{t+1}\hat{M}(t+1) + Q''_{t+1}V(t) \\ M(t) \end{bmatrix}$$

with terminal conditions:

$$\hat{G}(T) = G(T) \quad \text{and} \quad \hat{M}(T) = M(T)$$

where the matrix $Q_{t+1}=[Q'_{t+1} \vdots Q''_{t+1}]$ is defined by:

$$\{ \text{ rows of } Q_{t+1} \} = \{ \text{ generators of } R\begin{bmatrix}\hat{G}(t+1)B \\ F(t)\end{bmatrix}^{\perp} \cap P\}$$

and the blocks Q'_{t+1} and Q''_{t+1} are determined by the row dimensions of $\hat{M}(t+1)$ and $V(t)$.

PROOF:

The state constraint at T is:

$$G(T)\ x(T) \leq M(T)$$

Substituting for x(T):

$$G(T)\ (A\ x(T-1) + B\ u(T-1)) \leq M(T) \quad \text{or}$$

$$G(T)\ B\ u(T-1) \leq M(T) - G(T)\ A\ x(T-1)$$

Now, bearing in mind the original input constraint at T-1, we can write the inequality:

$$\begin{bmatrix}G(T)\ B \\ F(T-1)\end{bmatrix} u(T-1) \leq \begin{bmatrix}M(T)-G(T)\ A\ x(T-1) \\ V(T-1)\end{bmatrix}$$

In view of the dual nonvoidness condition (Theorem 7), the set of all bounds that make the latter equation feasible is given by:

$$Q_T\begin{bmatrix}M(T)\ -G(T)\ A\ x(T-1) \\ V(T-1)\end{bmatrix} \geq 0$$

where Q_T is the matrix, whose rows are the generators of the pointed polyhedral cone $R\begin{bmatrix} G(T)\ B \\ F(T-1) \end{bmatrix}^{\perp} \cap P$. Then, partitioning Q_T as $[Q_T' \vdots Q_T'']$, according to the dimension of M(T) and V(T-1), it is obtained:

$$Q_T' M(T) - Q_T' G(T) A x(T-1) + Q_T'' V(T-1) \geq 0 \text{ or}$$

$$Q_T' G(T) A x(T-1) \leq Q_T' M(T) + Q_T'' V(T-1)$$

The latter inequality represents the set of admissible states D_{T-1}, and, at the same time, shows that it is a polyhedron.

The set E_{T-1} is obtained intersecting D_{T-1} with the constraining set C_{T-1} and hence it is given by the polyhedron:

$$E_{T-1} = \{x: \hat{G}(T-1)x \leq \hat{M}(T-1)\}$$

where

$$\hat{G}(T-1) = \begin{bmatrix} Q_T' G(T) A \\ G(T-1) \end{bmatrix}$$

$$\hat{M}(T-1) = \begin{bmatrix} Q_T' M(T) + Q_T'' V(T-1) \\ M(T-1) \end{bmatrix}$$

At this point it is easy to see that generalizing the above formulas for the generic instant of time t, the desired expression of E_t is obtained. ∎

Next we pass on the reachability problem proper. To this purpose, it suffices to solve a special case of the above general constrained problem, defined by:

$$x(t) \in C_t \qquad \text{for all } t \in [0,\ T] \qquad (7.7)$$
$$0 \in D(0)$$

where

$$C_t = \{z\} \qquad t = T$$
$$C_t = \mathbb{R}^n \qquad \forall\ t \in [1,\ T\text{-}1]$$

Then $\mathscr{R}_T$ will be the set of all z that satisfy the above conditions.

The constraint at t=T may also be expressed as:

$$x(T) = z \qquad z \in \mathscr{R}_T$$

or also

$$x(T) \leq z$$
$$-x(T) \leq -z$$

and therefore

$$G(T) = \begin{bmatrix} I \\ -I \end{bmatrix} \qquad M(T) = \begin{bmatrix} z \\ -z \end{bmatrix}$$

$$G(t) = (0) \qquad M(t) = (0) \qquad t \in [1,T\text{-}1]$$

The set D(0) is described by the inequality (in this specific case D(0) is equal to E(0) and it is convenient to use the symbols introduced for this latter):

$$\hat{G}(0)x \leq \hat{M}(0)$$

where

$$\hat{G}(0) = Q_1' \hat{G}(1) A \qquad \hat{M}(0) = Q_1' \hat{M}(1) + Q_1'' V(0)$$

$$\hat{G}(1) = Q_2'\hat{G}(2)A \qquad \hat{M}(1) = Q_2'\hat{M}(2) + Q_2''V(1)$$

.....

$$\hat{G}(t) = Q_{t+1}'\hat{G}(t+1)\ A \qquad \hat{M}(t) = Q_{t+1}'\hat{M}_{t+1} + Q_{t+1}''V(t)$$

.....

$$\hat{G}(T) = G(T) = \begin{bmatrix} I \\ -I \end{bmatrix} \qquad \hat{M}(T) = M(T) = \begin{bmatrix} z \\ -z \end{bmatrix}$$

and where

$$\{ \text{ rows of } Q_{t+1} \} = \{ \text{ generators of } R \begin{bmatrix} \hat{G}(t+1)B \\ F(t) \end{bmatrix}^{\perp} \cap P \}$$

Note now that in the expression of the set D(0) the matrix $\hat{G}(0)$ depends on the known matrices of the dynamic constrained system, while the bound vector $\hat{M}(0)$ actually depends on the parameter z. Let us examine more in detail the structure of $\hat{M}(0)$

$$\hat{M}(0) = Q'\ \hat{M}(1) + Q''\ V(0)$$

substituting for $\hat{M}(1)$:

$$\hat{M}(0) = Q_1'Q_2'\hat{M}(2) + Q_1'Q_2''V(1) + Q_1''\ V(0)$$

and finally, after the last substitution for $\hat{M}(T-1)$, the following expression is obtained for M(0)

$$\hat{M}(0) = Q_1'Q_2'\ldots Q_T'\ \hat{M}(T) + Q_1'Q_2'\ldots Q_{T-1}'Q_T''\ V(T-1) + \ldots$$
$$+ Q_1'Q_2''\ V(1) + Q_1''\ V(0) \qquad (7.8)$$

At this point it is convenient to simplify the notations by means of the following positions:

$$
\begin{aligned}
Y_T &= Q'_1 Q'_2 \ldots Q'_{T-2} Q'_{T-1} Q'_T \\
Z_T &= Q'_1 Q'_2 \ldots Q'_{T-2} Q'_{T-1} Q''_T \\
Z_{T-1} &= Q'_1 Q'_2 \ldots Q'_{T-2} Q''_{T-1} \\
&\ldots\ldots \\
Z_2 &= Q'_1 Q''_2 \\
Z_1 &= Q''_1
\end{aligned}
\tag{7.9}
$$

Thus, remembering that $M(T) = \begin{bmatrix} z \\ -z \end{bmatrix}$, the bound vector $\hat{M}(0)$ can be rewritten as

$$
\hat{M}(0) = (Y'_T - Y''_T)\, z + \sum_{i=0}^{T-1} Z_{i+1} V(i) \tag{7.10}
$$

where Y_T has been partitioned as $[Y'_T \vdots Y''_T]$, according to the structure of $\hat{M}(T)$.

Clearly $0 \in D(0)$ if and only if $\hat{M}(0) \geq 0$. Hence, in view of the just found expression of $\hat{M}(0)$, we have established the following theorem:

THEOREM 9.

The set of reachable states $\mathscr{R}_T$ has the following expression:

$$
\mathscr{R}_T = \left\{ x: (Y''_T - Y'_T)\, x \leq \sum_{i=0}^{T-1} Z_{i+1} V(i) \right\} \tag{7.11}
$$

Some important remarks are now in order.

First of all note that the generalization to the case of initial state $x(0) \neq 0$ is straightforward. Moreover, with some additional computations, it is possible to describe the set of states reachable from a set of initial states [3].

The second and more interesting remark is the following: despite the numerical complexity of the computation of $\mathscr{R}_T$, an exact description of the reachable set is given in a very general case. Moreover, the main advantage of this approach is that the inequality which describes $\mathscr{R}_T$ is parameterized with respect to the vector bounds of the input constraining polyhedra. In fact the coefficient matrix $(Y''_T - Y'_T)$ and the matrices Z_{i+1}, $0 \leq i \leq T-1$, depend on the matrices of the constraints and the vector $\sum_{i=0}^{T-1} Z_{i+1} V(i)$ has the bounds V(i), i=0...T-1, as parameters. This means that with our technique we have solved the problem of finding the reachable set of an entire class of input constrained system, each element of the class corresponding to a set of bounds for the input constraining polyhedra.

8 - BOUNDED NORM REACHABILITY

In this section we outline the discrete time version of the theory of bounded norm reachability. Here we are interested to the possibility of steering the state from one point to another in a finite interval of time. Because we consider stationary systems such interval can be taken, once and for all, of the form [0, T]. Moreover, to simplify notations, we denote $\mathbf{u}(0,T)$ by $\mathbf{u}(T)$.

With such stipulation the problem in point is that of verifying the existence of a control $\mathbf{u}(T)$, such that $x(0) = x$ and $x(T) = z$, under the constraints $\|\mathbf{u}(T)\| \leq \rho$, for given x, z and ρ. Here, of course, the norm refers to the function space $\mathbb{R}^{pT}$ (incidentally, we use the same symbol for all norms leaving to the context to specify the space to which it refers). Thus we deal with a theory where the constraints are not pointwise in time.

The reachable set at T starting at time zero from the origin will be denoted by D.

The two cases of reachability from the origin and from x are quite similar since the set of states reachable from x is nothing but $A^T x + D$.

It will be useful in the sequel to bear in mind the following simple observation. The set D is the image under $C(T)$ of the closed sphere of radius ρ in the input function space, which is a convex and compact set. Thus such is D, and also the set $A^T x + D$.

Of course one may extend the theory to other norms. In the present finite dimensional context all norms are topologically equivalent, but it is interesting to note that norms with polytopic spheres (see also [31]) can be handled along the lines of polyhedral theory, rather than, for example, adopting an Lp approach.

At this point we can pass on to state a first important result of the theory. The machinery for proving the following theorem consists of inner product properties and a separation theorem.

THEOREM 10.

The state z is reachable at T, starting from the state x at 0, under $\|\mathbf{u}(T)\| \leq \rho$ if and only if

$$\forall\, y \in \mathbb{R}^n \quad (y,\, z\text{-}A^T x) \leq \rho \,\|\, C^*(T)\, y \,\| \qquad (8.1)$$

PROOF - Necessity: Suppose z is reachable at T from x at 0 for some given $\mathbf{u}(T)$ satisfying the constraints. Then

$$z = A^T x + C(T)\, \mathbf{u}(T)$$

Thus for an arbitrary $y \in \mathbb{R}^n$:

$$(y,\, z) = (y,\, A^T x) + (y,\, C(T)\, \mathbf{u}(T))$$

or

$$(y,\, z\text{-}A^T x) = (y,\, C(T)\, \mathbf{u}(T)) = (C^*(T)y,\, \mathbf{u}(T))$$

Applying Schwarz's inequality:

$$|\,(y,\, z\text{-}A^T x)\,| \leq \|C^*(T)y\| \;\; \|\mathbf{u}(T))\| \leq \|C^*(T)y\|\rho$$

whence, being the r.h.s. positive, the desired conclusion follows.

Sufficiency: suppose that, even though the condition holds, the point z in not reachable from x under the given constraint. In other words we have assumed that $z \notin A^t x + D$. Denote this latter set by C for the present purposes and for the sake of simplicity. In view of the fact that C is convex and compact and a well known separation result (e.g. Corollary 14.4 in [32]), there is a continuous linear functional strongly separating {z} and C. That is, if such a functional is represented by the vector v:

$$\sup \{(v, y): y \in C\} < (v, z)$$

Thus, there exists an α such that:

$$(v, y) < \alpha < (v, z) \qquad \forall\, y = A^t x + d \text{ with } d \in D$$

Hence:

$$(v, A^t x + d) < \alpha < (v, z) \qquad \forall\, d \in D$$

or:

$$(v, d) < \alpha - (v, A^t x) < (v, z - A^t x) \qquad \forall\, d \in D$$

This can be rewritten, letting $\beta = \alpha - (v, A^t x)$:

$$(v, d) < \beta < (v, z - A^t x) \qquad \forall\, d \in D$$

For the moment consider the first inequality. Bearing in mind that d has the form $C(T)\,u(T)$:

$$(C^*(T)v, u(T)) < \beta \quad \text{for } \|u(T)\| \leq \rho$$

And hence

$$\sup\{(C^*(T)v, u(T)): \|u(T)\| \leq \rho\} \leq \beta$$

On the other hand the norm of a functional, as is readily verified, is equal to the norm of the vector that represent the functional, so that this inequality yields:

$$\|C^*(T)v\|\, \rho \leq \beta$$

And using the second inequality too:

$$\|C^*(T)v\| \ \rho < (v,\ z\text{-}A^t x)$$

which contradicts the initial assumption thereby completing the proof. ∎

As a first remark note that, for the case of reachability from the origin, the condition of the theorem becomes:

$$\forall\ y \in \mathbb{R}^n \quad (y,\ z) \leq \rho\ \|\ C^*(T)\ y\| \tag{8.2}$$

Next it is useful to put this result in a different form. Let Z be either the matrix $z\ z^*$ or the matrix $(z\text{-}A^t x)\ (z\text{-}A^t x)^*$ for the case of reachability from the origin or from x respectively. For simplicity let's carry out the arithmetics for the first case, with the understanding that the substitution of z by $(z\text{-}A^t x)$ is all it is required to cover the second case too. The condition:

$$\forall\ y \in \mathbb{R}^n \quad (y,\ z) \leq \rho\ \|\ C^*(T)\ y\|$$

can be rewritten as:

$$\forall\ y \in \mathbb{R}^n \quad (y,\ z)^2 \leq \rho^2\ \|\ C^*(T)\ y\|^2 \tag{8.3}$$

$$\forall\ y \in \mathbb{R}^n \quad y^*\ z\ z^*\ y \leq \rho^2\ (C^*(T)\ y,\ C^*(T)\ y)$$

$$\forall\ y \in \mathbb{R}^n \quad (y,\ Z\ y) \leq \rho^2\ (C(T)\ C^*(T)\ y,\ y)$$

That is to say that the matrix $C(T)C^*(T) - 1/\rho^2\ Z$ is positive semidefinite. We may state this formally in the following:

COROLLARY 1.

The state z is reachable at T, starting from the state x (the origin) at 0, under $\|u(T)\| \leq \rho$, if and only if the matrix

$$C(T)\ C^*(T) - 1/\rho^2\ Z$$

is positive semidefinite, where Z denotes the matrix $(z\text{-}A^t x)\ (z\text{-}A^t x)^*$ (the matrix $z\ z^*$).

The matrix $C(T)C^*(T)$ is often called the Gramian matrix associated with the system. Notice that by its very definition, if $T_1 > T_2$, then $C(T_1)C^*(T_2) \geq C(T_2)C^*(T_2)$.

The result above readily implies that the system is reachable if and only if there exists a T such that the matrix $C(T)\ C^*(T)$ is positive definite. The key of the argument lies, for necessity, in the fact that, because z is arbitrary, taking z = y, we arrive to the conclusion that it must be $(y, C(T)C^*(T)y) \geq 1/\rho^2 \|y\|^2$ (from 8.3) for sufficiently large ρ and T. For sufficiency the key lies in the fact that if we assume that $C(T)C^*(T)$ is positive definite and, at the same time that $R(C(T)) \neq \mathbb{R}^n$ (i.e., the system is not reachable) a contradiction would arise. In fact in this case it would be $N(C^*(T)) \neq \{0\}$ so that it would exist an $y \neq 0$ such that $C^*(T)y = 0$. But there $(y,C(T)C^*(T)y)=0$ also.

8.1 - A bridge to optimization

As mentioned earlier, often optimization results are practically built into constrained reachability theory. This should not come to a surprise, since optimization and feasibility are akin to each other and reachability is, after all, a sort of feasibility. It is useful to digress a short while, just enough to taste the flavor of optimization.

From Corollary 1 and the fact that, as already noted, the Gramian matrix "increases" with time, it is clear that for any given bound to the norm, if a state is reachable at some T, then there exists a minimum reachability time. How about the minimum norm? Existence is settled by the following:

THEOREM 11.

Assume that the state z is reachable at T from the origin (or from x) by means of controls with norm less than or equal to ρ, and let U_ρ be the set of controls that effect the transfer of the state. Then there exists in U_ρ a control $\hat{u}$, that has minimum norm $\hat{\rho}$.

PROOF. The proof for the case of reachability from x is a trivial variation of the proof for the case of reachability from the

origin. Thus it is convenient to make reference to this latter case only. Let ρ be the infimum of the set of norms of members of U_ρ. Then there exists a sequence of positive numbers $\{\rho_n\}$, with $\rho_n \geq \hat{\rho}$ for any n, that converges to $\hat{\rho}$, and such that for each n there exists a control that effects the transfer and has norm less than or equal to ρ_n. Thus, in view of Theorem 9 we can write for each n

$$(y, z) \leq \rho_n \| C^*(T)\, y\| \qquad \forall\, y \in \mathbb{R}^n \tag{8.1.1}$$

Therefore, passing to the limit:

$$(y, z) \leq \hat{\rho} \| C^*(T)\, y\| \qquad \forall\, y \in \mathbb{R}^n \tag{8.1.2}$$

It follows, by the same theorem, that there exists a control $\hat{u}$ that makes the same transfer and such that $\|u\| \leq \rho$. Finally, since $u \in U_\rho$ and ρ is the infimum of norms of elements of U_ρ, it must be $\|\hat{u}\| = \rho$, so that the proof is complete. ∎

8.2 - Computation of optimal control

In this subsection we show that the machinery just developed allows a straightforward calculation of the optimal control.

THEOREM 12.

The minimum norm $\hat{\rho}$ of a control effecting the transfer in the interval [0, T] is given by:

$$\hat{\rho} = \sup \{(y, w): \| C^*(T)\, y\| = 1\} \tag{8.2.1}$$

where, as usual, $w = z$ for the case of reachability from the origin and $w = z - A^T x$ for the case of reachability from x.

PROOF - Being the proofs of the two statements similar it is convenient to give the proof for the case of reachability from the origin. The generalization will be immediate.

To this purpose we start from inequality (8.1.2):

$$(y, z) \leq \hat{\rho} \, \| C^*(T)\, y\| \qquad \forall\, y \in \mathbb{R}^n \tag{8.2.2}$$

This latter implies:

$$\hat{\rho} \geq \sup \{(y, z): \| C^*(T)\, y\| = 1\} = \rho' \tag{8.2.3}$$

Suppose that $\hat{\rho} > \rho'$ so that there exists σ such that $\hat{\rho} > \sigma > \rho'$. Because $\sigma < \rho$, z is no more reachable relative to the input bound norm σ. Therefore, in view of Theorem 9, there exists a vector v such that:

$$(v, z) > \sigma \, \| C^*(T)\, v\| \tag{8.2.4}$$

This inequality implies that $(v,z) \neq 0$, which in turn implies, in view of inequality we started with, that $\| C^*(T)\, v\| \neq 0$. Dividing both sides of the last inequality by $\| C^*(T)\, v\|$, we obtain for the vector $q = v / \| C^*(T)\, v\|$:

$$(q, z) > \sigma > \rho' \tag{8.2.5}$$

with $\| C^*(T)\, q\| = 1$. This contradicts the definition of ρ'. Thus $\hat{\rho} = \rho'$ and the proof is completed. ∎

Notice that if we consider reachability from x and $z = A^T x$, then $\rho = 0$. This is obvious because in this case the target state can be reached by free motion. It is obvious too that this is the only possibility for ρ to be zero, because the only zero norm input is the identically zero input. Thus it is contradictory that $z \neq A^T x$ and $\rho = 0$, because we would be left with the only free motion. In particular if we consider reachability from the origin and $z \neq 0$ then ρ cannot be zero too. In the sequel we exclude the trivial case $\rho = 0$, continuing to refer, without explicit mention, all the proofs to the case of reachability from the origin.

THEOREM 13.

The supremum of the preceding theorem is attained.

PROOF - By the expression of $\hat{\rho}$ given by the preceding theorem there exists a sequence $\{\rho_n\}$ converging to ρ from below such that, for

each n, $\rho_n > \sigma > 0$ and there exists an y_n with $\| C^*(T)\, y_n \| = 1$ with

$$\rho_n = \| C^*(T)\, y_n \|\, \rho_n < (y_n, z)$$

Notice that we can take each y_n in the subspace $R(C(T))$. In fact $\mathbb{R}^n = R(C(T)) + N(C^*(T))$, the two spaces being the orthogonal complement of each other. Moreover $z \in R(C(T))$, because it is a reachable state. Now if $y_n = y_{1n} + y_{2n}$ with $y_{1n} \in R(C(T))$ and $y_{2n} \in N(C^*(T))$, we can clearly write, substituting to y_n its decomposition:

$$\sigma < \rho_n = \| C^*(T)\, y_{1n} \|\, \rho_n < (y_{1n}, z)$$

and again $\| C^*(T)\, y_{1n} \| = 1$. The big difference is that now, since $C^*(T)$ is a linear isomorphism of $R(C(T))$ onto $R(C^*(T))$, and the sequence $\{y_{1n}\}$ is mapped into the unit ball of $R(C^*(T))$, it follows that the sequence is in some sphere and because this latter is compact it admits a convergent subsequence. Passing to the limit for this latter and eliminating the by now irrelevant subscript 1:

$$\sigma < \hat{\rho} = \| C^*(T)\, y \|\, \hat{\rho} \leq (y, z)$$

with $\| C^*(T)\, y \| = 1$, in view of an obvious continuity argument. But because z is reachable relative to ρ, our theorem on constrained reachability insures also that:

$$(y, z) \leq \hat{\rho} \| C^*(T)\, y \| = \hat{\rho}$$

which completes the proof. ∎

In view of this theorem there exists a vector $\hat{y}$ with $\| C^*(T)\, \hat{y} \| = 1$, such that $(\hat{y}, z) = |(\hat{y}, z)| = \hat{\rho} = \hat{\rho} \| C^*(T)\, \hat{y} \|$.

THEOREM 14.

The optimal control is given by:

$$\hat{u} = \hat{\rho}^2 \, C^*(T) \, v \tag{8.2.6}$$

and v can be found solving the equations

$$\begin{cases} \hat{\rho}^2 \, C(T) \quad C^*(T) \, v = z \\ | (v, \, z) | \quad = 1 \end{cases} \tag{8.2.7}$$

where $v = (1/\hat{\rho}) \, \hat{y}$.

PROOF - From the equality $| (\hat{y}, \, z) | = \hat{\rho} \, \| \, C^* \, (T) \, \hat{y} \|$ and since $z = C(T) \, \hat{u}$:

$$\hat{\rho} \, \| C^* \, (T) \, \hat{y} \| = | (\hat{y}, \, C(T) \, \hat{u}) | = | (C^* \, (T) \hat{y}, \, \hat{u}) |$$

From this we infer that the Schwartz inequality holds as equality and therefore $C^*(T) \, \hat{y}$ and $\hat{u}$ must be proportional. That is:

$$\hat{u} = K \, C^*(T) \, \hat{y}$$

where, of course:

$$K = \| \, \hat{u} \, \| / \| \, C^*(T) \, \hat{y} \, \| = \hat{\rho} \, / \, \| \, C^*(T) \, \hat{y} \, \| = \hat{\rho}$$

hence

$$\hat{u} = \hat{\rho} \, C^*(T) \, \hat{y} = \hat{\rho}^2 \, C^*(T) \, v$$

and because $z = C(T) \, \hat{u} = \hat{\rho}^2 \, C(T) \, C^*(T) \, v$ and $| (\hat{y}, z) | = \hat{\rho}$, so that $| (v, \, z) | = 1$, the proof is completed. ∎

9 - APPROXIMATED AND DISTURBED CONSTRAINED REACHABILITY

In this section we examine the problem of approximate reachability, which is particularly meaningful in view of the

presence of inputs constraints, and the problem of disturbed reachability. This is connected to the former since we wish to investigate whether, in presence of the constraints and of noise, (again characterized by the assumption of bounded norm) we can reach, if not the target point, at least some point of a ball around the target point.

Let's start with the first problem. This time our crucial result is based on the following immediate consequence of the usual separation theorem.

LEMMA 2.

If $\mathscr{C}$ and $\mathscr{D}$ are two convex and compact subsets of $\mathbb{R}^n$, then a necessary and sufficient condition for them to have a non empty intersection, is that:

$$\inf \{(x,d): d \in \mathscr{D}\} \leq \sup \{(x, c): c \in \mathscr{C}) \quad \forall x \in \mathbb{R}^n \qquad (9.1)$$

PROOF - If the two sets are not disjoint then, given $x \in \mathbb{R}^n$ for a point y in their intersection the two real numbers sets in our condition have a point in common, so that the inequality becomes obvious. On the other hand suppose that the inequality is true but the two sets do not intersect. Then in view of the already cited Corollary 14.4 in [32] there exists an $x \in \mathbb{R}^n$ such that

$$\inf \{(x, d): d \in \mathscr{D}\} > \sup \{(x, c): c \in \mathscr{C})$$

which is a contradiction. ∎

We can now state the following result of approximate reachability:

THEOREM 15.

There is at least one point in the closed sphere of radius ε about z, that is reachable at time T starting from the state x at time 0, under the constraint that the input functions have the norm bounded by ρ, if and only if:

$$(y, z\text{-}A^T x) - \rho \| C^*(T)\, y\| - \varepsilon \| y \| \leq 0 \qquad \forall y \in \mathbb{R}^n \qquad (9.2)$$

PROOF - As usual, we can give the proof for the only case of reachability from the origin, the general case requiring only trivial variations of the argument.

We already observed that the set D of all states reachable at time T, starting from the origin at time 0, is a convex and compact set. If this set intersects the closed sphere S of radius ε around z, then the present approximate reachability property prevails otherwise it doesn't. Thus, applying the Lemma, the first case occurs if and only if

$$\inf \{(y, d): d \in D\} \leq \sup \{(y, s): s \in S) \quad \forall\, y \in \mathbb{R}^n$$

But

$$\inf \{(y, d): d \in D\} = \inf \{(y, C(T)\, u(T)): \|u(T)\| \leq \rho\} =$$
$$- \sup \{(y, -C(T)\, u(T)): \|u(T)\| \leq \rho\} =$$
$$= - \sup \{(y, C(T)\, u(T)): \|u(T)\| \leq \rho\}$$

since the constraining set is symmetric.

Moreover

$$\sup \{(y, C(T)\, u(T)): \|u(T)\| \leq \rho\} =$$
$$\sup \{(C^*(T)\, y, u(T)): \|u(T)\| \leq \rho\} = \rho \|C^*(T)\, y\|$$

as is readily seen from Schwartz inequality and the fact that $u(T)$ can be taken proportional to $C^*(T)\, y$, and with norm ρ. By a similar argument, as to the second term of the inequality, we obtain:

$$\sup \{(y, s): s \in S) = (y, z) + \varepsilon \| y \|$$

Putting again together the terms

$$- \rho \|C^*(T)\, y\| \leq \varepsilon \| y \| + (y, z)$$

This latter, changing y into -y yields:

$$(y, z) - \rho \|C^*(T)\, y\| - \varepsilon \| y \| \leq 0$$

as we wanted to prove. ∎

The final bounded norm reachability problem we deal with is that in which not only the input has bounded norm, but, also, the system is affected by noise (that is, the matrix D is different from zero), and such noise disturbs our action of steering the state toward the target point. We may look at this case as a problem of robust reachability, since our purpose is that of establishing conditions under which there exist a control, that allows us to reach some point in the neighborhood of the target point, no matter what is the noise function, provided this latter has bounded norm too.

Let's be more precise. Let us call again ρ the bound for the input norm and δ the bound for the noise norm. The problem is that of determining whether there exists a control $\mathbf{u}(T)$, with $\|\mathbf{u}(T)\| \leq \rho$, such that the state is transferred from x at time 0 to a point of the closed ball S of radius ε around the target state z, no matter what is the noise function $\mathbf{d}(T)$, with $\|\mathbf{d}(T)\| \leq \delta$.

It is possible to reformulate the problem in such a way that it can be solved invoking the usual separation theorem. If we call N the set of all states in which the system is steered by all possible noise functions (satisfying the constraint) at time T, starting from the origin at time 0, then, for any given control $\mathbf{u}(T)$, the set of all states that will be reached is:

$$A^T x + C(T)\,\mathbf{u}(T) + N \tag{9.3}$$

Clearly we must verify if there exists an $\mathbf{u}(T)$, satisfying the constraints, such that this set be contained in S.

Next let F be the set

$$F = \{w: w + N \subset S\} \tag{9.4}$$

Notice that $F \subset S$. At this point we can equivalently express our condition saying that it must be $R \cap F \neq \phi$, where R is the set of all points reachable at T from x under zero noise and with inputs satisfying their constraint.

It is no more than a straightforward verification to ascertain that F is convex and closed. Therefore, since $F \subset S$, which is

compact, it follows that F is compact too. At this point the previously invoked separation theorem can be used again. Denoting by H the set F - z and reasoning along the same lines of the proof of the preceding theorem we can state the following:

THEOREM 16.

A point in the closed ε ball around z can be reached at time T starting from x at time 0 by means of an input $\mathbf{u}(T)$ with $\|\mathbf{u}(T)\| \leq \rho$, for any noise function $\mathbf{d}(T)$ with $\|\mathbf{d}(T)\| \leq \delta$ if and only if:

$$(y,\ z\text{-}A^t x) - \rho \parallel C^*(T)\ y\parallel + \min\ \{(v,\ y):\ v \in H\} \leq 0 \quad \forall y \in \mathbb{R}^n \tag{9.5}$$

PROOF: We again make reference to the case $x=0$, substituting R with D. In view of the non-emptiness condition, it must be:

$$\inf\ \{(y,\ d):\ d \in D\} \leq \sup\ \{(y,\ f):\ f \in F) \quad \forall\ y \in \mathbb{R}^n$$

We already computed the first term in the preceding proof. Thus:

$$-\ \rho\ \|C^*(T)\ y\| \leq \sup\ \{(y,\ f):\ f \in F)$$

Moreover

$$\begin{aligned}&\sup\ \{(y,\ f):\ f \in F) = \sup\ \{(y,w+z):\ w \in H\} = \\ &= (y,z) + \sup\ \{(y,w):\ w \in H\} = \\ &= (y,z) + \max\ \{(y,w):\ w \in H\}\end{aligned}$$

because H is compact. Finally:

$$-\ \rho\ \|C^*(T)\ y\| \leq (y,z) + \max\ \{(y,w):\ w \in H\}$$

and changing y into -y:

$$(y,\ z) - \rho \parallel C^*(T)\ y\| + \min\ \{(v,\ y):\ v \in H\} \leq 0 \quad \forall y \in \mathbb{R}^n$$

which is what we wanted to prove.

REFERENCES

[1] - A.A. Liapunov, "On completely additive vectorial functions", *Izvestia Akademii Nauk SSSR, Seria Matem.*, pp. 465-4784, (1940).

[2] - R. Conti, "Processi di controllo lineari in $\mathbb{R}^n$", *Quaderni UMI* **30**, Pitagora, Bologna, (1985) (In Italian)

[3] - E. De Santis, "On Reachability of constrained discrete time linear linear systems", *Research Report* **41**, Department of Electrical Engineering, University of L'Aquila, (1990).

[4] - P.d'Alessandro, E. De Santis, "General closed loop optimal solutions for linear dynamic systems with linear constraints and functional", *Research Report* **50**, Department of Electrical Engineering, University of L'Aquila, (1993) (revised version of *Research Report* **40**, (1990)).

[5] - P.d'Alessandro, E. De Santis, "Positiveness of Dynamic Systems with Non Positive Coefficient Matrices", *IEEE Transactions on Automatic Control*, (to appear), (1993).

[6] - A. Marzollo, "Controllability and Optimization", Lectures held at the Department for Automation and Information, University of Trieste, *Courses and Lectures* **17**, International Center for Mechanical Sciences, Udine (Italy), (1969).

[7] - H.A. Antosiewiez, "Linear control systems", *Arch. Rat. Mech. Anal.*, **12**, pp. 313-324 (1963).

[8] - D. Blackwell, "The range of certain vector integrals", *Proceedings of the Amer. Math. Society*, **2**, pp. 390-395, (1951).

[9] - J.P. LaSalle, "The time optimal problem", *in* "Theory of nonlinear oscillations", vol. 5, pp. 1-24, Princeton Press, (1960).

[10] - E.B. Lee, L. Markus, "Foundations of Optimal Control Theory", J. Wiley & Sons, (1967).

[11] - R.M. Bianchini, "Local Controllability, Rest States and Cyclic Points", *SIAM J. Control and Optimization*, vol. 21, No. **5**, (1983).

[12] - M.E. Evans, "Bounded control and discrete time controllability",*International Journal of Systems Science*, **17**, pp. 943-951, (1986).

[13] - J. E. Gayek and M.E. Fisher, "Approximating Reachable Sets for n-dimensional Linear Discrete Systems", *IMA Journal of Mathematical Control & Information* **4**, pp. 149-159, (1987).

[14] - M.E. Fisher and J.E. Gayek, "Estimating Reachable Sets for Two-Dimensional Linear Discrete Systems", *Journal of Optimization Theory and Applications*, Vol. 56, No. **1**, pp. 67-88, (1988).

[15] - M.E. Fisher and W.J. Grantam, "Estimating the effect of continual disturbances on discrete time population models", *J. Math. Biol.*, **22**, pp. 199-207, (1985).

[16] - P.O. Gutman and M. Cwikel, "Admissible Sets and Feedback Control for Discrete-Time Linear Dynamical Systems with Bounded Controls and States", *IEEE Transactions on Automatic Control AC*-**31**, No. **4**, pp. 373-376, (1986).

[17] - M. Cwikel and P.O. Gutman, "Convergence of an Algorithm to Find Maximal State Constraint Sets for Discrete-Time Linear Dynamical Systems with Bounded Controls and States", *IEEE Transactions on Automatic Control AC*-**31**, No. **5**, pp. 457-459, (1986).

[18] - P.O. Gutman and M. Cwikel, "An Algorithm to Find Maximal State Constraint Sets for Discrete Time linear Dynamical Systems with Bounded Controls and States", *IEEE Transactions on Automatic Control AC*-**32**, No. **3**, pp. 251-254, (1987).

[19] - V. G. Rumchev, "Constructing the reachability sets for positive linear discrete-time systems. The case of polyhedra". *Systems Science*, Vol. 15, No. **3**, (1989).

[20] - S.S. Keerthi and E.G. Gilbert, "Computation of Minimum Time Feedback Control Laws for Discrete-Time Systems with State-Control Constraints", *IEEE Transactions on Automatic Control AC*-**32**, No. **5**, pp. 432-435, (1987).

[21] - P. d'Alessandro, M. Dalla Mora and E. De Santis, "On consistency of linear linearly constrained discrete time systems", *International Journal of the Franklin institute*, vol. 319, no. 4, pp. 423-430, (1985).

[22] - P. d'Alessandro, M. Dalla Mora and E. De Santis, "On discrete time linear systems over cones" - *Systems & Control Letters*, **6**, pp.271-275, (1985).

[23] - P. d'Alessandro and E. De Santis, "Reachability in input constrained discrete time linear systems", *Automatica*, vol. 28 no. 1, pp. 227-230, (1992).

[24] - P. d'Alessandro and M. Dalla Mora, "Systems, memory, causality, evolution and recursive equations", *Computers and Mathematics with Applications*, Vol. 10, No. 1, pp. 61-69, (1984).

[25] - R.T. Rockafellar, "Convex Analysis", Princeton Mathematical Series, No. 28, (1970).

[26] - J. Stoer, C. Witzgall, "Convexity and Optimization in finite dimensions I", Springer-Verlag Berlin-Heidelberg-New York (1970).

[27] -D.G. Luenberger, "Introduction to dynamic systems", J. Wiley & Sons, New York, (1979).

[28] - A Berman, M. Neumann and R. J. Stern, "Nonnegative matrices in dynamic systems" J. Wiley & Sons, New York, (1989).

[29] - P. d'Alessandro, M. Dalla Mora and E. De Santis, "Techniques of linear programming based on the theory of convex cones", Optimization **20**, pp. 761-777, (1989).

[30] - P. d'Alessandro, "The conical approach to linear programming", *Research Report* **47**, Department of Electrical Engineering, University of L'Aquila, (1991).

[31] - P. d'Alessandro, M. Dalla Mora, "Fast projection method for a special class of polytopes with applications", *RAIRO O.R.*, vol. 22, no. 4, pp. 347-361, (1988).

[32] - J.L. Kelley, I. Namioka et al., "Linear topological spaces", Springer - Verlag, New York, (1963).

STABILIZATION, REGULATION, AND OPTIMIZATION OF MULTIRATE SAMPLED-DATA SYSTEMS

Patrizio Colaneri
Riccardo Scattolini
Nicola Schiavoni

Dipartimento di Elettronica e Informazione
Politecnico di Milano
Milano, Italy

I. INTRODUCTION

In classical digital control systems, it is usually assumed that both the plant inputs-updating and the plant outputs-measurement are performed at a unique constant rate and in a synchronous fashion. However, this hypotesis is sometimes not realistic, for economical and/or technological reasons, and, furthermore, relaxing it often allows the designer to obtain improved control performances. Hence, one is lead to consider the so-called multirate sampled-data control systems, which are characterized by the fact that each input is updated at an its own rate and each output is measured at an its own rate. The analysis and the design of such systems has recently received a great deal of attention. For an overview of the area see, e.g., [1]-[3].

There are two primary reasons of interest in multirate

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

digital control.

A first strong motivation behind their use is due to the possible presence of technological constraints which enforce the use of control schemes where sensor measurements and control calculations have to be performed at different sampling rates, see, e.g., [4]-[7]. This typically occurs in one of the following cases:

(i) Some sensors require a significant time before they supply the measurements of the plant output variables to the regulator. For example, such a situation occurs in controlling chemical plants where expensive chromatographs are used to measure composition products. These measurements are then infrequent and delayed with respect to those of other variables measured by sensors not suffering of such a limitation.

(ii) A small number of sensors is used to measure a large number of output variables at different times, or the sensors allow one to measure all the plant outputs at the same rate and time, but hardware constraints prevent one from transmitting data simultaneously from all the sensors to the control processing unit.

(iii) The plant outputs are all measured at the same rate and time, but this rate is less than that of the plant inputs updating allowed by the control apparatuses.

(iv) Some actuators are manipulated less frequently than others in order to reduce the effort of these apparatuses.

As a second reason, it has been shown that the use of multirate and periodically time varying controllers can significantly improve the closed-loop performance of a sampled-data system in terms of model matching, sensitivity reduction, disturbance rejection, pole and zero assignment with state feedback, see, e.g., [1], [8], [9]. However, these promising results usually refer to the sampled version of the system, while particular care has also to be paid to the intersample behavior which can be significantly deteriorated by the multirate input updating, see [10].

A deep difference exists in the two former classes of applications of multirate control: when a multirate approach is used to improve control performances, the frequencies and phases of inputs-updating and outputs-measurement are free design parameters to be determined by the control strategy in order to optimize the

required performances. On the contrary, when a multirate solution is enforced by technological considerations, the same parameters are problem data and must be faced by the adopted control synthesis technique. In this paper, attention will be focused on this last situation. Hence the problem addressed will be to design a multirate digital regulator once the inputs-updating and outputs-sampling mechanisms are fixed.

Research in multirate control can be traced back to the late fifties [11]-[13]; however, it has received more and more attention only in the past decade. For an overview of the most significant results of the area the reader is referred to [3], [14]. In [14] attention is focused on the analysis of a control structure where a different sampling rate is associated with any pair of input-output variables, then impulse modulation models are developed and criteria to assess closed-loop stability are presented. Several synthesis algorithms have recently been proposed in a linear time-invariant setting. Among them, the pole-placement approach has been considered in [15]-[21]. The Linear Quadratic Gaussian (LQG) technique has been applied in [4], [22]-[26], while some synthesis algorithms based on cost-function minimization have been presented for controllers with a prescribed structure [5], [27]-[29]. Some predictive and self-tuning multirate control algorithms have been proposed in [6], [7], [30], [31]. The output regulation problem, that is the problem of zeroing the steady-state error to the maximum possible extent in presence of exogenous signals of prescribed dynamics has been treated in [32]-[34]. Finally, the case of completely asynchronous sampling has been treated in [35].

The aim of this paper is to review the main aspects concerning the application of popular synthesis techniques, namely the pole-placement approach and the LQG method, to the multirate control problem. Specifically, the assumption is made that the plant under control is a discrete-time linear time-invariant system. It is also assumed that each output has its own frequency and phase of measurement and each input has its own frequency and phase of updating.

The paper is organized as follows. In Section 2, the discrete-time linear time invariant model of the plant under control is introduced and the sampling and updating

mechanisms are given a precise mathematical formulation. Further, it is also shown how multirate systems can be casted into the wider class of periodic systems. Then, in Section 3 some preliminary results on the structural properties (stabilizability, detectability and zeros) of the multirate system are given in terms of the original plant under control. Section 4 deals with the pole-placement and LQG methods when the system state is assumed to be available for control, while, since our main goal is to design output feedback controllers, in Section 5 the problem of state reconstruction is considered. In particular, two state observers are presented: in the first one the pole-assignment technique is again applied, while the second is derived by resorting to the Kalman filtering approach. In Section 6 the previous results on state feedback and state observers are joint together with the aim of deriving stabilizing feedback control laws. Finally, the classical output regulation problem is faced in Section 7 where, under some particular assumptions on the inputs updating, a suitable regulation structure is presented which guarantees the asymptotic tracking of given reference signals in spite of the presence of persistent disturbances and plant uncertainties.

II. THE PLANT, THE INPUTS-HOLDING AND THE OUTPUTS-SAMPLING MECHANISMS

Let the system under control be described by the following discrete-time linear time-invariant stochastic model

$$\mathcal{P}: \begin{cases} x(t+1) = Ax(t) + Bu(t) + M_1 w_1(t) & (1.a) \\ y(t) = Cx(t) + M_2 w_2(t) & (1.b) \end{cases}$$

where $A \in R^{n,n}$, $B \in R^{n,m}$, $M_1 \in R^{n,q}$, $C \in R^{p,n}$, $M_2 \in R^{p,p}$ and w_1 and w_2 are uncorrelated zero-mean gaussian white noises with identity covariance matrices, i.e.,

$w_1 \sim WGN(0,I)$, $w_2 \sim WGN(0,I)$

The standard hypothesis is made that the noise acting on the output variable y has a nonsingular covariance matrix, that is $\det(M_2)\neq 0$. Furthermore, the initial state $x(t_0)$ is supposed to be a gaussian random variable, uncorrelated with w_1 and w_2.

Now assume that the i-th component $u_i(t)$, i=1,2,...,m, of the input vector u(t) can be modified every $\tilde{T}_i$ time-instants, $\tilde{T}_i$ being a finite positive integer. Then, $u_i(t)$ can be viewed as the output of the following discrete-time periodic system, henceforth called the input-holding mechanism,

$$v_i(t+1) = s_i(t)\, v_i(t) + (1-s_i(t))\, r_i(t) \qquad (2.a)$$

$$u_i(t) = s_i(t)\, v_i(t) + (1-s_i(t))\, r_i(t) \qquad (2.b)$$

where $v_i(t)\in R^1$ is the state-variable and $r_i(t)\in R^1$ is the new input variable. As for $s_i(.)$, it is a $\tilde{T}_i$-periodic function defined as

$$s_i(t) := \begin{cases} 0 & t=k\tilde{T}_i+\tilde{\tau}_i \\ 1 & t\neq k\tilde{T}_i+\tilde{\tau}_i \end{cases}$$

where k is an integer and the integers $\tilde{\tau}_i$, $0\leq\tilde{\tau}_i<\tilde{T}_i$, describe the skew inputs-holding mechanism.

Now, letting

$$v(t) := [v_1(t)\ v_2(t)\ \dots\ v_m(t)]'$$

$$r(t) := [r_1(t)\ r_2(t)\ \dots\ r_m(t)]'$$

$$S(t) := \mathrm{diag}\{s_1(t), s_2(t), \dots\ s_m(t)\}$$

system (2) can be given the compact form

$$\mathcal{H}: \begin{cases} v(t+1) = S(t)\, v(t) + (I - S(t))\, r(t) & (3.a) \\ u(t) = S(t)\, v(t) + (I - S(t))\, r(t) & (3.b) \end{cases}$$

In view of the periodicity of the s_i's matrix $S(t)$ and system (3) are periodic of period $\tilde{T}$, where

$$\tilde{T} := \underset{i=1,2,\ldots,m}{\text{l.c.m.}} \{\tilde{T}_i\} \qquad (4)$$

As for the system outputs-sampling mechanism, let the i-th component $y_i(t)$, i=1,2,...,p, of the output vector $y(t)$ be measured every $\bar{T}_i$ time-instants, $\bar{T}_i$ being a finite positive integer. This can be given a mathematical formulation by defining the measured output variable $\zeta_i(t)$, i=1,2,...,p, as follows

$$\zeta_i(t) := \begin{cases} y_i(t) & t = k\bar{T}_i + \bar{\tau}_i \\ 0 & t \neq k\bar{T}_i + \bar{\tau}_i \end{cases}$$

where again k is an integer and the integers $\bar{\tau}_i$, $0 \leq \bar{\tau}_i < \bar{T}_i$, describe the skew outputs-sampling mechanism. Then, at any time-instant t, the output measured vector is given by

$$\mathcal{N}: \{ \zeta(t) := N(t)\, y(t) \qquad (5)$$

where

$$N(t) := \text{diag}\{\nu_1(t), \nu_2(t), \ldots \nu_p(t)\}$$

$$\nu_i(t) := \begin{cases} 1 & t = k\bar{T}_i + \bar{\tau}_i \\ 0 & t \neq k\bar{T}_i + \bar{\tau}_i \end{cases}$$

Matrix $N(t)$ is $\bar{T}$-periodic, where

$$\bar{T} := \underset{i=1,2,\ldots,p}{\text{l.c.m.}} \{\bar{T}_i\} \tag{6}$$

By combining the model (1), the inputs-holding system (3), (4) and the outputs-sampling mechanism (5), (6), the following overall system is obtained as

$$\xi(t+1) = \Phi(t)\ \xi(t) + \Gamma(t)\ r(t) + \Psi\ w_1(t) \tag{7.a}$$

$$\zeta(t) = \Delta(t)\ \xi(t) + \Omega(t)\ w_2(t) \tag{7.b}$$

where

$$\xi(t) := \begin{bmatrix} x(t) \\ \\ v(t) \end{bmatrix} \tag{8}$$

$$\Phi(t) := \begin{bmatrix} A & BS(t) \\ \\ 0 & S(t) \end{bmatrix}, \quad \Gamma(t) := \begin{bmatrix} B \\ \\ I \end{bmatrix} (I - S(t)), \quad \Psi := \begin{bmatrix} M_1 \\ \\ 0 \end{bmatrix}$$

$$\Delta(t) := [N(t)C \quad 0], \quad \Omega(t) := N(t)M_2$$

and 0 represents zero matrices of appropriate sizes.

In view of the $\tilde{T}$-periodicity of matrix $S(.)$ and $\bar{T}$-periodicity of matrix $N(.)$, system (7) is T-periodic, where

$$T := \text{l.c.m.}\ \{\tilde{T}, \bar{T}\} \tag{9}$$

Notice that at the fixed time instant t some output variables may be zero, due to the sampling mechanism $\mathcal{N}$, and some input variables do not affect the system dynamics, due to the holding mechanism $\mathcal{H}$. In order to avoid this degeneracy, it is convenient to reorder the input and output variables in the following way:

$$r(t) = P_u(t)\ \vartheta(t) \tag{10.a}$$

$$\eta(t) = P_y(t)\ \zeta(t) \qquad (10.b)$$

where $P_u(.)$ ($P_y(.)$) is a $\tilde{T}$-periodic ($\bar{T}$-periodic) permutation matrix such that

$$(I - S(t))\ P_u(t) = [\nabla_u(t) \quad 0]$$

$$P_y(t)\ N(t) = [\nabla_y(t)' \quad 0]'$$

and $\nabla_u(t)$ ($\nabla_y(t)$) is a full-rank matrix, whose dimensions change with time. Accordingly, partition ϑ and η as follows:

$$\vartheta(t) = [\vartheta_1(t)' \quad \vartheta_2(t)']' \qquad (11.a)$$

$$\eta(t) = [\eta_1(t)' \quad \eta_2(t)']' \qquad (11.b)$$

where the dimensions of $\vartheta_1(t)$ and $\eta_1(t)$ equal the number of columns of $\nabla_u(t)$ and rows $\nabla_y(t)$, respectively, and η_2=0. With these positions the overall system is easily obtained from Eqs. (7), (10) and (11) as

$$\xi(t+1) = \Phi(t)\ \xi(t) + [B' \quad I]'\ \nabla_u(t)\ \vartheta_1(t) + \Psi\ w_1(t) \qquad (12.a)$$

$$\eta_1(t) = [\nabla_y(t)C \quad 0]\ \xi(t) + \nabla_y(t)\ M_2\ w_2(t) \qquad (12.b)$$

Of course, system (12) is T-periodic with input and output having time-varying dimensions.

III. STRUCTURAL PROPERTIES AND ZEROS

This section is devoted to analyze the main features of the T-periodic system (7), namely zeros, stabilizability and detectability, in terms of the original data (system (1)). For the analysis of discrete-time periodic systems it

is often useful to resort to the time-invariant reformulation (TIR) associated with a periodic system first introduced in [36]. For a tutorial paper on the structural properties of periodic systems, the reader is referred to [37]. The zeros of periodic systems are defined in [38], [39] as those of their corresponding time-invariant reformulations. For the reader's convenience, the definition of such a reformulation for a general (periodic) system is briefly recalled in the Appendix. Given any (periodic) system $\mathcal{S}$, the simbol $\hat{\mathcal{S}}$ will henceforth denote its time-invariant reformulation.

Stabilizability and detectability are discussed first.

Theorem 1 [26]
If

(i) The pair (A,B) is stabilizable;
(ii) Do not exist two distinct eigenvalues of A, λ_i and λ_j, $|\lambda_i|\geq 1$, $|\lambda_j|\geq 1$, such that $\lambda_i^{\tilde{T}}=\lambda_j^{\tilde{T}}$;
(iii) Do not exist eigenvalues λ of A, $\lambda\neq 1$, $|\lambda|=1$, such that $\lambda^{\tilde{T}}=1$;

Then the pair $(\Phi(\cdot),\Gamma(\cdot))$ is stabilizable. ■

Theorem 2
The pair $(\Phi(\cdot),\Psi(\cdot))$ is stabilizable if and only if the pair (A,M_1) is stabilizable. ■

As for detectability, recall that the T-periodic pair $(\Phi(.),\Delta(.))$ is detectable if and only if there exists a T-periodic matrix D(.) such that $\Phi(.)+D(.)\Delta(.)$ is asymptotically stable (see, e.g., [20]). According to the structure of matrices $\Phi(.)$ and $\Delta(.)$, let $D(t):=[D_1(t) \quad 0]$, where $D_1(t)\in R^{n,p}$ is periodic of period $\bar{T}$. Correspondingly

$$\Phi(t)+D(t)\Delta(t) = \begin{bmatrix} A+D_1(t)N(t)C & ? \\ 0 & S(t) \end{bmatrix}$$

where ? denotes a block which does not need to be

specified. Hence, since the characteristic multipliers [37] of $S(.)$ lie all at the origin, i.e. $S(.)$ is asymptotically stable, detectability in $\bar{T}$ of $(A,N(.)C)$ implies detectability in T of $(\Phi(.),\Delta(.))$. Thus, the following result is proven.

Theorem 3 [20]
If
(i) The pair (A,C) is detectable;
(ii) Do not exist two distinct eigenvalues of A, λ_i and λ_j, $|\lambda_i|\geq 1$, $|\lambda_j|\geq 1$, such that $\lambda_i^{\bar{T}}=\lambda_j^{\bar{T}}$;
Then the pair $(A,N(.)C)$ is detectable.■

The eigenvalues of the T-periodic system (7) are defined as the characteristic multipliers of $\Phi(\cdot)$, i.e., the eigenvalues of the so-called monodromy matrix $\hat{\Phi}:=\Phi(T+\tau-1)\Phi(T+\tau-2)\ldots\Phi(\tau)$, where τ is any (arbitrary) time instant. Recalling the definition of $\Phi(\cdot)$, it follows that the eigenvalues of (7) are those of A^T along with m eigenvalues at the origin, since, for all τ, $S(T+\tau-1)S(T+\tau-2)\ldots S(\tau) = 0$.

All the above results, defined for system (7), immediately extend to system (12), since the two systems differs only for slack input and output variables.

The situation dramatically changes when considering the zeros of systems (7) and (12). Indeed, restricting the attention to system (1) free of disturbances and square, i.e., m=p, it is apparent that, apart from the trivial case where $S(\cdot)=N(\cdot)=I$, any complex number is a zero of system (7). This kind of degeneracy has been thrown away in system (12), where the slack input and output variables do not appear anymore. However, the zeros of system (12) depend on $\mathcal{H}$ and $\mathcal{N}$ in such a way that they cannot be directly argued from $\mathcal{P}$ only. Neverthless, the following partial results can be stated. The first of them concerns the relationship betweeen the zeros of $\mathcal{P}$ and those of its TIR $\hat{\mathcal{P}}$.

Theorem 4 [33]
(i) If λ is a transmission zero of $\mathcal{P}$, then λ^T is a transmission zero of $\hat{\mathcal{P}}$;

(ii) If $\mu \neq 0$ is a transmission zero of $\hat{\mathcal{P}}$, without being an eigenvalue of $\mathcal{P}$, then there exists λ, such that $\lambda^T = \mu$, which is a transmission zero of $\mathcal{P}$. ■

In the special case where $S(\cdot)=0$, i.e., the inputs are updated at any time instant, system (3) simply reduces to $u(t)=r(t)$. This allows one to neglect Eq. (3.a) and consider system (12), as far as the zeros are concerned, as the cascade connection of $\mathcal{P}$ and $\mathcal{N}$.

Theorem 5 [33]
If $S(\cdot)=0$, then the set of transmission zeros of (12) belongs to the set of the transmission zeros of $\hat{\mathcal{P}}$. ■

IV. STATE-FEEDBACK CONTROL LAWS

In this section, we extend the main classical stabilization techniques, namely Pole-Placement (PP) and Linear-Quadratic (LQ) control, to the case of multirate systems. To this end, we consider the $\tilde{T}$-periodic system (7.a) and assume that its state $\xi(.)$ is available for control. Since our ultimate goal is to design output feedback controllers, we make reference in the sequel to period T, instead of $\tilde{T}$.

A. POLE-PLACEMENT

The TIR of system (7.a) with $w_1 \equiv 0$, is easily determined as

$$\hat{\xi}(k+1) = \hat{\Phi}\, \hat{\xi}(k) + \hat{\Gamma}\, \hat{r}(k) \qquad (13)$$

where

$$\hat{\xi}(k) := \xi(kT)\ ,\ \hat{r}(k) := [r(kT)'\ \ r(kT+1)'\ \ \dots\ r(kT+T-1)']'$$

and $\hat{\Phi}$, $\hat{\Gamma}$ are suitable matrices, for which the following result holds.

Theorem 6 [37],[40]
The pair $(\hat{\Phi},\hat{\Gamma})$ is stabilizable if and only if the T-periodic pair $(\Phi(.),\Gamma(.))$ is stabilizable.■

Now let ν_R be the dimension of the reachability subspace of $(\hat{\Phi},\hat{\Gamma})$. Under the conditions of Theorem 1, and in view of Theorem 6, there exists a feedback control law for system (13)

$$\hat{r}(k) = \hat{K}\,\hat{\xi}(k) \tag{14}$$

such that the closed-loop matrix $\hat{\Phi}+\hat{\Gamma}\hat{K}$ has ν_R eigenvalues arbitrarily assigned, and $n+\nu_R+m$ eigenvalues (independent of $\hat{K}$) inside the unit circle. Notice that epression (14), in the light of the very definition of $\hat{\xi}$ and r, can be rewritten as follows:

$$r(kT+i) = \hat{K}_i\,\xi(kT)\ ,\ i=0,1,\ldots,T-1 \tag{15}$$

where

$$\hat{K} := [\hat{K}_0 \quad \hat{K}_1 \quad \ldots \quad \hat{K}_{T-1}]$$

Equation (15) is a so-called generalized sampled-data hold control function. Implemented on system (7.a) (with $w_1\equiv 0$), it leads to a T-periodic system whose monodromy matrix is exactly $\hat{\Phi}+\hat{\Gamma}\hat{K}$. Notice that the control law (15) corresponds to the following T-periodic state controller

$$z(t+1) = Q(t)\,z(t) + (I-Q(t))\,\xi(t) \tag{16.a}$$

$$r(t) = K(t)\,[Q(t)\,z(t) + (I-Q(t))\,\xi(t)] \tag{16.b}$$

where

$$Q(t) := \begin{cases} I & t\neq kT \\ 0 & t=kT \end{cases}$$

and

$$K(i) := \hat{K}_1 \ , \ i=0,1,\ldots,T-1 \qquad (17)$$

System (16) corresponds to eq. (15) since, as it is easily seen, (16.a) corresponds to $z(kT+i)=\xi(kT)$, $i=1,2,\ldots,T$, so that (16.b) becomes $r(kT+i)=K(i)\xi(kT)$, $i=0,1,\ldots,T-1$. Hence

$$\xi(kT+T) = (\hat{\Phi}+\hat{\Gamma}\hat{K})\ \xi(k)$$

$$z(kT+T) = \xi(kT)$$

so that the monodromy matrix of the T-periodic closed-loop system (7.a), (16) is

$$\begin{bmatrix} \hat{\Phi}+\hat{\Gamma}\hat{K} & 0 \\ I & 0 \end{bmatrix}$$

which corresponds to the dynamical matrix

$$\bar{\phi}_1(t) := \begin{bmatrix} \Phi(t)+\Gamma(t)K(t)(I-Q(t)) & \Gamma(t)K(t)Q(t) \\ I-Q(t) & Q(t) \end{bmatrix}$$

of system (7.a), (16).

The discussion above can be summarized in the following result.

Theorem 8

Suppose that the conditions of Theorem 1 hold. Then, there exists a periodic matrix K(.) such that the closed-loop system (7.a), (16) has ν_R characteristic multipliers arbitrarily assigned, m+n characteristic multipliers at the origin and the remaining $m+n-\nu_R$ characteristic multipliers inside the unit circle (in positions independent of K(.))■

Notice that, according to control law (15), the state $\xi(t)$ is fed-back only once in the period, so that the

control is open-loop in the intersampling. This fact is not completely satisfactory from a practical point of view. In the direction of obtaining a more robust state control law, one may consider

$$r(kT+i) = \Omega(i)\ \xi(kT+i) \tag{18}$$

instead of (15). It may be proven that the free motion of system (7.a), (18) coincides with that of system (7.a), (15) if and only if

$$\hat{K}_0 = \Omega(0)$$

$$\hat{K}_i = \Omega(i)\ (\Phi(i-1)+\Gamma(i-1)\Omega(i-1))\ldots(\Phi(0)+\Gamma(0)\Omega(0)) \qquad i=1,2,\ldots,T-1$$

Hence, if such equations are solvable with respect to $\Omega(.)$, then the control law (18) is easily derived from (15).

B. LINEAR-QUADRATIC CONTROL

The classical LQ problem for system (7.a) is based on the definition of a suitable quadratic performance index to minimized, i.e.,

$$J(t_0,t_1) := \sum_{t=t_0}^{t_1} (\xi(t+1)'\Theta(t)\xi(t+1) + r(t)'\Lambda(t)r(t)) \tag{19}$$

where $\Theta(.)$ and $\Lambda(.)$ are T-periodic matrices. Moreover, $\Lambda(t)=\Lambda(t)'>0$ and $\Theta(t)=\Theta(t)'\geq 0$, $\forall t$.

It is well known that this problem (finite horizon) is solved by the control law

$$r(t) = K(t)\ \xi(t) \tag{20}$$

where

$$K(t) := -[\Lambda(t)+\Gamma(t)'P_c(t)\Gamma(t)]^{-1}\Gamma(t)'P_c(t)\Phi(t) \tag{21}$$

and $P_c(.)$ is the solution of the difference T-periodic Riccati equation

$$P_c(t) = \Phi(t)'P_c(t+1)\Phi(t)+\Theta(t)+$$

$$-\Phi(t)'P_c(t+1)\Gamma(t)[\Lambda(t)+\Gamma(t)'P_c(t+1)\Gamma(t)]^{-1}\Gamma(t)'P_c(t+1)\Phi(t) \tag{22}$$

with the terminal condition $P_c(t_1)=\Theta(t_1)$. Moreover, the optimal performance index is $J^o(t_0,t_1)=\xi(t_0)'P_c(t_0)\xi(t_0)$.

The case where $t_1\to\infty$ (infinite horizon) is the argument of the following result.

Theorem 9 [41]
If the pair $(\Phi(.),\Gamma(.))$ is stabilizable and the pair $(\Phi(.),\Theta(.))$ is detectable, then

$$\lim_{t_1\to+\infty} P_c(t) = \bar{P}_c(t) \geq 0 \ , \ \forall\ t \tag{23}$$

which is the unique T-periodic positive semidefinite solution of (22) for any $P_c(t_1)\geq 0$. Moreover, the closed-loop system (7.a), (20) is asymptotically stable, i.e., all the characteristic multipliers of the T-periodic matrix $(\Phi(.)+\Gamma(.)K(.))$ are inside the unit disk. ■

The control law of Theorem 9 is also optimal when the minimum of the expected value of

$$\lim_{t_1\to\infty} \frac{1}{t_1-t_0} J(t_0,t_1)$$

is sought in the presence of a stochastic noise $w_1\not\equiv 0$ in eq. (7.a).

V. STATE-OBSERVERS

The problem of state reconstruction is a classical one

in control theory. Here, the two most common approaches are pursued, namely pole-assignement and Kalman filtering.

Reference will be made to system (7). However, the theory to be developed can easily be modified so as to apply it to system (1), which is very reasonable, since the state of (3) is always available for measure.

A. POLE-PLACEMENT

The TIR of system (7) with $w_1 \equiv 0$ and $w_2 \equiv 0$ is given by the state equation (13) together with the output transformation

$$\hat{\zeta}(k) = \hat{\Delta}\ \hat{\xi}(k) + \Xi\ \hat{r}(k) \tag{24}$$

where

$$\hat{\zeta}(k) := [\zeta(kT)' \quad \zeta(kT+1)' \ \dots\ \zeta(kT+T-1)']'$$

For system (13), (24) consider the observer

$$\hat{\xi}_o(k+1) = \hat{\Phi}\ \hat{\xi}_o(k) + \hat{\Gamma}\ \hat{r}(k) + \hat{L}\ [\hat{\Delta}\ \hat{\xi}_o(k) + \Xi\ \hat{r}(k) - \hat{\zeta}(k)] \tag{25}$$

where $\hat{L} \in R^{n+m,pT}$ is an arbitrarily chosen matrix, which can be partitioned as

$$\hat{L} := [\hat{L}_0 \quad \hat{L}_1 \ \dots\ \hat{L}_{T-1}]\ ,\ \hat{L}_i \in R^{n+m,p}\ ,\ i = 0,\ 1,\ \dots, T-1$$

System (25) is characterized by the dynamical matrix $\hat{\Phi}+\hat{L}\hat{\Delta}$.

The detectability of the pair $(\hat{\Phi},\hat{\Delta})$ is the object of the following result.

Theorem 10 [37],[40]
The pair $(\hat{\phi},\hat{\Delta})$ is detectable if and only if the pair $(\Phi(.),\Delta(.))$ is detectable. ■

Let ν_o be the dimension of the unobservability subspace of $(\hat{\phi},\hat{\Delta})$. Under the conditions of Theorem 3 and in view of

Theorem 10, there exists a matrix $\hat{L}$ such that $\hat{\Phi}+\hat{L}\hat{\Delta}$ has $n+m-\nu_o$ eigenvalues arbitrarily assigned and ν_o eigenvalues (independent of $\hat{L}$) inside the unit circle.

In order to enlight the behavior of system (25) in terms of the original T-periodic system (7), let L(.) be the T-periodic matrix such that

$$L(i) := \hat{L}_i \ , \ i = 0, 1, \ldots, T-1$$

Moreover, define the T-periodic system

$$\xi_a(t+1) = \Phi(t)\ \xi_a(t) + \Gamma(t)\ r(t) + V(t+1)\ q(t+1) \qquad (26.a)$$

$$q(t+1) = (I-V(t))\ q(t) + L(t)\ (\Delta(t)\ \xi_a(t) - \zeta(t)) \qquad (26.b)$$

where

$$V(t) := \begin{cases} 0 & i = 1, 2, \ldots, T-1 \\ I & i = 0 \end{cases}$$

Such a system is characterized by the dynamical matrix

$$\bar{\Phi}_2(t) = \begin{bmatrix} \Phi(t)+V(t+1)L(t)\Delta(t) & V(t+1) \\ L(t)\Delta(t) & I-V(t)) \end{bmatrix}$$

to which it corresponds the mpnodromy matrix

$$\begin{bmatrix} \hat{\Phi}+\hat{L}\hat{\Delta} & 0 \\ ? & 0 \end{bmatrix}$$

Hence, the following theorem holds.

Theorem 11
Suppose that the conditions of Theorem 3 hold. Then, there exists a T-periodic matrix L(.) such that system (26) has $n+m-\nu_o$ characteristic multipliers arbitrarily assigned, n+m

characteristic multipliers in the origin, and the remaining ν_o characteristic multipliers inside the unit circle (in positions independent of L(.)).■

Now define the error

$$\varepsilon_1(t) := \xi_a(t) - \xi(t)$$

and

$$\varepsilon(t) := [\varepsilon_1(t)' \quad q(t)']'$$

Then, simple computations yeld $\varepsilon(t+1)=\bar{\Phi}_2(t)\ \varepsilon(t)$.

Hence, if $\hat{\Phi}+\hat{L}\hat{\Delta}$ is asymptotically stable, the error $\varepsilon(t)$ asymptotically vanishes, so that $\xi_a(t)$ tends to $\xi(t)$ as $t\to\infty$. Further,

$$\hat{\varepsilon}_1(k+1) = (\hat{\Phi}+\hat{L}\hat{\Delta})\ \hat{\varepsilon}_1(k)$$

yelds, asymptotically, $\hat{\xi}_a(k)= \xi_a(kT)$.

B. KALMAN FILTERING

The goal of this section is to construct a Kalman predictor for the state of system (7). To this regard, first note that the covariance matrix of the noise affecting the output $\zeta(t)$ is singular due to the structure of matrix $\Omega(t)$. As such, standard theory cannot be applied directly. Instead of (7), consider system (12) and notice that the covariance matrix of the noise affecting η_1 is nonsingular for each t with time-varying dimension.

With the aim of obtaining for system (12) an output vector of fixed dimension p and affected by a noise with nonsingular covariance, it is possible, following the line of reasoning adopted in [42], to add a slack output

$$\eta_3(t) = w_3(t)$$

with $[\eta_1' \quad \eta_3']' \in R^p$, $w_3(t) \sim WGN(0, W_3(t))$, $W_3(t) > 0$, $\forall$ t, and w_3 independent from the other noises. Obviously, these new output components do not bring any piece of information of the system state $\xi(t)$.

Turning back to the output variable $\zeta(t)$ of (7), adding $\eta_3(t)$ to $\eta_2(t)$ in (11.b), and recalling (10.b), it follows that

$$\zeta(t) = \Delta(t)\ \xi(t) + w(t) \tag{27}$$

where

$$w(t) = \Omega(t)\ w_2(t) + P_y(t)^{-1}\ [0 \quad \eta_3(t)']'$$

The covariance matrix of w(t) is

$$W(t) = \Omega(t)\ \Omega(t)' + P_y(t)^{-1} \begin{bmatrix} 0 & 0 \\ 0 & W_3 \end{bmatrix} P_y(t)^{-1}, =$$

$$= P_y(t)^{-1} \begin{bmatrix} \nabla_y(t) M_2 M_2' \nabla_y(t)' & 0 \\ 0 & W_3 \end{bmatrix} P_y(t)^{-1}, > 0$$

The discussion above allows us to conclude that eq. (7) can be suitably adopted as the output transformation of (7.a). Notice, however, that we need not explicitely compute the permutation matrix $P_y(t)$, but it suffices to replace the zeros of the diagonal of

$$\Omega(t)\Omega(t)' = N(t)M_2M_2'N(t)$$

with positive numbers.

The one step ahead optimal predictor for system (7.a), (27) is then given by

$$\xi_a(t+1) = \Phi(t)\ \xi_a(t) + \Gamma(t)\ r(t) + L(t)\ [\Delta(t)\ \xi_a(t) - \zeta(t)] \tag{28}$$

where

$$L(t) := - \Phi(t)\, P_F(t)\, \Delta(t)' \,[W(t) + \Delta(t)\, P_F(t)\, \Delta(t)']^{-1} \qquad (29)$$

and $P_F(t)$ is the solution of the difference T-periodic Riccati equation

$$P_F(t+1) = \Phi(t)\, P_F(t)\, \Phi(t)' + \Psi\, \Psi' - L(t)\,[W(t) + \Delta(t)\, P_F(t)\, \Delta(t)']\, L(t)' \qquad (30)$$

with the initial condition

$$P_F(t_0) = E\{\xi(t_0)\, \xi(t_0)'\}$$

The case where $t_0 \to -\infty$ is the argument of the following result.

Theorem 12 [41]
If the pair $(\Phi(.),\Delta(.))$ is detectable and the pair $(\Phi(.),\Psi)$ is stabilizable, then

$$\lim_{t_0 \to -\infty} P_F(t) = \bar{P}_F(t) \geq 0 \,, \quad \forall\, t \qquad (31)$$

which is the unique T-periodic positive semidefinite solution of (30) for any $P_F(t_0) \geq 0$. Moreover, the predictor (28) is asymptotically stable, i.e., all the characteristic multipliers of the T-periodic matrix $(\Phi(.)+L(.)\Delta(.))$ are inside the unit disk. ■

VI. OUTPUT FEEDBACK CONTROL

In this section, the previous results on state feedback control laws and state observers are joint together with the aim of deriving stabilizing output feedback control laws. In general, it could be possible to

implement any state control law on any state reconstruction. However, according to a classical approach, in following purely deterministic output pole-assignement and optimal LQG techniques are pursued.

A. POLE-PLACEMENT

Consider the T-periodic regulator constituted by eqs. (26) along with eqs. (16), where ξ is substituted by ξ_a:

$$\xi_a(t+1) = \Phi(t)\,\xi_a(t) + \Gamma(t)\,r(t) + V(t+1)\,q(t+1) \tag{32.a}$$

$$q(t+1) = (I-V(t))\,q(t) + L(t)\,[\Delta(t)\,\xi_a(t) - \zeta(t)] \tag{32.b}$$

$$z(t+1) = Q(t)\,z(t) + (I-Q(t))\,\xi_a(t) \tag{32.c}$$

$$r(t) = K(t)\,[Q(t)\,z(t) + (I-Q(t))\,\xi(t)_a] \tag{32.d}$$

In eqs. (32), matrix K(t) [L(t)] is chosen as described in Section IV.A. [V.A.], so as to assign the eigenvalues of $\Phi+\Gamma K$ [$\Phi+L\Delta$].

By combining eqs. (7) and (33), disregarding again the exogenous signals w_1 and w_2, we obtain

$$\begin{bmatrix} \xi(t+1) \\ z(t+1) \\ \hdashline \varepsilon(t+1) \end{bmatrix} = \left[\begin{array}{c:c} \bar{\Phi}_1(t) & ? \\ \hdashline 0 & \bar{\Phi}_2(t) \end{array}\right] \begin{bmatrix} \xi(t) \\ z(t) \\ \hdashline \varepsilon(t) \end{bmatrix}$$

It is then apparent that the separation principle holds for the characteristic multipliers of the closed-loop system.

The previous results are summarized in the following theorem.

Theorem 13

If

(i) The pair (A,B) is stabilizable and the pair (A,C) is detectable;

(ii) do not exist two distinct eigenvalues of A, λ_i and λ_j, $|\lambda_i|\geq 1$, $|\lambda_j|\geq 1$, such that $\lambda_i^T=\lambda_j^T$;

(iii) do not exist eigenvalues λ of A, $\lambda\neq 1$, $|\lambda|=1$, $\lambda^T=1$;

Then, the closed-loop system (7), (30) has:

A. ν_R characteristic multipliers arbitrarily assignable by a proper choice of the T-periodic matrix K(.);

B. $n+m-\nu_0$ characteristic multipliers arbitrarily assignable by a proper choice of the T-periodic matrix L(.);

C. $2n+2m$ characteristic multipliers at the origin;

D. $n+m+\nu_0-\nu_R$ characteristic multipliers inside the unit circle. ■

B. LINEAR QUADRATIC GAUSSIAN CONTROL

Consider the following quadratic performance index for system (7):

$$J := \lim_{\substack{t_0\to-\infty \\ t_1\to\infty}} \frac{1}{t_1-t_0}\, \mathcal{E}\Big[\sum_{t=t_0}^{t_1} (\xi(t+1)'\Theta(t)\xi(t+1)+r(t)'\Lambda(t)r(t))\Big] \tag{33}$$

where $\Theta(.)$ and $\Lambda(.)$ are T-periodic matrices with $\Theta(t)=\Theta(t)'\geq 0$, $\Lambda(t)=\Lambda(t)'>0$, $\forall t$.

By applying the separation theorem, the overall optimal regulator is obtained by substituting the optimal prediction $\xi_a(t)$ to $\xi(t)$ in (20), that is, by letting

$$r(t) = K(t)\, \xi_a(t) \tag{34.a}$$

where

$$\xi_a(t+1) = \Phi(t)\, \xi_a(t) + \Gamma(t)\, r(t) + L(t)\, [\Delta(t)\, \xi_a(t) - \zeta(t)] \tag{34.b}$$

In eqs. (34), matrix K(t) [L(t)] is chosen as in eqs.

(21)-(23) [(29)-(31)] so as to solve an optimal LQ control problem [Kalman prediction problem].

Combining eqs. (7) and (34), one gets

$$\varphi(t+1) = \bar{\Phi}_3(t)\ \varphi(t) + \sigma(t) \tag{35}$$

where

$$\varphi(t) := [\xi(t)' \quad \varepsilon_1(t)']'$$

$$\bar{\Phi}_3(t) := \begin{bmatrix} \Phi(t)+\Gamma(t)K(t) & \Gamma(t)K(t) \\ 0 & \Phi(t)+L(t)\Delta(t) \end{bmatrix} \tag{36}$$

$$\sigma(t) := \begin{bmatrix} \Psi & 0 \\ -\Psi & -L(t)\Omega(t) \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}$$

Note that the noise $\sigma(.)$ is cyclostationary with period T, zero mean-value and covariance $\Sigma(t)$.

Due to the asymptotic stability of matrix $\bar{\Phi}_3(t)$, the stochastic process $\varphi(.)$ of eq. (35) converges to a cyclostationary one $\varphi_s(.)$, whose associated covariance matrix

$$F(t) := \mathcal{E}[\varphi_s(.)\varphi_s(.)']$$

is the unique positive semidefinite T-periodic solution of the T-periodic Lyapunov equation

$$F(t+1) = \bar{\Phi}_3(t)\ F(t)\ \bar{\Phi}_3(t)' + \Sigma(t) \tag{37}$$

The optimal value J^o of the performance index (33) can now be written in terms of matrix F(.). Thanks to the stability property of (35), (36), the sum of a finite number of terms in the performance index (33) can be dropped, so obtaining

$$J^{o} := \lim_{\substack{k_0 \to -\infty \\ k_1 \to +\infty}} \frac{1}{(k_1 - k_0)T+\gamma} \mathcal{E}\left[\sum_{t=k_0 T}^{k_1 T} (\xi_s(t)' \Theta(t) \xi_s(t) + r_s(t)' \Lambda(t) r_s(t) \right]$$

where $\xi_s(.)$ and $r_s(.)$ are the asymptotic cyclostationary processes relative to $\xi(.)$ and $r(.)$, respectively, and γ is an appropriate constant integer.

By recalling eqs. (34.a), (35) and using standard results, one obtains

$$J^{o} = \lim_{\substack{k_0 \to -\infty \\ k_1 \to +\infty}} \frac{1}{(k_1 - k_0)T+\gamma} \operatorname{tr}\left(\sum_{t=k_0 T}^{k_1 T} G(t)\, \mathcal{E}[\varphi_s(t)\varphi_s(t)'] \right)$$

where

$$G(t) := \begin{bmatrix} \Theta(t)+K(t)'\Lambda(t)K(t) & +K(t)'\Lambda(t)K(t) \\ +K(t)'\Lambda(t)K(t) & K(t)'\Lambda(t)K(t) \end{bmatrix}$$

Further, the periodicity of $F(.)$ implies that

$$J^{o} = \frac{1}{T} \operatorname{tr}\left(\sum_{t=0}^{T-1} G(t)F(t) \right) \qquad (38)$$

Finally, the arguments above can be summarized in the following theorem, which formally supplies the solution of the LQG control problem for the considered class of multirate sampled-data systems.

Theorem 14 [26]
If

(i) The pair (A,B) is stabilizable and the pair (A,C) is detectable;
(ii) The pair (A,M_1) is stabilizable;

(iii) Do not exist two distinct eigenvalues of A, λ_i and λ_j, $|\lambda_i| \geq 1$, $|\lambda_j| \geq 1$, such that $\lambda_i^T = \lambda_j^T$;

(iv) Do not exist eigenvalues λ of A, $\lambda \neq 1$, $|\lambda| = 1$, such that $\lambda^T = 1$;

(v) The pair $(\Phi(.), \Gamma(.))$ is detectable;

Then

(a) The solution of the LQG problem (7), (33) exists and is given by system (34);

(b) The closed-loop T-periodic system is asymptotically stable and given by eqs. (35), (36);

(c) The optimal performance index is given by (38), where F(.) is the unique T-periodic positive semidefinite solution of the T-periodic difference Lyapunov equation (37). ■

VII. OUTPUT REGULATION

The classical robust output regulation problem consists of determining a suitable regulator which guarantees the asymptotic tracking of given reference signals in spite of the presence of persistent disturbances and plant uncertainties.

For this kind of problem to have a solution, a well known fact is that the control signals must be free to cover the same functional class as that of the reference and disturbance signals. It is then apparent that problems generally arise when dealing with nonstandard updating mechanisms, apart from the particular case where the exogenous signals are constant functions. Strictly speaking, the exact solution of the output regulation problem does not exist, and only partial solutions can be achieved [33].

Hence, from now on, it will be assumed that the plant input is updated at any time instant.

A. STATEMENT

The plant $\mathcal{P}$ under control is assumed square and described by

$$x(t+1) = Ax(t) + Bu(t) + D_1 d(t)$$

$$y(t) = Cx(t) + D_2 d(t)$$

where $A \in R^{n,n}$, $B \in R^{n,m}$, $D_1 \in R^{n,r}$, $C \in R^{m,n}$, $D_2 \in R^{m,r}$. Again, the output vector y_p is periodically measured according to (5).

The reference signal y^o for y and the disturbance d are generated by the following system $\mathcal{E}$:

$$x_e(t+1) = A_e x_e(t)$$

$$y^o(t) = C_{ey} x_e(t)$$

$$d(t) = C_{ed} x_e(t)$$

where $A_e \in R^{n_e,n_e}$.

To avoid trivialities, some standing assumptions on $\mathcal{E}$ are in order:

(i) Matrix A_e is known, whereas vector $x_e(0)$ and matrices C_{ey} and C_{ed} are unknown;

(ii) The eigenvalues of A_e all have magnitude greater than or equal to 1.

The controller $\mathcal{C}$ to be synthesized is a linear T-periodic ($T=\bar{T}$) system, generating the control u as a function of

$$e_\zeta(t) := \zeta^o(t) - \zeta(t)$$

where ζ^o is the output of the system $\mathcal{N}^o$ defined by

$$\mathcal{N}^o : \begin{cases} \zeta^o(t) = N(t) y^o(t) \end{cases}$$

Then, letting the control system error e be

$$e(t) := y^{o}(t) - y(t)$$

the problem considered in the paper is:

Output Robust Asymptotic Regulation Problem (ORARP)

Find a controller $\mathcal{C}$ such that:

(i) The closed-loop system $(\mathcal{P},\mathcal{N},\mathcal{C})$ is asymptotically stable;

(ii) The output regulation constraint

$$\lim_{t\to\infty} e(t) = 0$$

holds true for any $y^{o}(t)$ and $d(t)$, generated by $\mathcal{E}$ with any $x_e(0)$, in a robust way, i.e., for all perturbations of matrices A, B, C, D_1 and D_2, which preserve the asymptotic stability of the closed-loop system $(\mathcal{P},\mathcal{N},\mathcal{C})$.■

The block scheme of the overall control system is reported in Fig. 1.

B. SOLUTION

Denote by $\mathcal{L}_e$ the set of the distinct eigenvalues of $\mathcal{E}$ and by $\mathcal{L}$ the union (without repetitions) of $\mathcal{L}_e$ and the set of the distinct eigenvalues of system $\mathcal{P}$. Then, letting 0 denote the zero matrix of any size, the following solvability condition for ORARP can be stated:

Theorem 15 [32],[34]

Suppose that:

(i) The pair (A,B) is stabilizable;

(ii) The pair (A,C) is detectable;

(iii) $$\mathrm{Det}\left(\begin{bmatrix} A-\lambda_e I & B \\ C & 0 \end{bmatrix}\right) \neq 0 \,, \quad \lambda_e \in \mathcal{L}_e$$

(iv) There does not exist a couple of elements of $\mathcal{L}$, λ_i and λ_j, $|\lambda_i| \geq 1$, $|\lambda_j| \geq 1$, such that $\lambda_i^T = \lambda_j^T$.

Then, ORARP admits a solution. ■

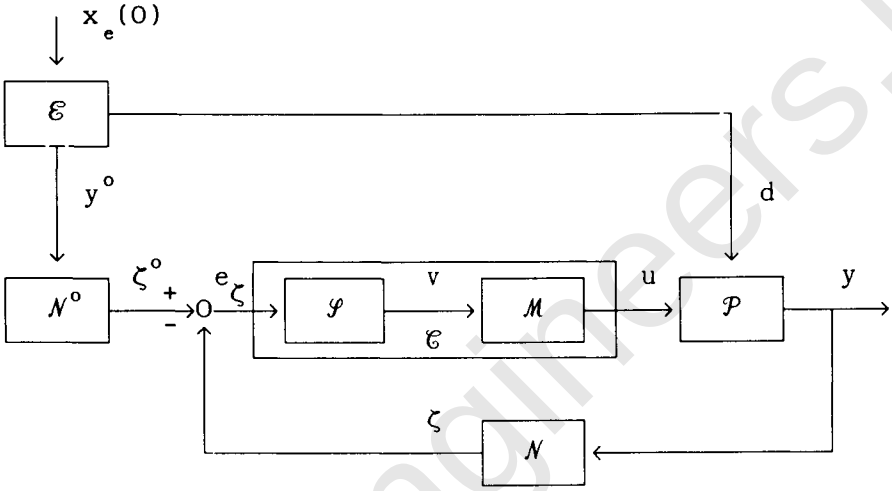


Fig. 1: The block scheme of the overall control system.

This theorem supplies a sufficient condition for the existence of $\mathcal{C}$; in order to state a result specifying its structure, some definitions are necessary.

Let the minimal polynomial of matrix A_e be

$$z^{\nu} + a_{e1} z^{\nu-1} + a_{e2} z^{\nu-2} + \ldots + a_{e\nu-1} z + a_{e\nu}$$

Then, define the matrices

$$\bar{A}_m := \begin{bmatrix} 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & \dots & 0 & 1 \\ -a_{e\nu} & -a_{e\nu-1} & \dots & -a_{e2} & -a_{e1} \end{bmatrix}$$

$$\bar{B}_m := \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \end{bmatrix}',$$

$$\bar{C}_m := \begin{bmatrix} -a_{e\nu} & -a_{e\nu-1} & \dots & -a_{e2} & -a_{e1} \end{bmatrix}$$

$$A_m := \text{diag}\,\{\bar{A}_m\} \in \mathbb{R}^{m\nu \times m\nu}$$

$$B_m := \text{diag}\,\{\bar{B}_m\} \in \mathbb{R}^{m\nu \times \nu}$$

$$C_m := \text{diag}\,\{\bar{C}_m\} \in \mathbb{R}^{\nu \times m\nu}$$

and call $\mathcal{M}$ the system

$$x_m(t+1) = A_m x_m(t) + B_m v(t)$$

$$u(t) = C_m x_m(t) + v(t)$$

which is an m-fold reduplication internal to $\mathcal{C}$ of the system $\mathcal{E}$ modelling the exogenous signals.

Further, let $\tilde{\mathcal{S}}$ be the set of all the T-periodic systems $\mathcal{S}$ with input e and output v such that the closed-loop system $(\mathcal{P},\mathcal{N},\mathcal{S},\mathcal{M})$ is asymptotically stable.

Theorem 16 [32],[34]
Suppose that conditions (i)-(iv) of Theorem 14 hold. Then:
(i) The set $\tilde{\mathcal{S}}$ is nonempty;
(ii) For any $\mathcal{S} \in \tilde{\mathcal{S}}$, the controller $\mathcal{C}$ constituted by the cascade connection of $\mathcal{S}$ and $\mathcal{M}$ solves ORARP. ■

Any stabilization method valid for this class of

multirate systems can be used to synthesize system $\mathcal{S}$, for instance the pole-placement or the LQG methods. These techniques should be applied to the cascade connection of systems $\mathcal{M}$, $\mathcal{P}$ and $\mathcal{N}$, with input v and output $e_\zeta=-\zeta$. Hence, the output tracking constraint is robust also with respect to variations of order and parameters of $\mathcal{S}$, as long as asymptotic stability is preserved. The same property holds for as concerns variations of the order n of the plant $\mathcal{P}$.

It is worth noticing that, though the input of $\mathcal{C}$ is the variable e_ζ, the controller is able to asymptotically bring to zero the difference e between the reference signal y^o and the plant output y at all time instants, not only at times where the output is measured.

The block scheme of Fig. 1 shows that the structure of the overall control system is very similar to a possible one for monorate systems. However, by comparing the sufficient conditions of Theorem 1 for the solvability of ORARP with the discrete-time version of the necessary and sufficient conditions for the solvability of the same problem in the monorate case, it may be observed that condition (iv) here does not have a counterpart there. As a matter of fact, this condition guarantees (but is not necessary for) the detectability from y of the state of the system $(\mathcal{M},\mathcal{P},\mathcal{N})$, along with the standard conditions (ii) and (iii). The consequence of that is the impossibility of asymptotically zeroing this way the system error for some exogenous signals for whom it can be brought to zero when the outputs are always measured.

VIII. CONCLUDING REMARKS

This paper has reviewed some recent results about stabilization and regulation of multirate sampled-data systems.

First, a precise mathematical formulation of the input and output mechanisms has been given in terms of a discrete-time periodic system. Then, its structural properties have been investigated and related to those of the underlying time-invariant plant.

The classical pole-placement and LQ techniques have

then been used for deriving stabilizing state feedback control laws and stable state observers.

Finally, it has been shown how to select a proper regulator structure, which, in some significant cases, guarantees zero-error regulation in the face of wide classes of exogenous signals, despite the possible lack of information due to the outputs-sampling.

APPENDIX

Let S be the discrete-time T-periodic system described by

$$x(t+1) = A(t)\,x(t) + B(t)\,u(t)$$

$$y(t) = C(t)\,x(t) + D(t)\,u(t)$$

where $t \in Z$, $A(t) \in R^{n,n}$, $B(t) \in R^{n,m}$, $C(t) \in R^{p,n}$, $D(t) \in R^{p,m}$.

Denote by $\Phi_A(t,\tau)$, $t>\tau$, the transition matrix of $A(.)$, i.e., $\Phi_A(t,\tau) := A(t-1)A(t-2)\ldots A(\tau)$. Matrix $\Phi_A(\tau+T,\tau)$ is the so-called monodromy matrix associated with $A(.)$. Its eigenvalues do not depend on τ and are called characteristic multipliers of S. System S is asymptotically stable if and only if all its characteristic multipliers are inside the open unit disk.

Now, define the lifted input, the sampled state and the lifted output as

$$\hat{u}(k) := [u(kT+\tau)' \quad u(kT+\tau+1)' \quad \ldots \quad u(kT+\tau+T-1)']'$$

$$\hat{x}(k) := x(kT+\tau)$$

$$\hat{y}(k) := [y(kT+\tau)' \quad y(kT+\tau+1)' \quad \ldots \quad y(kT+\tau+T-1)']'$$

respectively, where τ is a given initial sampling time $(0 \le \tau \le T-1)$ and $\hat{x}(\tau)=x(0)$. The lifted or time-invariant (TIR) reformulation $\hat{S}$ associated with S is easily obtained from the definitions above as

$$\hat{x}(k+1) = \hat{A}\,\hat{x}(k) + \hat{B}\,\hat{u}(k)$$

$$\hat{y}(k) = \hat{C}\,\hat{x}(k) + \hat{D}\,\hat{u}(k)$$

where

$$\hat{A} := \Phi_A(\tau+T,\tau)$$

$$\hat{B} := [\hat{B}_1 \quad \hat{B}_2 \quad \dots \quad \hat{B}_T] \,, \quad B_i \in R^{n,m}$$

$$\hat{C} := [\hat{C}_1' \quad \hat{C}_2' \quad \dots \quad \hat{C}_T']' \,, \quad C_i \in R^{p,n}$$

$$\hat{D} := \{\hat{D}_{ij},\ \hat{D}_{ij} \in R^{p,m},\ i=1,2,\dots,T,\ j=1,2,\dots,T\}$$

$$\hat{B}_i := \Phi_A(\tau+T,\tau+i)B(\tau+i-1)$$

$$\hat{C}_i := C(\tau+i-1)\Phi_A(\tau+i-1,\tau)$$

$$\hat{D}_{ij} := \begin{cases} C(\tau+i-1)\Phi_A(\tau+i-1,\tau+j)B(\tau+j-1) \,, & i > j \\ D(\tau+i-1) \,, & i = j \\ 0 \,, & i < j \end{cases}$$

Of course, matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ depend on the choice of the initial sampling time τ. It can be shown that the nonzero poles and zeros of $\mathcal{S}$ are independent of τ. From the very definition of the lifted system, it can immediately be recognized that $\mathcal{S}$ and $\hat{\mathcal{S}}$ share the same structural properties. For instance, the pair $(A(.),B(.))$ is stabilizable if and only if $(\hat{A},\hat{B})$ is stabilizable, and so forth.

REFERENCES

[1] B.A. Francis and T.T. Georgiou, "Stability theory for linear time-invariant plants with periodic digital controllers, " *IEEE Trans. Automat. Contr.*, AC-33, 820-832, 1988.

[2] T. Hagiwara and M. Araki, "Design of stable state

feedback controller based on multirate sampling of the plant output," *IEEE Trans. Automat. Contr.*, AC-33, 812-819, 1988.

[3] M. Araki, "Recent development in digital control theory," *Proc. 12th IFAC World Congr.*, 9, 251-260, 1993.

[4] D.P. Glasson, "Research in multirate estimation and control," *The Analytic Science Corp.*, Rep. TR1356-2, 1980.

[5] P. Colaneri, R. Scattolini and N. Schiavoni, "A design technique for multirate control with application to a distillation column," *Proc. 12th IMACS World Congr.*, 589-591, 1988.

[6] R. Scattolini, "Self-tuning control of systems with infrequent and delayed output sampling," *Proc. IEE Part D*, 135, 213-221, 1988.

[7] J.H. Lee, M.S. Gelormino and M. Morari, "Model predictive control of multi-rate sampled-data systems: a state-space approach, 55, 153-191, 1992.

[8] P.T. Kabamba, "Control of linear systems using generalized sampled-data hold functions," *IEEE Trans. Automat. Contr.*, AC-32, 772-783, 1987.

[9] T. Mita, Y. Chida, Y Kaku and H. Numasato, "Two-delay robust digital control and its applications - avoiding the problem of unstable limiting zeros", *IEEE trans. Automat. Contr.*, AC-35, 962-970, 1990.

[10] K.L. Moore, S.P. Bhattacharyya and M. Dahleh, "Capabilities and limitations of multirate control schemes," *Automatica*, 29, 941-951, 1993.

[11] G.M. Kranc, "Input-output analysis of multirate feedbac system," *IRE Trans. Automat. Contr.*, 3, 21-28, 1957.

[12] R.E. Kalman and J.E. Bertram, "A unified approach to the theory of sampling systems," *J. Franklin Inst.*, 267, 405-436, 1959.

[13] E. I. Jury, "A note on multirate sampled-data systems," *IEEE Trans. Automat. Contr.*, Ac-12, 319-320, 1967.

[14] M. Araki and K. Yamamoto, "Multivariable multirate sampled-data systems: state space description, transfer characteristics, and Nyquist criterion," *IEEE Trans. Automat. Contr.*, AC-31, 145-154, 1986.

[15] A.B. Chammas and C.T. Leondes, "On the design of linear time invariant systems by periodic output feedback: Part I-II. Discrete-time pole assignment," *Int. J. Contr.*, 27, 885-903, 1978.

[16] A.B. Chammas and C.T. Leondes, "Pole assignment by piecewise constant output feedback," *Int. J. Contr.*, 29, 31-38, 1979.

[17] P.P. Khargonekar, K. Poolla and A. Tannenbaum , "Robust control of linear time invariant plants using periodic compensation," *IEEE Trans. Automat. Contr.*, AC-30, 1088-1096, 1985.

[18] M. Araki and T. Hagiwara, "Pole assignment by multirate sampled-data output feedback," *Int. J. Contr.*, 44, 1661-1673, 1986.

[19] T. Hagiwara and M. Araki, "Design of a stable feedback controller based on multirate sampling of the plant output," *IEEE Trans. Automat. Contr.*, AC-33, 812-819, 1988.

[20] P. Colaneri, R. Scattolini and N. Schiavoni, "Stabilization of multirate sampled-data systems," *Automatica*, 26, 377-380, 1990.

[21] P. Colaneri, R. Scattolini and N. Schiavoni, "Regulation of Multirate Sampled-Data Systems," *C-TAT*, 7, 429-441, 1991.

[22] N. Amit, "Optimal control of multirate digital control systems," Stanford Univ, Rep. N. SUDAAR 523, 1980.

[23] T. Söderström and B. Lennartson, "On linear optimal control with infrequent output sampling," *Proc. 3rd IMA Conf. on Control Theory*, 605-624, 1981.

[24] M.C. Berg, N. Amit and J.D. Powell, "Multirate digital control system design," *IEEE Trans. Automat. Contr.*, AC-33, 1139-1150, 1988.

[25] H.M. Al-Ramany and G.F. Franklin, "A new optimal multirate control of linear periodic and time-invariant systems," *IEEE Trans. Automat. Contr*, 35, 406-415, 1990.

[26] P. Colaneri, R. Scattolini and N. Schiavoni, "LQG optimal control of multirate sampled-data systems, *IEEE Trans. Automat. Contr.*, AC-37, 675-682, 1992.

[27] J.R. Broussard and N. Halyo, "Optimal multirate output feedback," *Proc. of 23rd Conf. Decision Contr.*, 1984, 926-929.

[28] R. Scattolini and N. Schiavoni, "Design of multirate control systems via parameter optimization," *Proc. of 26th Conf. Decision Contr.*, 1556-1557, 1987.

[29] M. E. Sezer and D.D. Siljak, "Decentralized multirate control, *IEEE Trans. Automat. Contr.*, AC-35, 60-65, 1990.

[30] P. Carini, R. Micheli and R. Scattolini, "Multirate self-tuning predictive control with application to a binary distillation column," *Int. J. System Sc.*, 21, 51-64, 1990.

[31] R. Scattolini, "A multirate self-tuning controller for multivariable systems," *Int. J. System Sc.*, 23, 1347-1359, 1992.

[32] P. Colaneri, R. Scattolini and N. Schiavoni, "Stabilization and regulation of multirate sampled-data systems," *Proc. Int. Symp. MTNS-91*, 511-516, 1991.

[33] P. Colaneri, R. Scattolini and N. Schiavoni, "The output control problem for multirate sampled-data systems", *Proc. 31 Conf. Dec. Contr.*, 1768-1773, 1992.

[34] R. Scattolini and N. Schiavoni, "On the output control of multirate systems subject to arbitrary exogenous signals," *IEEE Trans. Automat. Contr.*, 643-646, 1993.

[35] V.S. Ritchey and G.F. Franklin, "A stability criterion for asynchronous multirate linear systems," *IEEE Trans. Automat. Contr.*, AC-34, 529-535, 1989.

[36] R.A. Meyer and C.S. Burrus, "A unified analysis of multirate and periodically time-varying digital filters," *IEEE Trans. Circuits Syst.*, CAS-22, 162-167, 1975.

[37] S. Bittanti, "Deterministic and stochastic linear periodic systems," *Time Series and Linear Systems*, Springer Verlag, 141-182, 1986.

[38] P. Bolzern, P. Colaneri and R. Scattolini, "Zeros of discrete-time linear periodic systems," *IEEE Trans. Automat. Contr.*, AC-31, 1057-1058, 1986.

[39] O.M. Grasselli and S. Longhi, "Zeros and poles of linear periodic discrete-time systems," *Circuits Syst. Signal Process.*, 7, 361-382, 1988.

[40] O.M. Grasselli and S. Longhi, "The geometric approach for linear periodic discrete-time systems," *Linear Algebra Appl.*, 158, 27-60, 1991.

[41] S. Bittanti, P. Colaneri and G. De Nicolao, "The difference periodic Riccati equation for the periodic prediction problem," *IEEE Trans. Automat. Contr.*, AC-33, 706-711, 1988.

[42] P. Colaneri and G De Nicolao, "Optimal stochastic control of multirate sampled-data systems," *Proc. 1st European Contr. Conf.*, 2519-2523, 1991.

ACKNOWLEDGEMENTS

This paper has partially been supported by CNR (Centro di Teoria dei Sistemi) and MURST (40% and 60% funds).

Maximizing the Fisher Information Matrix in Discrete-Time Systems

Wendy L. Poston
Carey E. Priebe
O. Thomas Holland

Naval Surface Warfare Center, Dahlgren Div, G33
Dahlgren, Virginia 22448-5000

I. INTRODUCTION

An important part of research in control theory is that of developing and verifying a mathematical model of a system. Experimental tests should be performed using the real system, and results from these tests should then be compared with results from the model using similar conditions. If they differ significantly, then the model should be modified accordingly.

An example of this is verifying a finite element model [1] of a structure. When a vibration test is performed, data are collected from the sensors, and modal parameters such as mode shapes, frequencies, and damping rates must be extracted from the data. This information is then used to validate the model of the structure. To correlate the measured data with the analytical model, it is necessary to place sensors at locations that keep the mode shapes independent in the spatial domain. If it is not possible to spatially differentiate between them, then mode shape correlation using orthogonality and cross-orthogonality methods cannot be used.

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

One application where it is vitally important to place sensors properly is in vibration tests of large flexible space structures such as the space station. These will be constructed in space so a ground level vibration test is not possible. A limited number of sensors must be placed on each piece of the space station before it is sent into space. Once the structure is built, a test could be performed by exciting it orbit and measuring the free-decay response using accelerometers. If it is found that the sensors are incorrectly placed, then it is very expensive to move the sensors and repeat the test. Therefore, it is very important to place sensors properly prior to performing the test.

Several methods have been presented in the literature that address the problem of optimal sensor placement. They can be classified into the areas of system identification, state estimation, optimal control, and structural parameter identification. A survey of these sensor location methods was published by Kubrusly and Malebranche [2]. The methods discussed in the survey are based on the optimization of different criteria. These include minimizing the trace of the error covariance matrix, maximizing the kinetic energy, and maximizing the determinant of the information matrix. The method presented here addresses the problem of sensor placement for structural parameter identification and for validation of a finite element model of the structure. It maximizes the determinant of the Fisher Information Matrix (FIM).

In this chapter, a method of choosing optimal sensor locations based on the Effective Independence Distribution (EID) will be presented. This technique uses the EID to rank the contribution of each sensor measurement to the linear independence of a structure's mode shapes. Sensor locations can then be kept that will keep the mode shapes spatially independent. After the EID is developed, it will be shown that retaining locations with large values will maximize the determinant of the Fisher Information Matrix, thus increasing the amount of information available from the sensors. The computational cost of the process is analyzed, and simple examples are given that provide some insight into how the EID ranks locations. Finally, some additional applications of this technique to other problems are discussed, and it will be shown that it optimizes the observability of a system. Before the references, a list of symbols and acronyms is provided.

II. THEORETICAL BACKGROUND

The EID method will be applied to the problem of sensor placement for the purposes of model verification described previously. For this reason, the EID will be derived using the following mode shape equation for any one time step

$$\mathbf{h} = \Phi\mathbf{q} + \varepsilon \tag{1}$$

where $\mathbf{h}$ is an n x 1 column vector containing the measurements from the sensors, Φ is an n x p matrix of mode shapes such that each column Φ_j corresponds to the *jth* mode shape, and $\mathbf{q}$ is a column vector of mode shape scale factors that must be estimated because they are unobservable. The vector ε denotes the noise in the measurements, and it is assumed that

$$E[\varepsilon] = \mu = 0 \tag{2}$$

and

$$E[(\varepsilon - \mu)^2] = \Sigma \tag{3}$$

It might be useful to discuss how the matrix Φ is obtained before continuing. For a complicated structure, it can be calculated using finite element analysis techniques and software [1]. However, for a simple structure it can be determined analytically. Consider a uniform bar of length L hinged at both ends. It can be shown [3] that the normal modes are given by

$$\Phi_m(x) = q \sin\frac{m\pi x}{L} \qquad m = 1,2,\ldots$$

where q is a constant. The first step is to divide the bar into possible sensor locations; these will be the x coordinates. The values for m are chosen and these will determine the number of target modes (i.e., $m = 1,2,\ldots,p$). The desired mode shape matrix can now be calculated using the above equation where each row corresponds to an x coordinate and each column corresponds to a target mode.

The mode shape matrix Φ has p target modes that have to be recovered in a vibration test and an initial set of n candidate sensor locations. It is usually the case that $n>>p$. It is assumed that the initial set of p locations must be reduced to some smaller set such that $n \geq p$.

Equation (1) is in the familiar form of a least squares problem [4,5]. The minimum variance estimate for the scale factors $\mathbf{q}$ can be written

$$\hat{\mathbf{q}} = (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi^T \Sigma^{-1} \mathbf{h} \tag{4}$$

where Φ^T denotes the transpose of the matrix Φ. The covariance matrix in Eq. (4) is given by

$$\mathbf{COV} = (\Phi^T \Sigma^{-1} \Phi)^{-1} \tag{5}$$

and it is the inverse of the Fisher Information Matrix

$$\mathbf{COV}^{-1} = \mathrm{FIM}$$

The covariance matrix given by Eq. (5) provides statistical information for the estimate $\mathbf{q}$ and the matrix Σ is the covariance of the noise vector ε. Without loss of generality and for simplicity, we will take the covariance Σ to be the identity matrix. This means that the noise is uncorrelated and each element has a variance of one.

To get a good estimate of the scale factors $\mathbf{q}$, sensor locations should be chosen that will minimize a norm (e.g., trace, determinant, etc.) of the covariance matrix. Calculating the covariance matrix involves determining a matrix inverse which is an $O(p^3)$[1] process [6,7]. Since it is computationally expensive to have an objective function containing a matrix inverse, it would be better to maximize a norm on the FIM. It is known from information theory [8] that the determinant of the FIM is a measure of the amount of information provided by the sensor locations. Therefore, this will be the norm optimized

[1] $O(f(p))$ is the set of all functions $g(p)$ such that there exist positive constants c and N_o with $|g(p)| \leq c\, f(p)$ for all $p \geq N_o$. Thus, this notation indicates functions that are at most as large as some constant times $f(p)$ and can be thought of as an upper bound.

here, and sensor locations will be kept that maximize the determinant of the FIM.

The Effective Independence Distribution is an n-dimensional vector where each element corresponds to one sensor location. The development [9,10] of the EID method given in the next section will show that the *ith* term of the vector is the contribution of the *ith* sensor location to all of the eigenvalues of the FIM. Since

$$|\text{FIM}| = \prod_{j=1}^{p} \lambda_j \tag{6}$$

where $|\bullet|$ denotes the determinant, then the eigenvalues are also a measure of the information. If sensor locations with large EID values are kept, then this will maximize the eigenvalues and hence maximize the determinant.

III. THE EFFECTIVE INDEPENDENCE DISTRIBUTION METHOD

A. DEVELOPING EID FROM AN EIGENVALUE PROBLEM

The Effective Independence Distribution can be derived from the following eigenvalue problem

$$(\text{FIM} - \lambda_j \mathbf{I})\Psi_j = 0$$

where $\mathbf{I}$ is a p x p identity matrix, λ_j is the *jth* eigenvalue, and Ψ_j is the *jth* eigenvector. It is important to remember that the eigenvalues found here are not the natural frequencies of the structure; they are the eigenvalues of the FIM.

It follows from the definition of the information matrix that the FIM is symmetric. Since the columns of Φ are linearly independent, this implies that the FIM is also positive definite [11]. Therefore, the eigenvectors Ψ_j can be chosen to be orthonormal, which implies that

$$\Psi_i^T \Psi_j = 0, \quad i \neq j \qquad \text{and} \qquad \Psi_i^T \Psi_i = 1$$

Hence, the following matrix properties hold

$$\begin{aligned} \Psi^T \Psi &= \mathbf{I} \\ \text{FIM}\ \Psi &= \Lambda \Psi \end{aligned} \tag{7}$$

where Ψ is an orthonormal matrix with each column containing an eigenvector and Λ denoting a diagonal matrix of eigenvalues.

Starting from the second property given above and substituting for the FIM, we have

$$\Phi^T \Phi \Psi = \Lambda \Psi$$

Pre-multiplying by Ψ^T and using the first property in Eq. (7) yields

$$\Psi^T \Phi^T \Phi \Psi = \Lambda$$

After grouping terms, this can be written as

$$(\Phi \Psi)^T (\Phi \Psi) = \Lambda$$

It can be seen from this that the *jth* eigenvalue has the form

$$\lambda_j = \sum_{i=1}^{n} \left(\sum_{k=1}^{p} \phi_{ik} \psi_{kj} \right)^2, \quad j = 1, \ldots p \tag{8}$$

The eigenvectors of the information matrix span the p-dimensional mode shape space, so Ψ can be used to transform the mode shape matrix Φ. The following matrix product is now formed

$$\mathbf{G} = (\Phi \Psi) \otimes (\Phi \Psi) \tag{9}$$

where $\otimes$ denotes an element by element matrix multiplication and $\Phi\Psi$ represents the transformed mode shape matrix. The *ij-th* element of **G** is given by

$$g_{ij} = \left(\sum_{k=1}^{p} \phi_{ik} \psi_{kj}\right)^2 \tag{10}$$

An examination of each element of **G** reveals that the sum of the *jth* column of **G** equals the *jth* eigenvalue given in Eq. (8)

$$\sum_{i=1}^{n} g_{ij} = \lambda_j$$

The next step is to post-multiply **G** by Λ^{-1} forming the following matrix

$$\mathbf{E} = \mathbf{G}\Lambda^{-1}$$

The purpose of this step is to divide each column of **G** by the corresponding eigenvalue (i.e., the *jth* column of **G** is divided by the *jth* eigenvalue). Each column in the matrix **E** sums to one, and the element e_{ij} represents the fractional contribution of the *ith* sensor to the *jth* eigenvalue.

The Effective Independence Distribution is calculated by summing the terms in the *ith* row of the matrix **E**

$$\mathrm{EID}_i \equiv \sum_{j=1}^{p} e_{ij}, \qquad i = 1, \ldots, n \tag{11}$$

Thus, EID_i represents the contribution of the *ith* sensor to the eigenvalues of the Fisher Information Matrix. Note that there are *n* elements in the Effective Independence Distribution corresponding to each sensor location.

B. AN ALTERNATIVE CALCULATION OF THE EID

The diagonal elements of the following matrix

$$\mathbf{P} = \Phi(\Phi^T\Phi)^{-1}\Phi^T$$

will also yield the Effective Independence Distribution. To derive this equation, start with the definition of the *ith* element of the EID.

$$\mathrm{EID}_i = \sum_{j=1}^{p} e_{ij} = \sum_{j=1}^{p} \frac{g_{ij}}{\lambda_j}$$

and substituting for the *ij-th* element of **G** from Eq. (10) yields

$$\mathrm{EID}_i = \sum_{j=1}^{p} \left(\sum_{k=1}^{p} \frac{\phi_{ik}\psi_{kj}}{\sqrt{\lambda_j}} \right)^2$$

These are the diagonal elements of the following matrix product

$$\mathbf{P} = (\Phi\Psi\Lambda^{-1/2})(\Phi\Psi\Lambda^{-1/2})^T$$

where $\Lambda^{1/2}$ is a diagonal matrix containing the square roots of the eigenvalues. Re-arranging the matrices yields

$$\mathbf{P} = \Phi\Psi\Lambda^{-1}\Psi^T\Phi^T \tag{12}$$

Using the properties in Eq. (7), it can be shown that

$$\mathrm{FIM}^{-1} = \Psi\Lambda\Psi^T$$

Therefore, using Eq. (5) the matrix **P** can be re-written as

$$\mathbf{P} = \Phi(\Phi^T\Phi)^{-1}\Phi^T$$

This matrix has some interesting properties. First of all, it is an idempotent[2] matrix. These matrices have the property that the trace equals the rank [11,12], therefore

$$\sum_{i=1}^{n} \mathrm{EID}_i = \mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\Phi) = p$$

[2] An idempotent matrix is one that equals its square; i.e., $\mathbf{A}^2 = \mathbf{A}$.

So, each term in the EID can be said to show the contribution of the *ith* sensor location to the rank of the mode shape matrix and thus the linear independence of the modes. Secondly, the matrix **P** is a projection matrix which can be used to project any vector onto the column space of Φ. This means that the error component can be calculated using

$$\mathbf{e} = \hat{\mathbf{q}} - \mathbf{q} = \mathbf{P}\mathbf{q} - \mathbf{q}$$

The elements of the vector **e** are called the residuals, and they are the difference between the observed values and the estimated values. When performing a least squares analysis of a problem, the residuals should always be examined for indications that assumptions about the errors or noise have been violated [4,5].

C. USING THE EID TO CHOOSE OPTIMAL SENSOR LOCATIONS

The Effective Independence Distribution can be used in an iterative manner to determine optimal sensor locations. This method employs the following steps:

1. Start with a large set of potential sensor locations and the mode shape matrix Φ.
2. Calculate the Effective Independence Distribution for all sensors in the current set of locations.
3. Delete the sensor location with the smallest EID value.
4. Repeat steps 2 and 3 until the desired number of sensors is reached.

The sensor location with the smallest EID value is deleted because that sensor contributes the least amount of information and contributes the least to the linear independence of the mode shapes.

To see how computationally intensive this method is the following algorithm analysis is provided. This considers the calculation of the EID vector as the diagonal elements of the projection matrix **P**. A similar analysis of the algorithm can be performed using the definition of the EID given in Eq. (11). This is left to the interested reader. These operation counts indicate the number of loops, each loop containing a multiplication and an addition. For one iteration of the EID method, the following calculations are needed:

- Calculate $\Phi^T\Phi$, which is an $O(np^2)$ operation
- Find $(\Phi^T\Phi)^{-1}$, which is $O(p^3)$
- Multiply $(\Phi^T\Phi)^{-1}\Phi^T$, which is $O(np^2)$
- Find the diagonal elements only of $\Phi(\Phi^T\Phi)^{-1}\Phi^T$, which is an $O(np)$ operation

To determine the upper bounds on this algorithm, the step with the largest operation counts must be determined. Examining the above counts, the first inclination is to state that the calculation of the EID is $O(p^3)$. However, in most applications of the method $n >> p$ and this must be taken into account. The other steps that could produce upper bounds are the two that are $O(np^2)$. To see which one is greater, it is instructive to look at the ratio of the counts

$$\frac{p^3}{np^2} = \frac{p}{n}$$

The numerator is obviously smaller, given the conditions mentioned above. The value for n decreases (i.e., the number of possible sensor locations decreases) at every iteration, but it will never be less than p. Therefore the method is $O(np^2)$ for one iteration with n observations.

IV. MAXIMIZING THE DETERMINANT OF THE FIM

It was stated previously that in order to get the best estimate of the scale factors **q**, the sensor locations should be chosen such that a norm on the Fisher Information Matrix is maximized. The norm optimized here is the determinant of the FIM. In this section, it will be shown that deleting the sensor location with the smallest EID value will cause the least change in the determinant of the FIM, thus maximizing the determinant.

The eigenvectors of the FIM are an orthonormal basis for the p-dimensional column space of the mode shape matrix Φ. They also represent an identification ellipsoid in p dimensions with the axes of the ellipsoid pointing toward the eigenvectors. The determinant of the FIM is proportional to the

volume of this ellipsoid and is a measure of the amount of information provided by the sensors. When sensors are deleted from the candidate set of measurement locations, it is desirable to delete those sensor locations that change the volume of the ellipse by the least amount. If locations are deleted such that the determinant of the information matrix is maximized, then the volume of the identification ellipse will be maximized also.

Recall the following relationship from Eq. (6)

$$|\text{FIM}| = \prod_{j=1}^{p} \lambda_j$$

The relative change in the determinant can be found from the above equation using the differential approximation

$$\frac{\partial|\text{FIM}|}{|\text{FIM}|} = \frac{\partial\lambda_1}{\lambda_1} + \ldots\ldots + \frac{\partial\lambda_p}{\lambda_p}$$

It will be shown that when a sensor is deleted with the smallest EID value, then that will yield the smallest relative change given above.

Also recall the formulation of the **G** matrix in Eq. (9)

$$\mathbf{G} = (\Phi\Psi) \otimes (\Phi\Psi)$$

and that each column of the matrix **G** sums to the corresponding eigenvalue

$$\sum_{i=1}^{n} g_{i1} = \lambda_1, \sum_{i=1}^{n} g_{i2} = \lambda_2, \ldots, \sum_{i=1}^{n} g_{ip} = \lambda_p$$

so that the *ij-th* term is the contribution of sensor *i* to the *jth* eigenvalue. Using the definition of the *ith* element of the EID, the relative change in the determinant of the FIM from deleting the *ith* sensor location (i.e., the *ith* row of Φ) is

$$\left(\frac{\partial|\text{FIM}|}{|\text{FIM}|}\right)_i = \sum_{j=1}^{p} \frac{g_{ij}}{\lambda_j} = \sum_{j=1}^{p} e_{ij} = \text{EID}_i \qquad (13)$$

Because the terms in **E** are the fractional contribution of the sensors to the eigenvalues, then using those values to delete sensors will change the eigenvalue by that amount.

It can be seen from Eq. (13) that deleting a sensor with a small EID value will yield the smallest relative change in the determinant. It will be shown in this section that this differential approximation for the relative change in the determinant is actually exact and that the following holds true

$$\left| \text{FIM}_{-i} \right| = (1 - \text{EID}_i) \left| \text{FIM} \right| \tag{14}$$

where FIM_{-i} denotes the Fisher Information formed with the *ith* row of $\mathbf{\Phi}$ removed. One can see from this relationship that deleting the sensor location with the smallest EID value will maximize the determinant as desired.

However, prior to proving the above relationship, some information from matrix theory is needed [11]. Because the determinant of a matrix is linearly dependent on each row separately, the following relationships hold

$$\begin{vmatrix} a+b & c+d \\ e & f \end{vmatrix} = \begin{vmatrix} a & c \\ e & f \end{vmatrix} + \begin{vmatrix} b & d \\ e & f \end{vmatrix} \tag{15}$$

and

$$\begin{vmatrix} ka & kb \\ c & d \end{vmatrix} = k \begin{vmatrix} a & b \\ c & d \end{vmatrix} \tag{16}$$

The first theorem presented is needed to prove the relationship in Eq. (14).

THEOREM 1. If $\mathbf{A} = \mathbf{B} - \mathbf{r}^T\mathbf{r}$ where **A** and **B** are p x p matrices and **r** is a 1 x n row vector, then

$$|\mathbf{A}| = |\mathbf{B}| - \mathbf{r}\mathbf{B}_{cof}^T\mathbf{r}^T$$

where $\mathbf{B}_{cof}$ is the cofactor matrix of **B** with the cofactors of the *ith* row of **B** entered in the *ith* column.

PROOF: Since $\mathbf{A} = \mathbf{B} - \mathbf{r}^T\mathbf{r}$, then

$$|\mathbf{A}| = |\mathbf{B} - \mathbf{r}^T\mathbf{r}|$$

This can be expanded as

$$|\mathbf{A}| = \begin{vmatrix} b_{11} - r_1 r_1 & \cdots & b_{1p} - r_1 r_p \\ \vdots & \ddots & \vdots \\ b_{p1} - r_p r_1 & \cdots & b_{pp} - r_p r_p \end{vmatrix}$$

Using the properties in Eqs (15)-(16), the determinant of $\mathbf{A}$ can be written

$$|\mathbf{A}| = \begin{vmatrix} b_{11} & \cdots & b_{1p} \\ b_{21} - r_2 r_1 & \cdots & b_{2p} - r_2 r_p \\ \vdots & \ddots & \vdots \\ b_{p1} - r_p r_1 & \cdots & b_{pp} - r_p r_p \end{vmatrix} - \begin{vmatrix} r_1 r_1 & \cdots & r_1 r_p \\ b_{21} - r_2 r_1 & \cdots & b_{2p} - r_2 r_p \\ \vdots & \ddots & \vdots \\ b_{p1} - r_p r_1 & \cdots & b_{pp} - r_p r_p \end{vmatrix}$$

Expanding both determinants on the right side about the second row yields four determinants, the last one having the following form

$$\begin{vmatrix} r_1 r_1 & \cdots & r_1 r_p \\ r_2 r_1 & \cdots & r_2 r_p \\ b_{31} - r_3 r_1 & \cdots & b_{3p} - r_3 r_p \\ \vdots & \ddots & \vdots \\ b_{p1} - r_p r_1 & \cdots & b_{pp} - r_p r_p \end{vmatrix}$$

The first two rows in this determinant are linearly dependent, so the determinant is zero. Continuing to expand the determinants in the same manner yields

$$|\mathbf{A}| = \begin{vmatrix} b_{11} & \cdots & b_{1p} \\ b_{21} & \cdots & b_{2p} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pp} \end{vmatrix} - \begin{vmatrix} r_1 r_1 & \cdots & r_1 r_p \\ b_{21} & \cdots & b_{2p} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pp} \end{vmatrix} - \ldots - \begin{vmatrix} b_{11} & \cdots & b_{1p} \\ b_{21} & \cdots & b_{2p} \\ \vdots & \ddots & \vdots \\ r_p r_1 & \cdots & r_p r_p \end{vmatrix}$$

The first term is the determinant of **B**, and using the definition of the determinant as the sum of the *ith* row times the cofactors of the *ith* row, the above equation can be written

$$|\mathbf{A}| = |\mathbf{B}| - r_1 \sum_{j=1}^{p} r_j \mathbf{B}'_{1j} - \ldots - r_p \sum_{j=1}^{p} r_j \mathbf{B}'_{pj}$$

where

$$\mathbf{B}'_{ij} = (-1)^{i+j} |\mathbf{M}_{ij}|$$

and $\mathbf{M}_{ij}$ is the sub-matrix formed by deleting the *ith* row and *jth* column of **B** [11]. Using the definition of the cofactor matrix, the determinant of **A** can be written in matrix form

$$|\mathbf{A}| = |\mathbf{B}| - \mathbf{r}\mathbf{B}_{cof}^{T}\mathbf{r}^{T}$$

and the theorem is proved.

It is now possible to prove the result [13] presented in Eq. (14) which is given below in Theorem 2. This provides a rigorous relationship between the determinants of the FIM before and after a sensor location is removed from the candidate set.

THEOREM 2. Given $\text{FIM} = \Phi^T \Phi$ and $\text{FIM}_{-i} = \text{FIM} - \mathbf{r}^T \mathbf{r}$ where **r** is the row vector corresponding to the *ith* sensor location, then

$$|\text{FIM}_{-i}| = (1 - \text{EID}_i)|\text{FIM}|$$

PROOF: An application of Theorem 1 yields

$$\left|\mathrm{FIM}_{-i}\right| = \left|\mathrm{FIM}\right| - \mathbf{r}\,\mathrm{FIM}_{cof}^{T}\,\mathbf{r}^{T}$$

Since the FIM is a symmetric matrix, its cofactor matrix is symmetric also, so

$$\left|\mathrm{FIM}_{-i}\right| = \left|\mathrm{FIM}\right| - \mathbf{r}\,\mathrm{FIM}_{cof}\,\mathbf{r}^{T}$$

Using the definition of a matrix inverse [11], the above equation can be written

$$\begin{aligned}\left|\mathrm{FIM}_{-i}\right| &= \left|\mathrm{FIM}\right| - \left|\mathrm{FIM}\right|\mathbf{r}\,\mathrm{FIM}^{-1}\,\mathbf{r}^{T} \\ &= (1 - \mathbf{r}\,\mathrm{FIM}^{-1}\mathbf{r}^{T})\left|\mathrm{FIM}\right|\end{aligned}$$

From Eq. (12), the *ith* element of the EID is seen to be

$$\mathrm{EID}_i = \mathbf{r}\,\mathrm{FIM}^{-1}\mathbf{r}^{T}$$

therefore,

$$\left|\mathrm{FIM}_{-i}\right| = (1 - \mathrm{EID}_i)\left|\mathrm{FIM}\right|$$

which is the desired result.

The determinant of the FIM with the *ith* sensor location removed can be calculated using Theorem 2, thus yielding a simple way of optimizing this norm on the information matrix. Since the determinant of the FIM is a measure of the information from the sensors, then the EID is proportional to the amount of information lost by deleting a sensor. So, Theorem 2 shows that deleting the location with the smallest EID value yields the smallest change in the determinant as previously stated.

It will be proven shortly that the range of EID values is given by

$$0 \le \mathrm{EID}_i \le 1$$

However, it would be instructive to see what happens if a sensor has one of the extreme values of zero or one. A sensor location with a value of one must be retained in order to preserve the linear independence of the matrix Φ. If this

sensor location is deleted, then from Theorem 2 we have $\left|\mathrm{FIM}_{-i}\right| = 0$ and the matrix becomes singular. This means that all of the mode shape scale factors cannot be determined. If a candidate sensor location has an EID value of zero, then the determinant is unchanged and no loss of information occurs.

The fact that an element in the EID is non-negative is readily apparent from the definition in Eq. (11). All of the elements in the matrix **G** are positive because they are squares, and all eigenvalues λ_j are non-negative because the FIM is positive definite. The following proposition will show that an element of the EID is also less than one.

PROPOSITION: EID_i is in the range $0 \le \mathrm{EID}_i \le 1$

PROOF: Since **P** is an idempotent matrix, this implies that

$$p_{ii} = (\mathbf{PP})_{ii} = \sum_{j=1}^{n} p_{ij} p_{ji}$$

Since **P** is also symmetric, the diagonal elements can be written

$$p_{ii} = \sum_{j=1}^{n} p_{ij} p_{ji} = \sum_{j=1}^{n} p_{ij}^2$$

Expanding the sum on the right-hand side yields

$$p_{ii} = p_{ii}^2 + \sum_{i \neq j} p_{ij}^2$$

This equality can only be true if $p_{ii} \le p_{ii}^2$ which implies that

$$0 \le p_{ii} \le 1$$

The results in this section show that sensor locations can be ranked using the EID. Those that are closer to one are more important and should be retained to keep the mode shapes linearly independent. Sensor locations with values close to zero contribute less information and less to the linear independence of the mode shapes and can be deleted.

V. APPLICATIONS OF THE EID METHOD

A. NUMERICAL EXAMPLES

Two simple numerical examples are provided to illustrate how sensor locations are ranked by the EID. One example will show that a sensor location with an EID value of one must be retained to keep the columns of the mode shape matrix linearly independent. The other example will show that the EID tends to give a higher rank to those locations with a larger magnitude, if none of the locations are essential in keeping the problem non-singular (i.e., the columns of Φ linearly independent).

Suppose the following is a mode shape matrix

$$\Phi = \begin{bmatrix} 1 & 1 \\ 0 & 0.0001 \\ 2 & 2 \end{bmatrix} \tag{17}$$

and one sensor measurement must be removed. The EID values for each of the locations are

$$\text{EID}_1 = 0.2$$
$$\text{EID}_2 = 1.0$$
$$\text{EID}_3 = 0.8$$

Notice that the sum of the EID values is 2, which is the rank of the mode shape matrix. According to these numbers, the second row of the mode shape matrix must be kept so that the columns remain linearly independent. An examination of the matrix Φ reveals that to be true. If the second row is deleted, then the two columns are the same and the FIM formed with Eq. (17) is singular. Also note that the EID value for sensor 3 is larger than sensor 1. This means that sensor 1 can be deleted yielding the smallest change in the determinant of the FIM.

Now suppose that the mode shape matrix is

$$\Phi = \begin{bmatrix} 1 & 2 \\ 1 & 0.5 \\ -4 & -5 \end{bmatrix} \tag{18}$$

None of the rows of this matrix are needed to keep the columns linearly independent. The EID values in this case are

$$\mathrm{EID}_1 = 0.5556$$

$$\mathrm{EID}_2 = 0.5556$$

$$\mathrm{EID}_3 = 0.8889$$

One can see from these values that the EID ranks sensor 3 the highest and that either sensor 1 or 2 can be removed to yield the smallest change in the determinant of the FIM formed using Eq. (18). This example shows that if no row is critical to the independence of the target modes, then the rows with a larger magnitude tend to receive a higher value in the EID. Note, however, that in the first example the row with the smallest magnitude received the highest EID value because it was needed to keep the problem at full rank.

B. PLACING SENSORS ON A UNIFORM PLATE

To illustrate the problem of determining mode shape scale factors for a structure, a rectangular plate of uniform thickness will be used. This model of a uniform plate is a simple structure, making it easier to evaluate the effectiveness of using the EID values to choose optimal sensor locations. The EID will be used to choose 9 optimal sensor locations from 184 candidate locations. Results will show that the method described in Section III chooses optimal sensor locations in keeping with engineering judgment.

The rectangular plate extends over the domain $0 < x < a$ and $0 < y < b$, and the plate is simply supported. The free-decay response of a uniform plate to N concentrated unit impulse forces can be written as

$$w(x,y,t) = \sum_{k=1}^{N} \sum_{s=1}^{\infty} \sum_{m=1}^{\infty} \Phi(x_m, y_s) f(x_{mk}, y_{sk}, t) \tag{19}$$

where the mode shapes are given by

$$\Phi(x,y)=\frac{2}{\sqrt{\rho ab}}\sin\frac{m\pi x}{a}\sin\frac{s\pi y}{b} \tag{20}$$

and the forcing function is

$$f(x_{mk},y_{sk},t)=\frac{2}{\omega_{d_{ms}}\sqrt{\rho ab}}e^{-\xi_{ms}\omega_{ms}t}\sin\frac{m\pi x}{a}\sin\frac{s\pi y}{b}\sin\omega_{d_{ms}}t \tag{21}$$

The natural frequencies of the plate are given by

$$\omega_{ms}=\pi^2\sqrt{\frac{D_E}{\rho}\left[\left(\frac{m}{a}\right)^2+\left(\frac{s}{b}\right)^2\right]}$$

with

$$\omega_{d_{ms}}=\omega_{ms}\sqrt{1-\xi_{ms}^2}$$

denoting the damped natural frequency. The mass per unit area is given by ρ, and the plate stiffness is denoted by D_E. To simplify things further, the values for ρ and the plate stiffness will be set to one. The x and y coordinates in Eq. (20) refer to the sensor locations and those in Eq. (21) refer to the locations of the actuators (i.e., places where the concentrated impulse forces are located). For more information on the response equation given in Eq. (19), any graduate text on vibration theory can be consulted [14].

The mode shape matrix Φ must be calculated to model the plate. This can then be used with the EID technique to optimally locate sensors for a vibration test. The normal modes of a simply supported plate are given by Eq. (20). In order to construct the mode shape matrix, maximum values are chosen for the frequency parameters m and s. These values determine the number of target modes or columns in the mode shape matrix. For example, maximum values of $m = s = 3$ yield $p = 9$ target modes. The next step is to choose x and y coordinate pairs for candidate sensor locations. These coordinate pairs are used in Eq. (20) along with the frequency indices to evaluate each element of the matrix Φ.

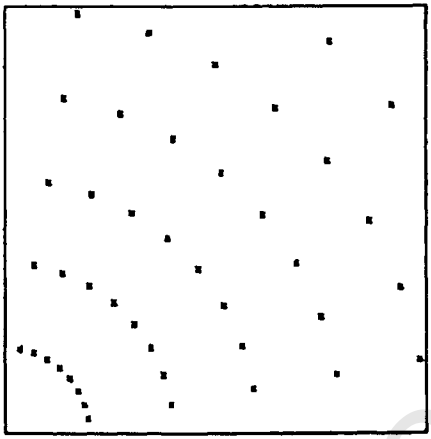


Fig. 1. A set of 42 candidate sensor locations on a square plate.

For the application presented here, the initial sensor locations were found using the following method. The rectangular plate is divided into angular and radial distances, and the x and y coordinate pairs are found using

$$x = r\cos\theta$$
$$y = r\sin\theta$$

An initial set of 42 sensor locations is shown in Fig. 1. An application of the EID technique yields the sensor locations illustrated in Fig. 2. The parameters used for this example are given in Table I. This example shows that deleting the sensor location with the smallest EID value chooses sensor locations in keeping with engineering judgment. The sensor locations are evenly spaced on the interior of the plate, which makes sense for a vibration test of this type of structure.

This is a simple example,[3] but it is illustrative of the power of the method. In practical use, an analyst would not need to use a technique such as the one described here for simple structures. However, for large, complicated structures with many target modes engineering judgment alone is not very

[3] For an example of the EID technique applied to a more complicated structure where the finite element model is used, see Kammer [9] and Poston [10].

Table I. Parameters Used in Plate Application

Parameter	Value
Maximum m	3
Maximum s	3
a	1.001
b	1.0
Number of Candidates	184
Number of Final Locations	9

effective. It should also be noted that taking measurements is a costly process, so the sensors should be placed correctly and in the most effective manner possible. Relying solely on the expertise of an engineer or analyst can be risky and subjective. A combination of engineering expertise and the EID technique should produce sensor locations that will yield the maximum amount of information and result in a meaningful test. For example, an initial set of possible sensor locations is determined, and final sensor locations are chosen using the EID technique. At this point, the engineer can use his experience to decide whether the sensor locations make sense. This can be repeated using different starting locations until the engineer is satisfied with the results.

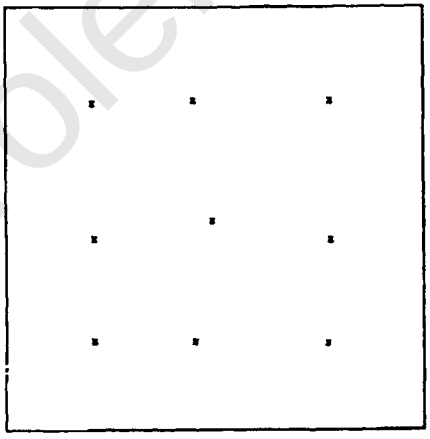


Fig. 2. Final sensor configuration chosen using the EID technique from 184 possible sites.

VI. SUMMARY AND OTHER APPLICATIONS

The EID distribution can be used in applications other than the one presented here. Since it is derived from a least squares problem, it is applicable to any situation that can be represented in this context. For example, a series of experiments are performed and measurements are taken. Regression analysis [4,5,11] is used to fit equations to the values observed in the experiments. This is a least squares problem under certain assumptions with n observations and p response variables. If, for some reason, all of the observations cannot be processed, then the EID technique can be used to optimally choose the best observations for analysis.

Another application for this method is Kalman filtering [4] and other state space representations of random processes. In this system, the state summarizes all of the information from the past that is needed to predict future states. The state space representation of a discrete-time linear system is described by two equations. One is a system equation that pertains to the evolution of the state, and the other is a measurement equation that denotes the observations obtained from a given state [4,15,16]. The measurement equation is

$$\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \varepsilon(k)$$

where k denotes a time step, $\mathbf{y}$ is an observation vector, $\mathbf{x}$ is an unobservable state vector that summarizes the state of the system at time step k, $\mathbf{C}$ is a known matrix, and ε is a white-noise process as described in Eqs (2) and (3). This is in the same form as Eq. (1) and the EID method can be used in this context also, where $\mathbf{C}$ takes on the role of the mode shape matrix.

A linear system is observable [15,16] at time step k_0 if $\mathbf{x}(k_0)$ can be determined from the output function or sequence $\mathbf{y}(k_0, k_1)$ for $k_0 \le k_1$, where k_1 is a finite time step. A system is completely observable if this is true for all k_0 and $\mathbf{x}(k_0)$. If the system is not completely observable, then the initial state cannot be determined from the output. Observability was originally defined by Kalman and Bucy [17,18,19] for linear, lumped parameter systems, and it can be seen from the definition given above that it indicates the ability to recover the prior state of a system based on observations of the state over a period of time. Therefore, observability should be a major consideration when determining where measurements should be taken on a system.

It was shown in Section IV using Theorem 2 that a sensor location with an EID value of 1 must be retained to keep the problem non-singular and of full rank. If the matrix C is not of rank p, then the system will not be observable because the prior state is unrecoverable. Therefore, it can also be said that using the EID method ensures the observability of the system.

A deterministic method for optimally locating sensors for the purposes of model verification has been presented. This method optimizes the determinant of the FIM as a means of maximizing the information obtained from the sensors. The FIM is determined from the finite element model of a complicated structure or analytically for a simple structure. The EID distribution can be calculated and used to determine a subset of measurement locations from a large candidate set of locations. The candidate sensor with the smallest EID value is deleted, the EID is calculated again, and another location is deleted. This continues in an iterative manner until the desired subset of locations is chosen. It is possible to delete more than one candidate location at each calculation of the EID. However, for optimal results only one should be deleted at any iteration.

Given the computer resources available today, this is not a computationally intensive algorithm. Thus, it is feasible to use this technique for large, complicated systems and still delete only one location at each iteration. The final configuration is dependent on the initial starting point, because the EID technique is a greedy algorithm and is not guaranteed to find the global optimum. This technique could be used several times with different sets of candidate locations. Engineering judgment or other modal analysis [10] would then be used to choose the final sensor locations.

VII. SYMBOLS AND ACRONYMS

λ	Eigenvalue of the FIM
$\boldsymbol{\varepsilon}$	Vector of noise in measurements
$\mathbf{h}$	Vector containing sensor output
$\mathbf{q}$	Vector of unobservable scale factors
$\hat{\mathbf{q}}$	Vector of estimated scale factors
$\mathbf{r}$	Row vector
COV	Covariance matrix corresponding to $\hat{\mathbf{q}}$

EID	Effective Independence Distribution; vector of values for each sensor location
FIM	Fisher Information Matrix
P	Projection matrix, diagonal elements are the EID
Φ	Matrix with each column corresponding to a mode shape
Λ	Diagonal matrix of eigenvalues
Ψ	Orthonormal matrix with each column corresponding to an eigenvector

VIII. REFERENCES

1. K. H. Huebner and E. A. Thornton, *The Finite Element Method for Engineers*, John Wiley & Sons, New York (1982).

2. C. S. Kubrusly and H. Malebranche, "Sensors and Controllers Location in Distributed Systems - A Survey," *Automatica* **21**, p 117 (1985).

3. L. Meirovitch, *Elements of Vibration Analysis,* McGraw Hill, New York, (1986).

4. B. Abraham and J. Ledolter, *Statistical Methods for Forecasting,* John Wiley & Sons, New York (1983).

5. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York (1987).

6. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C,* Cambridge Univ Press, New York (1992).

7. F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, New York (1985).

8. R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, New York (1987).

9. D. C. Kammer, "Sensor Placement for On-Orbit Modal Identification and Correlation of Large Space Structures," *Journal of Guidance, Control, and Dynamics* **14**, p 251 (1991).

10. W. L. Poston, *Optimal Sensor Locations for On-Orbit Modal Identification of Large Space Structures,* Master's Thesis, George Washington Univ, (1991).

11. G. Strang, *Linear Algebra and Its Applications,* Harcourt Brace Jovanovich, San Diego (1988).

12. A. Ben-Israel and T. Greville, *Generalized Inverses: Theory & Applications,* John Wiley & Sons, New York (1974).

13. W. L. Poston and R. H. Tolson, "Maximizing the Determinant of the Information Matrix with the Effective Independence Distribution Method," *Journal of Guidance, Control, and Dynamics* **15**, p 1513 (1992).

14. L. Meirovitch, *Analytical Methods in Vibrations,* MacMillan, New York (1967).

15. C. T. Chen, *Introduction to Linear System Theory,* Holt Rinehart and Winston, New York (1970).

16. L. W. Brogan, *Modern Control Theory*, Prentice-Hall, New Jersey (1982).

17. T. K. Yu and J. H. Seinfeld, "Observability and Optimal Measurement Location in Linear Distributed Parameter Systems," *Int J Control* **18**, p 785 (1973).

18. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Eng.* **82**, p 35 (1960).

19. R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," *J. Basic Eng.* **83**, p 95 (1961).

This Page Intentionally Left Blank

DISCRETE TIME CONSTRAINED LINEAR SYSTEMS

Jean-Claude Hennet

Laboratoire d'Automatique et d'Analyse des Systèmes du C.N.R.S.,
7, avenue du Colonel Roche, 31077 Toulouse FRANCE

I. INTRODUCTION

The existence of hard constraints on state and control variables has often generated problems in practical implementation of control laws. As pointed out by E.G.Gilbert [1], modern multivariable control designs has increased the risks of performance degradation and of failure, due to unmodelled phenomena such as the existence of practical bounds on control variables and of physical limitations on state variables. Integration of such constraints in the formulation of an optimal control problem is generally possible and may lead to tractable solutions, as for instance in Model Predictive Control [2], [3], [4], and in Linear Quadratic Control [5], [6]. However, such control schemes generally imply considerable off-line computations and are not always robust enough. In practice, anti-windup schemes [7] can often compensate for the undesired evolutions due to actuators saturations, but their properties are unperfectly analyzed. Also, it is often interesting to avoid state or input saturations as much as possible, and thus to generate controlled trajectories staying in the interior of the constrained domain, without reaching the border of this domain. It is only recently that some attempts have been made to analytically integrate constraints within the design of feedback control laws. The basic mathematical tools for such a direct integration are mainly the positive invariance concept and its analytical characterizations.

Several algorithms are now available to obtain the positive invariance of the constrained domain or of a domain included in the constrained domain by construction of linear state feedback regulators. Some extensions of this ap-

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

proach to dynamic output feedback have also been recently studied in the literature [8]. Two basic methods have been developed for constructing positively invariant regulators: one by Linear Programming, and the other one by eigenstructure assignment. The advantages of the first method are in its greater generality and in the possibility of easily integrating other constraints, such as performance and robustness requirements in the optimization program [9]. On the other hand, the major advantages of the second method is the simplicity of its implementation, and the insights it gives into the structural and spectral properties of the system.

II. POSITIVE INVARIANCE RELATIONS FOR LINEAR SYSTEMS

A. Positively invariant domains

Definition II.1 : Positive invariance
Positive invariance is a property characterizing some function of time generated by a dynamical system : any trajectory of this function starting in a region of space always remains in that region along the system evolution.

In particular, consider a discrete-time multivariable system represented by a state-space equation of the following type (1):

$$x_{k+1} = f(x_k, w_k) \text{ for } k = 0, 1, .. \tag{1}$$

with $x_0 \in \Re^n$, $w_k \in \mathcal{W} \subset \Re^p$ and $f(.)$ a mapping from $\Re^n \times \mathcal{W}$ onto $\Re^n$.

Let Ω be a subset of $\Re^n$. Within the framework of set-theory , Ω is said to be a positively invariant domain of system (1) if and only if :

$$f(\Omega, \mathcal{W}) \subset \Omega.$$

The well-known "geometric approach" (W.M.Wonham [10]) was initially based on the properties of invariance of some subspaces for linear systems (A-invariance) and on the use of these properties for the design of feedback control laws ((A,B)-invariance). Extension of this approach to some classes of non-linear systems also proved highly successful. The natural extension of the notion of invariant linear subspace to the case of smooth non-linear systems is the property of invariance of affine distributions [11].

In view of the application of invariance properties to linear constrained control problems, two classes of positively invariant domains of the state space have recently been explored: polyhedral cones, as studied by C.Burgat et al. ([12], [13]), and polyhedral domains including the origin point in the works of G.Bitsoris et al. [14], [15], and J.C.Hennet et al. [16], [17].

B. Positive invariance of polyhedral domains of the state space

Definition II.2 : Polyhedral Set [18]
Any polyhedral set of $\Re^n$ can be characterized by a matrix $Q \in \Re^{r*n}$ and a vector $\phi \in \Re^r$, r and n being positive integers. It is defined by:

$$R[Q, \phi] = \{x \in \Re^n ; Q.x \leq \phi\}. \tag{2}$$

Definition II.3 : Polyhedral Cone [18]
A polyhedral cone of $\Re^n$ is characterized by a matrix $Q \in \Re^{r*n}$, with r a positive integer. It is defined by: $R[Q, 0_r] = \{x \in \Re^n ; Q.x \leq 0_r\}$, where, by convention, 0_r denotes the null vector of $\Re^r$. From this definition, it is clear that polyhedral cones are a particular class of polyhedral sets.

Definition II.4 : Simplicial Polyhedral Cone [19]
For $Q \in \Re^{n \times n}$ and non-singular, the polyhedral cone $R[Q, 0_n]$ is called a simplicial polyhedral cone. It is a proper cone with exactly n extremal rays.

Definition II.5 : Simplicial Proper Polyhedra
If rank $Q = n$ and if ϕ is a vector of $\Re^n$ with strictly positive components, then, the polyhedral set $R[Q, \phi]$ is a simplicial polyhedron with the origin as an interior point. It is called here a simplicial proper polyhedron.

Consider a discrete-time linear system described by the state equation:

$$x_{k+1} = A_0 x_k \tag{3}$$

This equation may represent an autonomous system or a system controlled by state feedback. In the latter case, matrix A_0 can be decomposed as $A_0 = A + BF$, with A the open-loop state matrix, B the control matrix and F the gain matrix. The purpose of chapters IV and V will mainly be to give algorithms to construct F so as to obtain the desired positive invariance properties.

Definition II.6 : Positive invariance of polyhedral sets
A polyhedral set $R[Q, \phi]$ is a positively invariant set of system (3) if and only if : $x_k \in R[Q, \phi] \Longrightarrow x_{k+p} \, (= A_0^p x_k) \in R[Q, \phi]$, $\forall p \in \mathcal{N}, \forall k \in \mathcal{N}$.

C. Basic invariance property

Lemma 1
A necessary and sufficient condition for $R[Q, \phi]$ to be positively invariant for system (3) is :

$$QA_0 x \leq \phi \ , \ \forall x \in \Re^n ; Qx \leq \phi. \tag{4}$$

This property directly follows the foregoing definition. From this lemma, positive invariance of polyhedral domains in the state space for system (3) can be analyzed as a special case of inclusion of polyhedral domains and characterized using the extended Farkas' lemma presented in J.C.Hennet 1989 [20].

D. The Extended Farkas' Lemma

The set of linear inequalities

$$Q.x \leq \phi \text{ with } Q \in \Re^{r*n} \text{ and } \phi \in \Re^{r} \tag{5}$$

defines a convex polyhedral set of $\Re^n$, denoted $R[Q, \phi]$. Under what conditions any point of $R[Q, \phi]$ also belongs to an other convex polyhedral set, $R[P, \psi]$ defined by the set of linear inequalities:

$$Px \leq \psi \text{ with } P \in \Re^{p*n} \text{ and } \psi \in \Re^{p} \tag{6}$$

An extension of Farkas' lemma to the non-homogeneous matrix case provides a set of necessary and sufficient conditions on Q, P, ϕ, ψ under which:

$$R[Q, \phi] \subseteq R[P, \psi]. \tag{7}$$

Lemma II.2
The system $Px \leq \psi$ is satisfied by any point of the non-empty convex polyhedral set defined by the system $Qx \leq \phi$ if and only if there exists a (dual) matrix U of $\Re^{p*r}$ with non-negative coefficients satisfying conditions :

$$UQ = P \tag{8}$$
$$U.\phi \leq \psi. \tag{9}$$

Proof
This Lemma can easily be proven by concatenation of necessary and sufficient conditions related to each row P_i of matrix P. For $i = 1, .., p$, consider the row vector P_i of matrix P and its associated component ϕ_i of vector ϕ. Condition (7) is equivalent to the joint p conditions (C_i) defined by:

$$(C_i) \quad P_i x \leq \psi_i \forall x \ ; \ Qx \leq \phi.$$

From Farkas' Lemma [18], under the assumption $R(Q, \phi) \neq \emptyset$,

$$(C_i) \iff \begin{cases} \exists U_i \ ; & {U_i}^T \in \Re^q \\ U_{ij} \geq 0 & \forall j = 1, ..., q \\ U_i.Q = P_i & \\ U_i.\phi \leq \psi_i & \end{cases} \tag{10}$$

Then, joint satisfaction of p conditions (C_i) for $i = 1, ..., p$ is equivalent to the joint existence of p row vectors U_i satisfying (10). Therefore, as stated in the theorem, condition (7) is equivalent to the existence of a matrix U having the U_i as row vectors. It clearly satisfies (8) and (9).
□

It is worth pointing out that the derivation of this property by the theory of duality in linear programming does not require any particular assumption on the rank of Q and components of vector ϕ.

As for the classical Farkas' Lemma, many different versions of the extended Farkas' Lemma can be obtained for different values of ϕ and ψ and for equalities instead of inequalities in (5) and/or (6). A particularly interesting version of this Lemma is obtained when replacing (5) and (6) by equalities with null right-hand terms.

Lemma II.3
The system $Px = 0_p$ is satisfied by any point of the non-empty convex polyhedral set defined by the system $Qx = 0_r$ if and only if there exists a (dual) matrix U of $\Re^{p*r}$ satisfying condition :

$$UQ = P \tag{11}$$

The proof is similar to the proof of Lemma II.2, with the usual dual correspondences.

Note that relation (8) with U non-negative is a particular instance of relation (11). Therefore, the *primal* condition (7) implies the *primal* implication:

$$Px = 0_p \Longrightarrow Qx = 0_r. \tag{12}$$

E. Invariance relations

The following Proposition and its Corollary are direct consequences of Lemma II.2.

Proposition II.1
Positive invariance of the polyhedral set $R[Q, \phi]$ for system (3) is equivalent to the following properties:

$$\exists K = ((K_{ij})) \in \Re^{r*r}, K_{ij} \geq 0, \ \forall i = 1, ..., r, \forall j = 1, ..., r \tag{13}$$

$$KQ = QA_0 \tag{14}$$

$$K\phi \leq \phi \tag{15}$$

Corollary II.1

Positive invariance of the polyhedral cone $R[Q, 0_r]$ *for system (3) is equivalent to the existence of a non-negative matrix :*
$K = ((K_{ij})) \in \Re^{r*r}$, $K_{ij} \geq 0$, $\forall i = 1, ..., r$, $\forall j = 1, ..., r$, *such that :*

$$KQ = QA_0. \tag{16}$$

Some elementary results can be derived from the conditions stated in Proposition II.1.

F Homothesis property

Note that if $R[Q, \phi]$ is an invariant set of system (3), any domain homothetic to $R[Q, \phi]$, $R[Q, e\phi]$ with $e > 0$ is also invariant. Indeed, in Proposition II.1, the only relation in which ϕ appears is (15) and for any strictly positive value of e, this inequality is equivalent to:

$$Ke\phi \leq e\phi \tag{17}$$

G. Stability properties

1. Positive invariance of a polyhedral cone

Positive invariance of a polyhedral cone $R[Q, 0_r]$ with respect to system (3) is simply characterized by (16) with matrix K non-negative. Matrix K being non negative, its spectral radius, ρ_K, is an eigenvalue of K and, from the Perron-Frobenius Lemma [19], an associated eigenvector of K, denoted v_ρ, is non-negative and non-null. Consider the projected system :

$$y_{k+1} = Ky_k \text{ with } y_k = Qx_k \tag{18}$$

Asymptotic stability of system (18) is equivalent to $\rho(K) < 1$. Then, from (16) and

$$Kv_\rho = \rho_k v_\rho \leq v_\rho, \tag{19}$$

$R[Q, v_\rho]$ is a positively invariant polyhedron for system (3).

Lemma II.4

Under the assumption of positive invariance of $R[Q, 0_r]$ *with respect to system (3), asymptotic stability of (18) implies the existence of a positively invariant polyhedron* $R[Q, \chi]$ *for system (3), with* χ *a non-negative and non-null vector in* $\Re^r$.

Conversely, under the assumption of positive invariance of the polyhedral cone $R[Q, 0_r]$, the following matrix can then be constructed :

$$Z = I_{r\times r} - K \tag{20}$$

where $I_{r\times r}$ is the unity matrix of $\Re^r$. All the off-diagonal terms of Z are non-positive. Then, from a classical result on matrices ([19], [21], [22]), system (18) is asymptotically stable if matrix Z is a non-singular M-matrix. An equivalent characterization of Z is the existence of a strictly positive vector,

$$\chi \in \Re^r \text{ with } \chi_i > 0 \quad \forall i = 1, ..., r, \text{ such that } Z\chi > 0_r.$$

But this property can be written :

$$K\chi < \chi \tag{21}$$

This relation, together with (16) then shows the positive invariance of the polyhedral set $R[Q, \chi]$. The following result can then be formulated :

Proposition II.2
If system (3) admits a polyhedral cone $R[Q, 0_r]$ as a positively invariant set, the projection of (3) on Range(Q^T) is asymptotically stable if there exists a polyhedron $R[Q, \chi]$ with $\chi > 0_r$ which is positively invariant (with strict contractivity (21)) with respect to (3).

2. Contractive invariance of a simplicial proper polyhedron

Lemma II.5
If rank $Q = n$ and $R[Q, \phi]$ is a simplicial proper polyhedron (vector ϕ has strictly positive components), then, contractive invariance of $R[Q, \phi]$ for system (3) implies asymptotic stability of system (3) .

Note that positive invariance of $R[Q, \phi]$ implies the existence of a non-negative matrix $K \in \Re^{n\times n}$ such that : $KQ = QA_0$. And this condition is equivalent to positive invariance of the simplicial cone $R[Q, 0_n]$. Matrices K and A_0 being similar, they have the same spectrum. Then, Lemma II.4 directly derives from Proposition II.2.

H. Symmetrical and non-symmetrical invariant domains

Polyhedral domains which are symmetrical with respect to the origin point are of special importance for the analysis of invariance sets of a linear system. They are also well adapted to the regulator design problem in the current

case of symmetrical constraints. A specialized version of invariance relations (13),(14),(15) can be formulated as follows :

Proposition II.3

A necessary and sufficient condition for the symmetrical polyhedral set

$$S(G,\omega) = \{x \in \Re^n; -\omega \leq Gx \leq \omega\} \quad \textit{with } G \in \Re^{s*n}, \omega \in \Re^s{}_+$$

to be invariant for system (3) is the existence of a matrix $H \in \Re^{s*s}$ *such that:*

$$HG = GA_0 \tag{22}$$

$$|H|\omega \leq \omega \text{ with } |H| = ((|H_{ij}|)) \tag{23}$$

Proof

This Proposition can be derived from Proposition II.1 using the following analysis. Symmetrical inequalities: $-\omega \leq Gx \leq \omega$ with $G \in \Re^{s*n}$, $\omega \in \Re^s{}_+$, can be re-written as

$$\begin{pmatrix} G \\ -G \end{pmatrix} x \leq \begin{pmatrix} \omega \\ \omega \end{pmatrix} \tag{24}$$

Application of positive invariance relations (13),(14),(15) to formulation (24) can be expressed as follows:

$$\begin{aligned} \exists H^+ \in \Re^{s*s}, (H^+)_{ij} &\geq 0 \; \forall i = 1,..,s, \forall j = 1,..,s \\ \exists H^- \in \Re^{s*s}, (H^-)_{ij} &\geq 0 \; \forall i = 1,..,s, \forall j = 1,..,s \\ (H^+ - H^-)G &= GA_0 \\ (H^+ + H^-)\begin{pmatrix} \omega \\ \omega \end{pmatrix} &\leq \begin{pmatrix} \omega \\ \omega \end{pmatrix} \end{aligned}$$

And , by setting $H = H^+ - H^-$, the preceding relations can be re-written as stated in Proposition II.3 (G. Bitsoris [14]).

To show the sufficiency of the conditions, take for instance

$$H^+{}_{ij} = \max(H_{ij}, 0) \text{ and } H^-{}_{ij} = \max(-H_{ij}, 0).$$

□

Similarly, positive invariance w.r.t. system (3) of non-symmetrical domains of the following form:

$$S(G,\omega^-,\omega^+) = \{x \in \Re^n \text{ such that } -\omega^- \leq Gx \leq \omega^+\}$$

with $\omega^+ \in \Re^s{}_+$, $\omega^- \in \Re^s{}_+$ and rank $G = s$ is equivalent to the existence of two non-negative matrices of $\Re^{s*s}$, H^+ and H^- such that [23], [16] :

$$(H^+ - H^-)G = GA_0 \tag{25}$$

$$\begin{pmatrix} H^+ & H^- \\ H^- & H^+ \end{pmatrix} \begin{pmatrix} \omega^+ \\ \omega^- \end{pmatrix} \leq \begin{pmatrix} \omega^+ \\ \omega^- \end{pmatrix}. \tag{26}$$

I. Invariance of the subspace $\mathcal{K}er\ G$

Proposition II.4

A_0-Invariance of the subspace $\mathcal{K}er\ G$ is a necessary condition for positive invariance, with respect to system (3), of any of the non-empty sets $R[G,\omega]$, $S(G,\omega)$, $S(G,\omega^-,\omega^+)$, for $G \in \Re^{s\times n}$, rank $G = s$, $\omega \in \Re^s{}_+$, $\omega^- \in \Re^{s+}$, $\omega^+ \in \Re^{s+}$.

Proof

The first relation (22) characterizing the positive invariance property of the non-empty symmetrical polyhedron $S(G,\omega)$ is purely structural. It is a canonical projection equation (W.M. Wonham [10]). From Lemma II.3, the existence of a matrix H satisfying (22) is necessary and sufficient for the following implication to hold true : $Gx = 0_s \implies GA_0x = 0_s$. Such an implication characterizes the invariance of the subspace $\mathcal{K}er\ G$ with respect to system (3).

Note that equation (25) obtained in the case of a non-symmetrical domain $S(G,\omega^+,\omega^-)$ is equivalent to (22). Also, relation (16) applied to $R[G,\omega]$ is a particular case of (22) with matrix H non-negative. In this last case, positivity of the components of vector ω is not required.

□

J. Invariant domains of similar systems

Relation (22) defines similarity relationships between cyclic subspaces of A_0 and K. A possible way to construct invariant domains for system (3) can then proceed from the design of invariant domains for the "similar" system:

$$y_{k+1} = Uy_k \tag{27}$$

with $T \in \Re^{n*n}$, rank $T = n$ and

$$U = T^{-1}A_0T \tag{28}$$

Lemma II.6

A necessary and sufficient condition for system (3) to admit $R[Q,\phi]$ as a positively invariant polyhedron is that system (27) admits as a positively invariant domain $R[Q',\phi]$, with $Q' = QT$.

Proof

The proof is elementary. From Lemma II.1, positive invariance of $R[Q,\phi]$ is characterized by :

$$QA_0x \leq \phi \quad \forall\ x \text{ such that } Qx \leq \phi.$$

The change of variable $x = Ty$ implies $QA_0Ty \leq \phi$ and with $A_0T = TU$, $QTUy \leq \phi$ for all y such that $QTy \leq \phi$.

Conversely , by the similarity relationship and the change of variable $y = T^{-1}x$, positive invariance of $R[Q', \phi]$ for system (27) implies positive invariance of $R[Q'T^{-1}, \phi]$ for system (3).

Note

This lemma is also valid for symmetrical polyhedral sets. In the statement of the lemma, the domains $R[Q, \phi]$, $R[Q', \phi]$ can be replaced by $S(G, \omega)$ and $S(G', \omega)$ respectively, with $G' = GT$..

Lemma II.7

Existence of a positive vector of $\Re^n$, ψ, *such that* $|U|.\psi \leq \psi$ *is a necessary and sufficient condition for system (3) to admit as an invariant set a symmetrical n-parallelotope with facets parallel to the hyperplanes of the system of column vectors of matrix* T.

Proof

The similarity relationship, (28) can be re-written:

$$UT^{-1} = T^{-1}A_0 \tag{29}$$

If it is possible to find a strictly positive vector,$\psi \in \Re^n$ such that :

$$|U|\psi \leq \psi \tag{30}$$

then, from the symmetrical invariance relations (22),(23) of Proposition II.3, the domain :

$$S(T^{-1}, \psi) = \{x \in \Re^n; -\psi \leq T^{-1}x \leq \psi\}$$

is positively invariant for system (3). This polytope is symmetrical with respect to the origin. It has 2n facets. In an orthonormal basis, the symmetrical hyper-rectangle $S(\mathcal{I}_{n*n}, \psi)$ is positively invariant w.r.t. system $y_{k+1} = Uy_k$. Therefore, its image by T^{-1} is a symmetrical n-parallelotope (A. Brøndsted [24]).

Conversely, if system (3) admits as a positively invariant domain a symmetrical n-parallelotope $S(V, \psi)$, with $\psi \in \Re^n{}_+$, $V \in \Re^{n*n}$, rank $V = n$ and $U = VA_0V^{-1}$, then, by Lemma II.5, $S(\mathcal{I}_{n\times n}, \psi)$ is a positively invariant set of system $y_{k+1} = Uy_k$ and therefore $|U|\psi \leq \psi$.

□

III. POSITIVELY INVARIANT DOMAINS OF LINEAR SYSTEMS

A. Existence of Polyhedral Positively Invariant Domains for a Stable System

From a classical result by R.E.Kalman et al., 1960 [25], any stable linear system admits elliptic positively invariant domains associated with its quadratic Lyapunov functions. Therefore, if A_0 is a stable matrix, there exists a positive definite mapping from $\Re^n$ onto $\Re^+$ such that

$$v(x) = x^T P x \text{ satisfies } \Delta v(x) = v(A_0 x) - v(x) \leq 0 \ \forall x \in \Re^n.$$

Then any ellipsoid

$$v(x) \leq \mu, \text{ with } \mu \in \Re^+, \ \mu \neq 0, \tag{31}$$

is a positively invariant domain with respect to the linear system (3).

Similarly, the linear system (3) may admit as a Lyapunov function a generalized L_∞-norm function, called a polyhedral norm, of the following type:

$$v(x) = \max_i \frac{|(Gx)_i|}{\omega_i} \text{ with } G \in \Re^{s*n}, \ \text{rank} G = n, \text{ and } \omega \in \Re^s, \omega_i > 0 \tag{32}$$

In this case, $\Delta v(x) = v(Ax) - v(x)$ is negative semi-definite, and any convex polytope defined by

$$Gx \leq \nu\omega \text{ with } \nu \in \Re^+ \ , \nu \neq 0$$

is a positively invariant domain of system (3).

If the spectral radius of A_0 is less than 1, from the existence of elliptic positively invariant domains of the type (31) and from the classical property of equivalence between norms (see e.g. [26]), that there exists a polyhedral norm such that the unit ball for this norm (a symmetrical convex polytope) is positively invariant, i.e. the operator A_0 is contracting in this norm.
A basic consequence of this property is that any asymptotically stable linear system admits non-quadratic Lyapunov functions of the afore mentioned type (32), introduced by H.N.Rosenbrock [27].
The following existence proposition can thus be stated :

Proposition III.1 : *Asymptotic stability of system (3) implies that this system admits as positively invariants sets some closed and bounded symmetrical polytopes* $S(G,\omega)$, *with* $G \in \Re^{s*n}$, $rank G = n$, *and* $\omega \in \Re^s, \omega_i > 0$. *Each of*

these polytopes is associated with a polyhedral Lyapunov function (32) of system (3).

The main purpose of Chapter III is to construct such positively invariant polytopes using the information on the spectrum of A_0, and not only its spectral radius. Some of the results presented in this chapter can also be found in J.C Hennet et al., 1993 [28].

B. The Jordan decomposition of the problem

The construction technique which will now be described uses the real Jordan representation of matrix A_0 . It distinguishes the types of invariant domains associated with each diagonal block depending on the location of the associated eigenvalues.

Consider the Jordan decomposition in $\Re^n$ of matrix A_0 :

$$\tilde{A}_0 = P^{-1} A_0 P \tag{33}$$

with:

$$(\tilde{A}_0) = \begin{bmatrix} L_1 & 0 & . & & & . & 0 \\ 0 & . & & & & & . \\ . & 0 & L_{p1} & 0 & & & \\ & & & D_1 & 0 & . & \\ & & & 0 & & D_{p2} & 0 \\ & & & & & 0 & \Delta_1 \\ . & & & & & & 0 \\ 0 & . & & & & . & 0 \end{bmatrix} \quad \begin{matrix} & \\ & \\ & \\ & \\ & \\ . & \\ & \\ \Delta_{p3} & \end{matrix} \tag{34}$$

Matrix A_0 is assumed to have all its eigenvalues strictly inside the unit disk of the complex plane. The space $\Re^n$ can be decomposed into p cyclic independent eigensubspaces associated with the p Jordan blocks of $\tilde{A}_0$,with: $p = p_1 + p_2 + p_3$.

From lemma II.5, existence of a positively invariant polytope for system (3) is equivalent to the existence of a positively invariant polytope for system

$$y_{k+1} = \tilde{A}_0 y_k \text{ with } y_k = P^{-1} x_k. \tag{35}$$

Then, if it is possible to construct a positively invariant polytope $S(G', \omega)$ for system (35), the domain $S(G'P, \omega)$ is a positively invariant polytope of the initial system (3). The proposed technique then consists of constructing a positively invariant polytope in each cyclic subspace associated with a Jordan block of $\tilde{A}_0$, and of constructing a global positively invariant polytope of (35) from its positively invariant projections.

1. Blocks associated with real eigenvalues

The blocks $L_1, \ldots, L_{p1}$ are associated with the p_1 (not necessarily distinct) real eigenvalues of A_0. Each such block L_i concerns the eigenvalue λ_i ($|\lambda_i| < 1$) with the order of multiplicity q_i in this block.

$$L_i = \begin{bmatrix} \lambda_i & 0 & . & & . & 0 \\ 1 & . & & & & . \\ 0 & . & . & & & \\ . & & . & . & & . \\ . & & & . & . & 0 \\ 0 & . & . & 0 & 1 & \lambda_i \end{bmatrix}$$

Under condition $|\lambda_i| < 1$, it is possible to construct a strictly positive vector l_i, such that the symmetrical polyhedral set $S(I_{q_i \times q_i}, l_i)$ is positively invariant for the subsystem:

$$z_{k+1} = L_i z_k \tag{36}$$

Such a vector l_i can be constructed as follows:

- If the order of multiplicity of λ_i in the block L_i is 1, l_i can be any positive number. In this case, $z_k \in \Re$, and under the assumption $|\lambda_i| \leq 1$,

$$-l_i \leq z_k \leq l_i \Longrightarrow -l_i \leq \lambda_i z_k \leq l_i$$

- If $q_i > 1$, under the assumption $|\lambda_i| < 1$, define the strictly positive number $\epsilon_i = 1 - |\lambda_i|$ and set:

$$l_i{}^T = (l_{i1}, \ldots, l_{iq_i})$$

such that:
l_{i1} is any positive real number
the positive numbers $l_{i2}, \ldots, l_{iq_i}$ should satisfy the relations:

$$l_{i2} \geq \frac{l_{i1}}{\epsilon_i}, \ldots, l_{iq_i} \geq \frac{l_{iq_i-1}}{\epsilon_i} \tag{37}$$

Inequality (38) is then obviously verified:

$$|L_i|.l_i < l_i \text{ for } i = 1, \ldots, p_1 \tag{38}$$

Then, by Proposition II.3, $S(I_{q_i \times q_i}, l_i)$ is a positively invariant symmetrical polytope of system (36).

Remark
The same result applies for systems stable in the sense of Lyapunov having -1 or (and) $+1$ as eigenvalues.
This result obviously derives from the fact that stability of the system implies that each eigenvalue -1 or $+1$ is necessarily simple in each Jordan block associated with it.

2. Blocks associated with simple complex eigenvalues

The blocks D_m $(m = 1, \ldots, p_2)$ correspond to couples of simple complex eigenvalues: $\begin{cases} \rho_m[\cos(\beta_m) + j\sin(\beta_m)] \\ \rho_m[\cos(\beta_m) - j\sin(\beta_m)] \end{cases}$, with order of multiplicity 1 in this block, and such that $\rho_m^2 < 1$, $\rho_m > 0$, $0 \leq \beta_m \leq 2\pi$.

The real Jordan block associated with such a conjugated pair of eigenvalues is :

$$D = \begin{pmatrix} \rho\cos(\beta) & \rho\sin(\beta) \\ -\rho\sin(\beta) & \rho\cos(\beta) \end{pmatrix} \tag{39}$$

Consider the following linear dynamical system of $\Re^2$:

$$z_{k+1} = Dz_k \tag{40}$$

Lemma III.1
A necessary and sufficient condition for system (40) to admit as a positively invariant domain any regular polygone with N edges $(N \geq 3)$ and its center at the origin is:

$$\rho\cos(\frac{(2K+1)\pi}{N} - \beta) \leq \cos(\frac{\pi}{N}) \tag{41}$$

with the integer K $(0 \leq K < N-1)$ uniquely defined by :

$$\frac{2K\pi}{N} \leq \beta < \frac{2(K+1)\pi}{N} \tag{42}$$

Proof
The two relations above have a very simple geometrical interpretation, illustrated in figure 1. They describe the interior of the regular polygone with N edges having one vertex in point 1 and inscribed in the unit disk of the complex plane.

Now, consider the geometrically identical regular polygone with N edges in $\Re^2$. This polygone, denoted (Π), can be defined in polar coordinates by:

$$\Pi = \{Y = (r, \theta);\ r\cos[\frac{(2k+1)\pi}{N} - \theta] \leq \cos(\frac{\pi}{N})\} \text{ with } k \in (0, 1, \ldots, N-1) \tag{43}$$

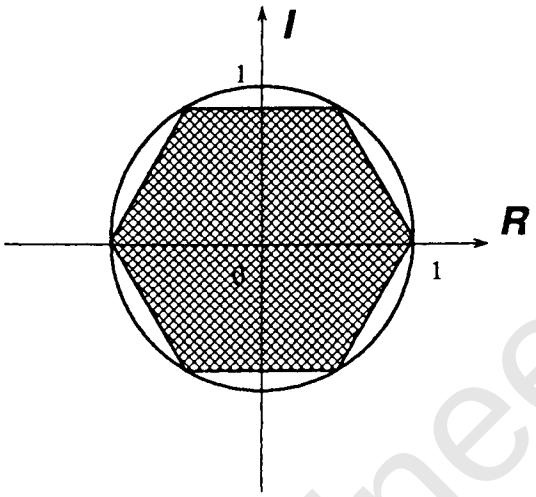


Figure 1: The polygone of admissible eigenvalues for N=6

Note that this definition of Π is similar to the relations (41) and (42) since the maximal value of $\cos[\frac{(2k+1)\pi}{N} - \theta]$ for $0 \leq k < N$, is obtained for k^* such that the absolute value of the angle is minimal, that is precisely for k^* such that:

$$\frac{2k^*\pi}{N} \leq \theta < \frac{2(k^*+1)\pi}{N}.$$

In cartesian coordinates, Π can be equivalently defined by a set of N linear inequalities:

$$\Pi = \{z = \begin{bmatrix} x \\ y \end{bmatrix}; Wz \leq w\} \tag{44}$$

with:

$$W = \begin{bmatrix} \cos(\frac{\pi}{N}) & \sin(\frac{\pi}{N}) \\ \vdots & \vdots \\ \cos(\frac{(2k+1)\pi}{N}) & \sin(\frac{(2k+1)\pi}{N}) \\ \vdots & \vdots \\ \cos(\frac{(2N-1)\pi}{N}) & \sin(\frac{(2N+1)\pi}{N}) \end{bmatrix}, w = \begin{bmatrix} \cos(\frac{\pi}{N}) \\ \vdots \\ . \\ \vdots \\ \cos(\frac{\pi}{N}) \end{bmatrix} = \cos(\frac{\pi}{N})\mathbf{1}_N$$

Since such a polygone is convex, a necessary and sufficient condition for its positive invariance is that the image of each of its vertices belongs to it. And this property can easily be checked. Let $S_k = (1, \theta_k)$, with $\theta_k = \frac{2k\pi}{N}$ be any of the N vertices of Π. Its image by D is:

$$S'_k = (\rho, \theta'_k) \text{ with } \theta'_k = \beta + \frac{2k\pi}{N}(mod 2\pi)$$

Note that :

$$\frac{2(K+k)\pi}{N}(mod 2\pi) \leq \theta' \leq \frac{2(K+k)\pi}{N}(mod 2\pi) + \frac{2\pi}{N}.$$

Therefore, by condition (41),

$$\rho\cos(\frac{2(K+k)+1)\pi}{N} - \theta'_k) = \rho(\cos(\frac{(2K+1)\pi}{2N}) - \beta) \quad (45)$$

$$\leq \cos(\frac{\pi}{N}). \quad (46)$$

And thus S'_k to belong to domain Π.

Conversely, if the conditions of the lemma are not satisfied,

$$\rho\cos(\frac{(2K+1)\pi}{N} - \beta) > \cos(\frac{\pi}{N})$$

and by (45), the image of any vertex S_k of (Π) is outside (Π). Now, if a domain (Π) is invariant by D in $\Re^2$, any homothetic domain is also invariant (section II.F). Furthermore, since D is a rotation matrix, any rotation of the domain around the origin preserves the invariance property. Indeed, the proof of invariance is the same as above if the polar coordinates of the current vertex is $(1, \frac{2k\pi}{N} + b)$ instead of$(1, \frac{2k\pi}{N})$.

In an orthonormal basis of $\Re^2$, any regular polygone Π' with N edges and its center at the origin is invariant by matrix D.

□

3. Blocks associated with multiple complex eigenvalues

The blocks Δ_m $(m = 1, \ldots, p_3)$ correspond to couples of multiple complex eigenvalues: $\begin{cases} \rho[\cos(\beta_m) + j\sin(\beta_m)] \\ \rho[\cos(\beta_m) - j\sin(\beta_m)] \end{cases}$, with order of multiplicity r_m in this block, and such that $\rho_m^2 < 1$, $\rho_m > 0$, $0 \leq \beta_m \leq 2\pi$.

The real Jordan block associated with these multiple complex eigenvalues take the form :

$$\Delta = \begin{bmatrix} D & 0 & . & . & . & . & 0 \\ I_{2\times 2} & . & . & & & & . \\ . & I_{2\times 2} & . & . & & & . \\ & & & & & & \\ . & & & & . & . & . \\ . & & & & & . & 0 \\ 0 & . & . & . & . & I_{2\times 2} & D \end{bmatrix}$$

with D defined as above, in paragraph III.B.2.

Consider the linear system of $\Re^{2r}$:

$$z_{k+1} = \Delta z_k. \tag{47}$$

A condition slightly stronger than (41) will now be used:

$$\rho \cos(\frac{(2K+1)\pi}{N} - \beta) < \cos(\frac{\pi}{N}) \tag{48}$$

with the integer K defined by (42).

Lemma III.2

Under conditions (42) and (48) on ρ and β, system (47) admits as a positively invariant domain $R[\mathcal{W}, \alpha]$, with $\mathcal{W} \in \Re^{Ns*2s}$ defined by

$$\mathcal{W} = \begin{bmatrix} W & 0 & . & . & . & . & 0 \\ 0 & W & . & & & & . \\ . & . & . & . & & & . \\ & & & & & & \\ . & & & & . & . & . \\ . & & & & & . & 0 \\ 0 & . & . & . & . & 0 & W \end{bmatrix}$$

and $\alpha \in \Re^{Nr}$ defined by $\alpha = \begin{bmatrix} a_1 w \\ \vdots \\ a_s w \end{bmatrix}$ with $w = \cos(\frac{\pi}{N})\mathbf{1}_N$, under conditions

$$\begin{cases} a_1 > 0 \\ a_i \geq \frac{a_{i-1}\cos(\frac{\pi}{N})}{\xi} \; ; \; i = 2, ..r \end{cases}$$

, W and w defined as in Lemma III.1, and

$$\xi = \cos(\frac{\pi}{N}) - \rho \cos(\frac{(2K+1)\pi}{N} - \beta). \tag{49}$$

Proof

Take any point $z \in R[\mathcal{W}, \alpha]$; $z = \begin{bmatrix} z_1 \\ \vdots \\ z_r \end{bmatrix}$, with $z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \in R[W, a_i w]$ for $i = 1, \ldots, s$. Let $S_{i1}, \ldots, S_{iN}$ be the N vertices of the polygone $R[W, a_i w]$.

z_i can be written : $z_i = \sum_{j=1}^{N} \chi_{ij} S_{ij}$, with $\sum_{j=1}^{N} \chi_{ij} = 1$, $\chi_{ij} \geq 0$.

Then, $\mathcal{D}z = \begin{bmatrix} & & Dz_1 \\ z_1 & + & Dz_2 \\ & \vdots & \\ z_{s-1} & + & Dz_s \end{bmatrix}$; $\mathcal{W}\mathcal{D}z = \begin{bmatrix} & & WDz_1 \\ Wz_1 & + & WDz_2 \\ & \vdots & \\ Wz_{s-1} & + & WDz_s \end{bmatrix}$

By construction, for $i = 1, \ldots, s$

$$Wz_i \leq a_i w \text{ and } WDz_i = \sum_{j=1}^{N} \chi_{ij} WDS_{ij}$$

but, from the proof of Lemma III.1,

$$WDS_{ij} \leq \rho \cos[\frac{(2K+1)\pi}{N} - \beta] a_i \mathbf{1}_N$$

thus

$$WDS_{ij} \leq (\cos(\frac{\pi}{N}) - \xi) a_i \mathbf{1}_N$$

and , under condition $a_i \xi \geq a_{i-1} \cos(\frac{\pi}{N})$ for $i = 2, \ldots, s$,

$$\begin{aligned} Wz_{i-1} + WDz_i &\leq (a_{i-1} \cos(\frac{\pi}{N}) + a_i(\cos(\frac{\pi}{N}) - \xi))\mathbf{1}_N \\ &\leq \cos(\frac{\pi}{N}) a_i \mathbf{1}_N. \end{aligned}$$

□

C. Symmetrical polytopes of the Jordan system

The positively invariant polygones of $\Re^2$ which have been considered in Lemmas III.1 and III.2 are regular and have their center at the origin, but they are not necessarily symmetrical with respect to the origin. The symmetrical case is simply obtained when the number of edges,N, is even ($N = 2\gamma$). with $\gamma \geq 2$. Then, as for any angle $\mathbf{a}$ ($0 \leq \mathbf{a} \leq \pi$), $\begin{cases} \cos(\mathbf{a} + \pi) = -\cos \mathbf{a} \\ \sin(\mathbf{a} + \pi) = -\sin \mathbf{a} \end{cases}$, and the analytical expression of $R[W, a_i w]$ can be re-written symmetrically as follows :

$$S(\Gamma, a_i c) = \{z_i \in \Re^2; -a_i c \leq \Gamma z_i \leq a_i c\} \tag{50}$$

$$\text{with } \Gamma \in \Re^{\gamma * 2s}; \Gamma = \begin{bmatrix} \cos(\frac{\pi}{2\gamma}) & \sin(\frac{\pi}{2\gamma}) \\ \vdots & \vdots \\ \cos(\frac{(2\gamma-1)\pi}{2\gamma}) & \sin(\frac{(2\gamma+1)\pi}{2\gamma}) \end{bmatrix}, \; c \in \Re^{\gamma}_{+}; c = \cos(\tfrac{\pi}{2\gamma})\mathbf{1}_{\gamma}.$$

The constructed positively invariant symmetrical polytope for subsystem (40) is now denoted $S(\Gamma, c)$, and the positively invariant symmetrical polytope for system (47) can be re-written $S(\mathcal{C}, \eta)$ with $\mathcal{C} \in \Re^{r\gamma \times 2r}$ and $\eta \in \Re^{\gamma r}$ defined by

$$\mathcal{C} = \begin{bmatrix} \Gamma & 0 & . & & . & 0 \\ 0 & \Gamma & & & & . \\ . & & . & & & . \\ . & & & & & 0 \\ 0 & . & & . & 0 & \Gamma \end{bmatrix}, \quad \eta = \begin{bmatrix} a_1 c \\ \vdots \\ a_r c \end{bmatrix}.$$

By construction, the number of constraints defining the symmetrical polytope of $\Re^{2r}$, $S(\mathcal{C}, \eta)$, is $2\gamma r$.

Because of the block-diagonal structure of matrix $\tilde{A}_0$, a positively invariant polytope of system (35) can be constructed from the positively invariant symmetrical polytopes of all its blocks. This polytope takes the form $S(G', \omega)$, with, in the most general case when $p_1 \neq 0$, $p_2 \neq 0$, $p_3 \neq 0$:

$$G' = \begin{bmatrix} \mathcal{I}_{q_1} & 0 & . & & & & . & 0 \\ 0 & . & 0 & & & & & . \\ . & & \mathcal{I}_{q_{p1}} & . & & & & \\ & & . & \Gamma_1 & . & & & \\ & & & & . & & & \\ & & & & & \Gamma_{p2} & 0 & \\ & & & & & . & \mathcal{C}_1 & . \\ . & & & & & & . & 0 \\ 0 & . & & & & . & 0 & \mathcal{C}_{p3} \end{bmatrix} \text{ and } \omega = \begin{bmatrix} l_1 \\ . \\ l_{p1} \\ c_1 \\ . \\ c_{p2} \\ \eta_1 \\ . \\ \eta_{p_3} \end{bmatrix}.$$

Concatenation of the invariance relations associated with each block implies the existence of a matrix H satisfying :

$$HG' = G'\tilde{A}_0 \tag{51}$$

$$|H|h < \omega \tag{52}$$

And thus, from Proposition II.3, $S(G', \omega)$ is a positively invariant symmetrical polytope of system (35). $G' \in \Re^{s \times n}$ and $\omega \in \Re^{s}$ with, by Jordan decomposition of A_0 and construction of the positively invariant domain ,

$$\begin{cases} n = \sum_{i=1}^{p_1} q_i + 2p_2 + 2\sum_{m=1}^{p_3} r_m \\ s = \sum_{i=1}^{p_1} q_i + \sum_{j=1}^{p_2} \gamma_j \sum_{m=1}^{p_3} r_m \gamma_m \end{cases}.$$

D. Positively invariant polyhedral sets for stable linear systems

Lemma II.5 and the decomposed construction of sections III.B, III.C now allow to state the following result.

Proposition III.2

Any asymptotically stable system (3) admits some symmetrical polytopes as positively invariant sets. The number of linear constraints defining such polytopes is

$$u = 2s = 2(\sum_{i=1}^{p_1} q_i + \sum_{j=1}^{p_2} \gamma_j \sum_{m=1}^{p_3} r_m \gamma_m)$$

q_i is the order of multiplicity of the real eigenvalue λ_i (with $|\lambda_i| < 1$) in the block L_i.

γ_m is the smallest integer (necessarily, $\gamma_m \geq 2$) associated with the pair of eigenvalues $\rho_m(\cos(\beta_m) \pm j\sin(\beta_m))$ with order of multiplicity r_m in the block D_m of $\tilde{A}_0$, such that:

$$\frac{K\pi}{\gamma_m} \leq \beta_m < \frac{(K+1)\pi}{\gamma_m} \tag{53}$$

$$\rho_m \cos[\frac{(2K+1)\pi}{2\gamma_m} - \beta_m] < \cos\frac{\pi}{2\gamma_m}. \tag{54}$$

Proof

Consider the real Jordan form of matrix A_0, $\tilde{A}_0$, defined by (34),under the asumption that system (3) is asymptotically stable, or that all its eigenvalues have their module strictly less than 1. From Lemma II.5, a positively invariant symmetrical polytope $S(G,\omega)$ for system (3) can easily be derived from the construction of the positively invariant symmetrical polytope $S(G',\omega)$ for system (35). It suffices to compute $G \in \Re^{s\times n}$ by :

$$G = G'P^{-1}. \tag{55}$$

to obtain from (33) and (51)

$$HG = GA_0. \tag{56}$$

Conditions (56) and (52) guarantee positive invariance of $S(G,\omega)$ w.r.t. system (3).

□

E. The case of simplicial invariant symmetrical polytopes

A case of particular interest is obtained when $s = n$. The two polytopes $S(G', \omega)$ and $S(G, \omega)$ are then simplicial, that is they are defined by non-singular matrices of $\Re^{n \times n}$, G' and G, respectively. Furthermore, from proposition III.2, this case is obtained for :

$|\lambda_i| < 1$ for any real eigenvalue of A_0,

$\gamma_m = 2$ for any complex eigenvalue $\rho_m(\cos(\beta_m) + j \sin(\beta_m))$.

In the special case when $\gamma_m = 2$ is feasible for all the blocks D_m, condition (48) on the location of the complex poles takes the simple form:

$$|\rho \cos \beta_m| + |\rho \sin \beta_m| < 1$$

for each elementary block: $\begin{pmatrix} \rho\cos(\beta_m) & \rho\sin(\beta_m) \\ -\rho\sin(\beta_m) & \rho\cos(\beta_m) \end{pmatrix}$.

The basic positively invariant square associated with each such block is $S(\Gamma, c)$, with :

$$\Gamma = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \text{ and } c = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}.$$

By a rotation of $-\frac{\pi}{4}$ and a homothesis of $\sqrt{2}$, an other elementary positively invariant square of system (40) is obtained : $S(I_2, 1_2)$. Similarly, for any block Δ_i associated with a couple of complex eigenvalues $\mu_i \pm j\sigma_i$ with multiplicity order $r_i > 1$ in this block, it is always possible to select r_i positive scalars $a_{i1}, ..., a_{ir_i}$ satisfying the conditions of Lemma III.2:

$$\begin{cases} a_{i1} > 0 \\ a_{ik} \geq \frac{a_{ik-1}\cos(\frac{\pi}{N})}{\xi_i} \; ; \; k = 2, ..r_i \end{cases}$$

with $\xi_i = 1 - |\mu_i| - |\sigma_i| > 0$, to guarantee positive invariance of $S(I_{2r_i}, \bar{a}_i)$, for

$$\bar{a}_i = \begin{bmatrix} a_{i1} 1_2 \\ . \\ a_{ir_i} 1_2 \end{bmatrix}. \tag{57}$$

Matrix G' then takes the form of the identity matrix of $\Re^n$. And relation (55) reduces to : $G = P^{-1}$. The row-vectors of matrix G form a set of left generalized real eigenvectors of matrix A_0. The following result, due to G. Bitsoris 1988 [14], then becomes a direct consequence of Proposition III.2.

Proposition III.3

A sufficient condition for the existence of a simplicial symmetrical invariant polytope $S(G,\omega)$, with $G \in \Re^{n*n}$, $\omega \in \Re^n{}_+$ for system (3) is that all the eigenvalues of A_0 (real and complex), denoted $\mu_i + j\sigma_i$, are such that:

$$|\mu_i| + |\sigma_i| < 1 \tag{58}$$

The spectral domain defined by (58) is represented on Fig.2. In general,

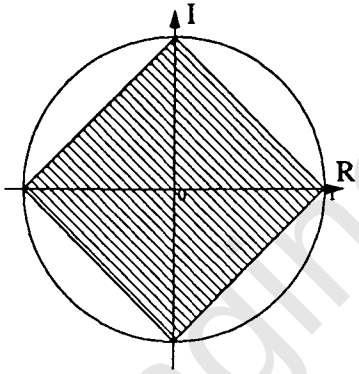


Figure 2: The spectral domain

condition (58), which is associated with the selected construction technique, is sufficient but not necessary for the existence of such a polytope $S(G,\omega)$. However, it is not difficult to show that condition (58) is also necessary for the particular choice $G = P^{-1}$, that is if matrix H of relation (56) is selected under the real Jordan form. A candidate vector ω can then be constructed as:

$$\omega^T = [l_1{}^T, ..., l_{p_1}{}^T, d_1{}^T, ..., d_{p_2}{}^T, f_1{}^T, ..., f_{p_3}{}^T]$$

with $l_i{}^T = (l_{i1}^T, ..., l_{iq_i}^T)$ under the conditions (37), for $i = 1, ..., p_1$,
$d_i = \delta_i 1_2$, with δ_i any positive number, for $i = 1, ..., p_2$,
$f_i = \phi_i \bar{a}_i$ with ϕ_i any positive number and the vector $\bar{a}_i \in \Re^{2r_i}$ constructed as in (57).

IV. FEEDBACK CONTROL OF STATE-CONSTRAINED LINEAR SYSTEMS

A. The Design Approach

Consider now the case of a linear system described by the state equation :

$$x_{k+1} = Ax_k + Bu_k \text{ for } k \in \mathcal{N} \tag{59}$$

with $x_k \in \Re^n$, $u_k \in \Re^m$, $A \in \Re^{n*n}$, $B \in \Re^{n*m}$, $m \leq n$. The state vector x_k is subject to symmetrical linear constraints:

$$-\omega \leq Gx_k \leq \omega \text{ for } k \in \mathcal{N} \tag{60}$$

with $G \in \Re^{r*n}$, $r \leq n$, $\text{rank}(G) = r$, $\omega \in \Re^r$ and $\omega_i > 0$ for $i = 1, ..., r$. These constraints are supposed to be satisfied by the initial state of the system, x_0.

A possible design technique to solve such a state-constrained control problem consists of constructing a closed-loop linear regulation law:

$$u_k = Fx_k \text{ with } F \in \Re^{m*n} \tag{61}$$

asymptotically stabilizing the system while maintaining its state vector in a domain Ω such that :

$$\Omega \subset S(G, \omega). \tag{62}$$

The problem will then be solved by restricting the set of allowed initial states to Ω and by imposing positive invariance of Ω with respect to the controlled system :

$$x_{k+1} = A_0 x_k \text{ with } A_0 = A + BF. \tag{63}$$

Clearly, the domain Ω should be constructed as large as possible, and the best possible choice in terms of allowable initial states is $\Omega = S(G, \omega)$. But, as it will now be shown, such a choice is not always possible, for structural or for stability reasons.

The structural limitation derives from Proposition II.4 of section II.I. Using the definition of (A,B)-invariance given in W.M. Wonham [10], it can be formulated as follows :

Lemma IV.1

(A,B)-invariance of the subspace Ker G is a necessary condition for the existence of a feedback gain matrix F such the closed-loop system (63) admits $S(G, \omega)$ as a positively invariant domain .

Proof

From Proposition II.3, positive invariance of $S(G,\omega)$ with respect to (63) is equivalent to the existence of a matrix $H \in \Re^{s\times s}$ such that :

$$HG = G(A+BF) \quad (64)$$

$$|H|\omega \leq \omega. \quad (65)$$

As shown in the proof of Proposition II.4, existence of a matrix H satisfying relation (64) is equivalent to invariance of $\mathcal{K}er\ G$ with respect to (63). Equivalently, the gain matrix F should be a "friend" of $\mathcal{K}er\ G$ [10] relatively to the pair (A,B).

□

If the domain $S(G,\omega)$ is a polytope, that is if $s \geq n$ and rank$G = n$, it is always possible, for any $F \in \Re^{m\times n}$, to satisfy (64). Select, for instance, $H = G(A+BF)(G^TG)^{-1}G^T$. In other words, using Lemma IV.1, relation (64) can be trivially satisfied since $\mathcal{K}er\ G = \emptyset$.

If the structural condition of Lemma IV.1 is satisfied, the second condition for positive invariance of $S(G,\omega)$, (65), and additional stability condition will be required to solve the constrained regulation problem. At this point, two approaches have been developed : one by Linear Programming, which directly provides an admissible solution whenever it exists, and one by eigenstructure assignment which not only solves the problem if possible, but also indicates, if the problem is unfeasible, how to select a better domain for positive invariance with stability.

The scope of application of these two approaches are different. The use of the eigenstructure assignment technique is probably better adapted to the case rank $G < n$, when the condition of Lemma IV.1 has to be obtained, and the Linear Programming technique is appropriate when rank $G = n$, because relation (64) is then trivially satisfied. In this case, satisfaction of relation (65) both implies positive invariance of $S(G,\omega)$ and closed-loop stability of (63), as it is shown in the next section.

B. Positive invariance of a symmetrical polytope

1. Positive invariance and stability

Lemma II.4 of section II.G establishes that, for rank $Q = n$ and $\phi > 0$, contractive invariance of a simplicial proper polyhedron $R[Q,\phi]$ with respect to system (63) implies asymptotic stability of system (63). An other stability result is obtained if the positively invariant domain is a symmetrical polytope $S(G,\omega)$, with rank $G = n$ and $\omega > 0$.

Lemma IV.2
Positive invariance with respect to (63) of $S(G,\omega)$, with rank $G = n$ and $\omega > 0$ implies Lyapunov stability of (63).

To proove this result, it suffices to show that the function

$$w(x) = \max_{i=1,..,s} \{\frac{|(Gx)_i|}{(\omega)_i}\} \tag{66}$$

is a Lyapunov function of system (63).

If rank $G = n$, the constrained regulation problem can thus be reduced to the determination of a gain matrix F such that the positive invariance conditions (64),(65) are satisfied.

2. Resolution by Linear Programming

A possible design technique was proposed in [29]. It is also related to the algorithm presented in [30]. Any matrix $H \in \Re^{r\times r}$ can be decomposed into:
-the matrix of its non-negative elements :H^+, with $H^+{}_{ij} = max(H_{ij}, 0)$
-the matrix of its non-positive elements :$-H^-$, with $H^-{}_{ij} = max(-H_{ij}, 0)$
so that

$$H = H^+ - H^- \text{ and } |H| = H^+ + H^-. \tag{67}$$

Using matrices H^+ and H^-, relation (65) is then formulated as a linear relation. A possible performance index to be optimized is the contraction rate of function $w(x)$ defined by (66). As shown in [29], optimizing such an index corresponds, in some sense, to maximizing the convergence of (63) and the robustness of the stability property with respect to uncertainties on A_0. An other advantage of selecting such a linear index is to formulate and to solve the design problem as a Linear Programming problem.

It is now interesting to show that in relation (64), matrix H can be simply replaced by $X - Y$ and in relation (65) matrix $|H|$ replaced by $X + Y$ with $X \leq 0_{s\times s}$ and $Y \leq 0_{s\times s}$ to obtain the following formulation of the L.P., with ϵ and the coefficients of matrices X, Y, F as unknown variables:

$$\begin{array}{lrcl} \text{Minimize} & \epsilon & & \\ \text{under} & (X+Y)\omega - \epsilon\omega & \leq & 0_s \\ & (X-Y)G - GBF & = & PA \\ & \epsilon,\ X,\ Y & \geq & 0_{s\times s} \end{array} \tag{68}$$

If the pair (X, Y) is feasible, set $H = X - Y$. Then the pair (H^+, H^-) also satisfies:

$$(H^+ + H^-)\omega - \epsilon\omega \leq 0_s,$$

since the following (componentwise) inequalities are always satisfied:

$$H^+ \leq X \text{ and } H^- \leq Y.$$

Furthermore, the value of ϵ obtained for the pair (H^+, H^-) is always less than or equal to the one obtained for (X, Y). To avoid non-minimal solutions, that is solutions with $X_{ij} \neq 0$ and $Y_{ij} \neq 0$ for some pairs of indices (i, j), it suffices to add in the criterion vector some very small weights on each variable X_{ij} and Y_{ij}. The minimum of the modified criterion is then obtained only if $X = H^+$, $Y = H^-$. Therefore, the optimal solution of this problem can be denoted $(\epsilon^*, H^{+*}, H^{-*}, F^*)$,.

If this solution is such that $\epsilon^* < 1$, the closed-loop system admits $S(G, \omega)$ as a positively invariant domain . The Lyapunov function (66) is strictly decreasing, and matrix $A + BF^*$ is asymptotically stable.

If the optimal solution of (68) is such that $\epsilon^* > 1$, then, positive invariance of $S(G, \omega)$ cannot be obtained by any static state feedback.

If $\epsilon^* = 1$, positive invariance of $S(G, \omega)$ can be obtained but not together with asymptotic stability of the closed-loop system.

The two last cases do not necessarily mean that system (63) cannot be stabilized. These cases can even occur with the pair (A,B) controllable. In this case, a solution $\epsilon^* > 1$ would simply mean that the function $w(x)$ of relation (66) cannot be used as a Lyapunov function for system (63).

Some additional design requirements can be introduced in the formulation of the Linear Program (68). In particular, if the desired dynamics of the system should not be too fast, it suffices to impose :

$$\epsilon \geq \epsilon_0. \tag{69}$$

Such a constraint, with $0 < \epsilon_0 < 1$ guarantees that the spectral radius of the closed-loop matrix A_0 is not smaller than ϵ_0. This result derives from the fact that the function $w(x)$ can be interpreted as an induced polyhedral norm of matrix A_0, and from the property of the spectral radius of a matrix to be always greater than or equal to any norm of this matrix.

An other classical additional requirement is to impose some constraints on the control vector $u_k = Fx_k$. This case of state and control constrained problem was treated in details in [30] and, with a different algorithm, in [16]. A simple solution to limit the magnitude of the control vector is to impose a constraint on the L_∞ norm of matrix F :

$$\|F\|_\infty \leq f \tag{70}$$

Such a constraint can be translated into a linear constraint by setting $F = F^+ - F^-$ with both F^+ and F^- matrices of $\Re^{m \times n}$ with non-negative components. Constraint (70) can then be replaced by

$$(F^+ + F^-)1_n \leq f1_n$$

and added to the constraints of problem (68), in which the gain matrix F is replaced by $F^+ - F^-$.

3. Example

Consider the following data:

$$A = \begin{bmatrix} 0.4 & 1.55 & -0.625 \\ -0.1 & 0.4 & -0.25 \\ -0.7 & -0.1 & 0.25 \end{bmatrix} \text{ and } B = \begin{bmatrix} -0.5 \\ 1.0 \\ 2.2 \end{bmatrix}$$

The open-loop system has one unstable eigenvalue and two stable ones :

$$\lambda(A) = \begin{bmatrix} 1.1282 \\ -0.0391 + j0.3731 \\ -0.0391 - j0.3731 \end{bmatrix}$$

The state constraints are defined by

$$G = \begin{bmatrix} 0.0 & 1.9 & -0.9 \\ 0.8 & 0.2 & -0.7 \\ 0.2 & -0.8 & 1.6 \end{bmatrix} \; ; \; \omega = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Resolution of the Linear Program (68) gives :

$$\epsilon^* \simeq 0.8987$$

for

$$H = \begin{bmatrix} 0.3864 & 0.4938 & 0.0185 \\ 0.4814 & -0.1909 & 0.2264 \\ 0.4096 & 0.4891 & 0.0 \end{bmatrix} \text{ and } F^T = \begin{bmatrix} 0.5158 \\ 0.3992 \\ -0.4527 \end{bmatrix}$$

The controlled system is stable. The eigenvalues of $A_0 = A + BF$ are also the eigenvalues of H :

$$\lambda(A_0) = \begin{bmatrix} 0.7692 \\ -0.0207 \\ -0.5529 \end{bmatrix}$$

The control $u_k = Fx_k$ can be applied to any state vector $x_k \in S(G, \omega)$. The resulting trajectory of the state vector remains in this symmetrical polytope which is positively invariant. The closed-loop system is stable, and this stability property is robust to any "small" perturbation of any component of A and B. This robustness property of regulators letting a symmetrical polytope positively invariant will be further analyzed in section IV.D.

C Positive invariance obtained by eigenstructure assignment

1. Assignment of closed-loop poles to system zeros

Assume now that matrix $G \in \Re^{s*n}$, with $s \leq n$, is full rank. As stated in Lemma IV.1, (A,B)-invariance of the subspace $\mathcal{K}er\ G$ is a necessary condition for obtaining positive invariance of the symmetrical polyhedron $S(G,\omega)$ with respect to the closed-loop system (63).

Consider the system matrix of the triplet (A, B, G) [27]:

$$P(\lambda) = \begin{bmatrix} \lambda I - A & -B \\ G & 0_{s*m} \end{bmatrix} \tag{71}$$

(A,B)-invariance of $\mathcal{K}er\ G$ requires and implies the existence of a gain matrix F such that $n - s$ independent generalized eigenvectors of $A + BF$ belong to $\mathcal{K}er\ G$. Equivalently, there should exist a system of $n - s$ independent generalized real eigenvectors $(v_1, ..., v_{n-s})$ and a matrix $J_1 \in \Re^{(n-s)\times(n-s)}$ having the real Jordan canonical form, satisfying:

$$\begin{cases} (A + BF)V_1 = V_1 J_1 & \text{with } V_1 = (v_1, ..., v_{n-r}), \\ GV_1 = 0 \end{cases} \tag{72}$$

Classically, the zeros of (A, B, G) are defined as the set of complex numbers λ_i for which there exist vectors $v_i \in \mathcal{C}^n$ and $w_i \in \mathcal{C}^m$ such that:

$$P(\lambda_i) \begin{bmatrix} v_i \\ w_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{73}$$

v_i is called a state zero direction, and w_i an input zero direction.

Taking into account the possible existence of zeros with a multiplicity order greater than 1, the zero subspace of of (A, B, G) can be defined as the subspace spanned by all the vectors v_i solutions of (73) for any possible complex value of λ_i. It is then possible to state the following Proposition :

Proposition IV.1

A necessary condition for the existence of a gain matrix F such that $S(G,\omega)$ is positively invariant w.r.t. the closed-loop system (63) is that the dimension of the zero subspace of (A,B,G) is not less than the dimension of $\mathcal{K}er\ G$.

If this condition is satisfied, then it is possible to obtain the $(A + BF)$-invariance of $\mathcal{K}er\ G$ by selecting F such that :

$$FV_1 = W_1 \tag{74}$$

This equation always admits the pseudo-inverse solution : $F = W_1(V_1^T V_1)^{-1} V_1^T$. But this solution is generally not satisfactory for obtaining positive invariance of $S(G,\omega)$. A more complete eigenstructure assignment scheme will actually be needed.

Conversely, if the dimension of the zero subspace of (A,B,G) is less than the dimension of $\mathcal{K}er\ G$, then it is easy to show by contradiction that the subspace $\mathcal{K}er\ G$ cannot be (A,B)-invariant.

The problem of the existence and of the number of zeros of a linear system is solved in the literature (for a survey of the main results, see Mac Farlane et al. [31]). Depending of the numbers of input variables, m, of output variables, s, and of state variables, n, the possible cases can be outlined as follows:

1. If $s = n$, then $\mathcal{K}er\ G = \{0\}$. The (A,B)-invariance of $\mathcal{K}er\ G$ is automatically satisfied.

2. If $m < s < n$, in general the dimension of the zero subspace is less than $n - s$ and therefore the zero subspace is strictly included in $\mathcal{K}er\ G$. Thus, from Proposition IV.1, **positive invariance of $S(G,\omega)$ cannot generally be obtained.** The particular case when the system has $n - s$ zeros is similar to the following case.

3. (a) If $s = m$ but $\text{rank}(GB) = m - d$, with $d > 0$, the system has only $n - m - d$ finite zeros and d infinite zeros, which cannot be used as closed-loop poles. $\mathcal{K}er\ G$ is not (A,B)-invariant.

(b) If $s = m$, and $\text{rank}(GB) = m$, then the system has exactly $n - m$ finite zeros. If these zeros are stable, they can be selected as closed-loop eigenvalues, to obtain the (A+BF)-invariance of $\mathcal{K}er\ G$. But if any of these zeros is unstable, closed-loop stability and positive invariance of $R(G,g)$ cannot both be obtained.

4. If $s < m$, equation (73) has solutions for any complex value λ_i. These "controllable" solutions generate the maximal controllability subspace of (A,B) included in $\mathcal{K}er\ G$. The system may also admit invariant zeros. Their associated zero-directions generate the maximal (A,B)-invariant subspace of $\mathcal{K}er\ G$ not intersecting $Im(B)$. Under the condition $\text{rank}(GB) = s$, the direct sum of these two independent subspaces is $\mathcal{K}er\ G$.

The three last cases can be summarized in the following Proposition [32]:

Proposition IV.2

If $s \leq m$ and $s < n$, condition $\text{rank}(GB) = s$ is sufficient for the (A,B)-invariance of $\mathcal{K}er\ G$. In order to obtain the (A+BF)-invariance of $\mathcal{K}er\ G$, it is necessary to locate closed-loop eigenvalues at all the invariant zeros of the system. If any of these zeros is unstable, (A+BF)-invariance of $\mathcal{K}er\ G$ and closed-loop stability will not be simultaneously obtained.

Consider the more general case $\text{rank}(G) = s \leq m$, when matrix GB can be rank deficient. Let d be the rank deficiency of matrix GB; $\text{rank}(GB) = s - d$. To find the zeros of (A, B, G), define, respectively, right and left annihilators of matrices G and B, $M \in \Re^{n*(n-r)}$ and $N \in \Re^{(n-m)*n}$, satisfying the following relations:

$$GM = 0_{r,n-r} \ , \ NB = 0_{n-m,m}.$$

The degrees of freedom on the choice of matrices N and M, and the property $\text{rank}(NM) = n - m - d$, allow to choose N and M so that [33]:

$$NM = \left[\begin{array}{c|c} I_{n-m-d,n-m-d} & 0_{n-m-d,m+d-s} \\ 0_{d,n-m-d} & 0_{d,m+d-s} \end{array}\right] \tag{75}$$

Any vector $v_i \in \mathcal{C}^n$ satisfying relation (73) belongs to $\mathcal{K}er\ G \subset \mathcal{C}^n$. Therefore, it is uniquely defined by the vector $z_i \in \mathcal{C}^{n-s}$ such that:

$$v_i = M z_i \tag{76}$$

Relation (73) can then be equivalently replaced by:

$$[\lambda_i I - A \ \ -B] \left[\begin{array}{c} M z_i \\ w_i \end{array}\right] = 0 \tag{77}$$

The m components of w_i can be eliminated by left multiplication of (77) by matrix N, yielding:

$$[\lambda_i NM - NAM] z_i = 0 \tag{78}$$

Equations (73) and (78) have the same solutions $\lambda_i \in \mathcal{C}$, which are the finite zeros of (A, B, G). The polynomial matrix $[\lambda NM - NAM]$ is called the zero pencil [33]. It completely characterizes the finite zeros and the associated zero directions of (A, B, G). Using for matrix NAM the same partitionning as for NM, we can write the zero pencil as follows [34]:

$$\lambda NM - NAM = \left[\begin{array}{c|c} \lambda I - (NAM)_1 & -(NAM)_2 \\ -(NAM)_3 & -(NAM)_4 \end{array}\right]. \tag{79}$$

This decomposition indicates that the zeros of system (A, B, G) are also the zeros of the non-proper system $[(NAM)_1, (NAM)_2, -(NAM)_3, -(NAM)_4]$. Any zero-direction associated to a value of $\lambda \in \mathcal{C}$ can be defined from a vector $z \in \mathcal{C}^{n-s}$ belonging to the transmission subspace $Tr(\lambda)$ of system $((NAM)_1, (NAM)_2)$. By extension of the classical definition of (state) transmission subspaces [35], $Tr(\lambda)$ is defined as the kernel of the pole pencil: $[\lambda I_{n-m-d,n-m-d} - (NAM)_1 \mid -(NAM)_2]$. Then, in order to satisfy relation (78) for some value of λ_i, vector z_i has to satisfy the two following relations:

$$[\lambda_i I_{n-m-d,n-m-d} - (NAM)_1 \mid -(NAM)_2] z_i = 0 \tag{80}$$

$$[-(NAM)_3 \mid -(NAM)_4]z_i = 0 \tag{81}$$

We are now ready to state the following proposition:

Proposition IV.3
A necessary and sufficient condition for $\mathcal{K}er\ G$ *to be spanned by zero-directions is:*

$$\delta = dim\{Image[-(NAM)_3 \mid -(NAM)_4]\} = 0$$

(or equivalently, $dim\{\mathcal{K}er[-(NAM)_3 \mid -(NAM)_4]\} = n - s$*)*
This condition can be split into two alternatives:

- *either* $d = 0$
- *or* $d > 0$ *but* $[-(NAM)_3 \mid -(NAM)_4] = 0_{d,n-s}$.

Then, the condition of $(A + BF)$*-invariance of* $\mathcal{K}er\ G$ *can be satisfied but stability of the closed-loop system is also obtained if and only if all the invariant zeros of* (A, B, G) *are stable (located in the unit-circle of the complex plane).*

Proof
necessity
Invariant and controllable zeros have to satisfy the two relations (80) and (81). But since the second condition does not depend on the value of λ, it constrains all the candidate vectors z to belong to $\mathcal{K}er[-(NAM)_3 \mid -(NAM)_4]$. If this kernel is not the whole space $\mathcal{C}^{n-s}$ but has dimension $n - s - \delta$ with $\delta > 0$, then the associated zero directions $v = Mz$ can at most generate a subspace of $\mathcal{K}er\ G$ with dimension $n - s - \delta$.
sufficiency
Let us now assume that the condition above ($\delta = 0$) is satisfied. Then, condition (81) can be suppressed. Only relation (80) has to be verified. As noted above, the considered singular pencils satisfy $s < m + d$. Then for any complex value of λ, the transmission subspace $Tr(\lambda)$ for system with state-matrix $((NAM_1)$ and input matrix (NAM_2),

$$Tr(\lambda) = \{z; z \in \mathcal{C}^{n-s}\ ;\ [\lambda I - (NAM)_1 \mid -(NAM)_2]z = 0\}$$

has a dimension greater or equal to $m + d - s$. (a) $\dim[Tr(\lambda_i)] > m + d - s$ if λ_i is an invariant zero of (A, B, G). Note that the invariant zeros of (A, B, G), when they exist, are the "input decoupling" zeros of system $((NAM)_1, (NAM)_2)$ [33]. If the system $((NAM)_1, (NAM)_2)$ has q input decoupling zeros, these zeros are uncontrollable poles of $(NAM)_1$ and the maximal controllability subspace of the pair $((NAM)_1, (NAM)_2)$ has dimension $n - m - d - q$.

(b) $\dim[Tr(\lambda_i)] = m + d - s$ if λ_i is a "controllable" zero of (A, B, G), that is simply any complex value which is not an invariant zero.

A sufficient condition for the existence of $n - s$ zero-directions spanning $\mathcal{K}er\ G$ is that the union of all the transmission spaces $Tr(\lambda)$ (for at most $n-s$ values of λ) span $\mathcal{C}^{n-s}$.

Any $z \in \mathcal{C}^{n-s}$ can be written $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$, with $z_1 \in \mathcal{C}^{n-m-d}$ and $z_2 \in \mathcal{C}^{m+d-s}$.
If it belongs to $Tr(\lambda)$, it satisfies:

$$(NAM)_1 z_1 = \lambda z_1 - (NAM)_2 z_2 \tag{82}$$

Under the assumption $\delta = 0$, the assignment of $n - s$ closed-loop eigenvectors in $\mathcal{K}er\ G$ can be obtained in two stages:

stage 1

First, apply a state feedback Φ_1 to $((NAM)_1, (NAM)_2)$ to locate the $n - m - d - s$ eigenvalues of matrix $(NAM)'_1 = (NAM)_1 + (NAM)_2\Phi_1$ in the stable region. These eigenvalues can be selected all different , and different from the uncontrollable poles (if there are any). In order for the set of vectors z_1 solutions of (82) to span $\Re^{n-m-d}$, the set of closed-loop poles, λ_i, must include the uncontrollable poles of $(NAM)'_1$ (and $(NAM)_1$), with their order of multiplicity. As mentionned above, these uncontrollable poles are precisely the invariant zeros of (A, B, G).

The eigenvectors of $(NAM)'_1$ define $n-m-d$ independent zero directions such that $\begin{cases} z_1 \neq 0 \\ z_2 = \Phi_1 z_1 \end{cases}$. The corresponding zero-directions obtained from these vectors through relation (76) span a subspace S_1 in $\mathcal{K}er\ G$ with dimension $n - m - d$.

stage 2

Any solution of equation (82) is also a solution of

$$(NAM)'_1 z_1 = \lambda z_1 - (NAM)_2 z'_2 \text{ with } z'_2 = z_2 - \Phi_1 z_1 \tag{83}$$

Controllability properties are invariant by state-feedback. They are the same for $((NAM)'_1, (NAM)_2)$ than for $((NAM)_1, (NAM)_2)$. And under the change of coordinates in (83), the $n-s-d$ independent eigenvectors of $(NAM')_1$ satisfy

$$\begin{cases} z_1 \neq 0 \\ z'_2 = 0 \end{cases} \tag{84}$$

Consider the Jordan form associated to the controllable subspace of $(NAM)'_1$, $\Lambda = \Pi(NAM)'_1\Gamma$, where the ith line of Π is the left-eigenvector of $(NAM)'_1$ and the ith column of Γ the corresponding right eigenvector for eigenvalue λ_i, with all λ_i distinct by construction for $i = 1, .., n - m - d - q$.

Now, we can use a basis of $\Re^{m+d-s}$ as $m+d-s$ independent input vectors, z'_{2i}, of $m+d-s$ different transmission subspaces $Tr(\mu_i)$ for any set of selected distinct eigenvalues $(\mu_1, ..., \mu_{m+d-s})$.
For $i = 1, .., m+d-s$, each couple (z'_{2i}, μ_i) of an input vector and of a selected eigenvalue generates a state vector of $\Re^{n-m-d}$, $z_{1i} = \Pi(\mu_i I - \Lambda)^{-1}\Gamma(NAM)_2 z'_{2i}$. By construction, all the vectors $\begin{bmatrix} z_{1i} \\ z'_{2i} \end{bmatrix}$ are independent and independent from vectors satisfying relation (84). The associated coordinates of vectors z_i satisfying equation (82) are $\begin{bmatrix} z_{1i} \\ z'_{2i} + \Phi_1 z_{1i} \end{bmatrix}$. The corresponding zero-directions obtained from these vectors through relation (76) span a subspace S_2 of $\mathcal{K}er\ G$ independent of S_1, with dimension $m+d-s$. Therefore, it is such that its direct sum with S_1 generates $\mathcal{K}erG$: $\mathcal{K}er\ G = S_1 \oplus S_2$.
□

Note that the decomposition of $\mathcal{K}er\ G$ used in this proof is far from being unique. In fact, as it will be illustrated in the examples, we have a free choice of the $n-r-s$ controllable zeros. And the eigenstructure assignment problem can practically be solved in a single stage.

2. Assignment in a complementary subspace of $\mathcal{K}er\ G$

Whenever the (A,B)-invariance of $\mathcal{K}er\ G$ is satisfied with rank $G = s$, the existence of positively invariant domains $S(G, \omega)$ in $\Re^n$ for system (63) reduces to the existence of positively invariant domains $S(I_s, \omega)$ in $\Re^s$ for the restriction of $(A+BF)$ to $(\frac{\Re^n}{\mathcal{K}er\ G})$, matrix F being constrained to be a "friend" of $\mathcal{K}er\ G$. Matrix H in equation (64) can precisely be interpreted as the map induced in $(\frac{\Re^n}{\mathcal{K}er\ G})$ by the map $(A+BF)$ in $\Re^n$. Let us now assume the (A,B)-invariance of the subspace $\mathcal{K}er\ G$, and wonder about the existence of positively invariant polyhedra $S(D, \eta)$, with $\mathcal{K}er\ D = \mathcal{K}er\ G$ and D to be constructed and not necessarily full-rank. The following result can easily be shown from Proposition III.1 and the results of section III.

Proposition IV.4
Under the assumption that $\mathcal{K}er\ G$ is (A,B) invariant and F a "friend" of $\mathcal{K}er\ G$, the asymptotic stability of the restriction of $A+BF$ to the quotient space $\Re^n/\mathcal{K}er\ G$ is a necessary and sufficient condition for the existence of a positively invariant polyhedral domain $S(D, \eta)$ of system (63) such that $\mathcal{K}er\ D = \mathcal{K}er\ G$.

Proof
From Proposition II.4, a necessary and sufficient condition for the (A,B)-invariance of $\mathcal{K}er\ G$ is the existence of a matrix $H \in \Re^{s*s}$ satisfying relation

(64) : $G(A+BF) = HG$. The dynamics of the projection of system (63) on $\Re^n/\mathcal{K}er\ G$ are described by the evolution of vector $y \in \Re^s$ such that

$$y_{k+1} = Hy_k, \text{ with } y_k = Gx_k. \tag{85}$$

Then, from Proposition III.1, asymptotic stability of system (85) implies the existence of a matrix $D' \in \Re^{r\times s}$ and of a vector $\eta \in \Re^r$, with strictly positive components, and $r \geq s$ such that the polytope of $\Re^s$, $S(D',\eta)$ is a positively invariant symmetrical polytope of (85). By application of Proposition II.3, there exists a matrix $K \in \Re^{r\times r}$ such that :

$$KD = DH \tag{86}$$

$$|K|\eta < \eta \tag{87}$$

Then, combining relation (86) with (64), we obtain, for $D = D'G$,

$$KD = D(A+BF) \tag{88}$$

Relation (87) then implies positive invariance of $S(D,\omega)$ with $\mathcal{K}er\ D = \mathcal{K}er\ G$ for system (63).

□

In order to replace matrix D by matrix G in relation (88) matrix H should be such that Eq. (86), (87) would be satisfied for $D' = I_s$. But from Proposition III.3 and for matrix H under the real Jordan form, this result is obtained when the eigenvalues of the restriction $(A+BF)|(\Re^n/\mathcal{K}er\ G)$, denoted $\mu_i+j\sigma_i$, satisfy relation (58). The following proposition is based on this result.

Proposition IV.5

If the pair (A,B) is controllable and Ker G an (A,B)-invariant subspace, and under the conditions rank(GB)=rank(G)=s ≤m, there exists a positive vector ω such that S(G,ω) is positively invariant. The existence of such a vector, ω, is induced by the existence of a gain matrix $F \in \Re^{m*n}$ such that:

(a) Ker G is an invariant subspace of system (63).

The real generalized eigenvectors of the restriction $(A+BF)|\mathcal{K}er\ G$ are the column-vectors of a matrix V_1 satisfying:

$$GV_1 = 0_{s*(n-s)} \tag{89}$$

(b) The real generalized eigenvectors associated to the eigenvalues of $(A+BF)|(\Re^n/\mathcal{K}er\ G)$ span a subspace $R \subset \Re^n$ such that $R \oplus \mathcal{K}er\ G = \Re^n$. They can be selected as the column-vectors of a matrix V_2 satisfying:

$$GV_2 = I_s \text{ with } I_s \text{ the unity-matrix in } \Re^{s*s} \tag{90}$$

The corresponding eigenvalues, $\mu_i + j\sigma_i$, satisfy $\mu_i + |\sigma_i| < 1$.

Proof

The eigenstructure assignment problem can be solved by the same decomposition technique as in Wonham [10]. Let F_0 be a friend of $\mathcal{K}er\ G$. The real generalized eigenvectors of $(A + BF_0)|\mathcal{K}er\ G$ are the column-vectors of a matrix V_1 satisfying relation (89). Any feedback gain matrix $F = F_0 + F_1 G$, with $F_1 \in \Re^{m*s}$ is also a friend of $\mathcal{K}er\ G$, and $(A + BF)|\mathcal{K}er\ G$ is identical to $(A + BF_0)|\mathcal{K}er\ G$.

Consider the real Jordan canonical form of $(A + BF)$:

$$J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} \tag{91}$$

J_1 is the real Jordan canonical form of $(A + BF)|\mathcal{K}er\ G$ and we have:

$$(A + BF)V_1 = V_1 J_1 \tag{92}$$

J_2 is the real Jordan canonical form of $(A+BF)|(\Re^n/\mathcal{K}er\ G)$. The restriction of $(A + BF_0)$ to $(\Re^n/\mathcal{K}er\ G)$ is denoted $\bar{A}_0$ and defined by the canonical projection equation:

$$\bar{A}_0 G = G(A + BF_0) \tag{93}$$

Define $\bar{B} = GB$. Under the assumption that (A,B) is controllable, then $(\bar{A}_0, \bar{B})$ is also controllable in $(\frac{\Re^n}{\mathcal{K}er\ G})$. The eigenvalues of $(\bar{A}_0 + \bar{B}F_1)$ can be selected so as to satisfy relation (58). Their associated generalized real Jordan form is matrix J_2. Moreover, if $\text{rank}(GB) = \text{rank}(G) = s \leq m$, $s \leq m$, we can select as generalized real eigenvectors in $(\frac{\Re^n}{\mathcal{K}er\ G})$ the canonical basis constituting the columns of the identity matrix I_s. The corresponding generalized real eigenvectors in $\Re^n$ are the column-vectors of matrix V_2 defined by relation (90). To do so, it suffices to select F_1 such that:

$$\bar{B}F_1 = J_2 - \bar{A}_0 \tag{94}$$

And in particular,

$$F_1 = \bar{B}^T(\bar{B}\bar{B}^T)^{-1}(J_2 - \bar{A}_0) \tag{95}$$

Let us now define matrix $G' \in \Re^{(n-s)*n}$ such that:

$$G' = [I_{n-s} \mid 0_{(n-s)*s}][V_1 \mid V_2]^{-1} \tag{96}$$

And consider the nonsingular matrix $\begin{bmatrix} G' \\ G \end{bmatrix} \in \Re^{n*n}$. It is the inverse of matrix $[V_1|V_2]$. Then, we have:

$$\begin{bmatrix} G' \\ G \end{bmatrix}(A + BF) = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}\begin{bmatrix} G' \\ G \end{bmatrix} \tag{97}$$

And in particular,

$$G(A + BF) = J_2 G \tag{98}$$

Relation (98) is equivalent to relation (64) when selecting J_2 as a candidate matrix H. Under conditions (58) on the eigenvalues of J_2, from Proposition III.3, there exists a positive vector $\omega \in \Re^s$ for which

$$|J_2|\omega < \omega. \tag{99}$$

And, from Proposition II.3, the polyhedron $S(G, \omega)$ is a positively invariant set of the closed-loop system.

Note that, as shown in section III.E, ω cannot generally be freely chosen for any admissible pole assignment in $(\frac{\Re^n}{\mathcal{K}er\ G})$. But it can be freely chosen if all the selected closed-loop poles in $(\frac{\Re^n}{\mathcal{K}er\ G})$ are selected stable, real and simple. □

3. Implementation and Examples

To compute the eigenvectors of the closed-loop system, we use Moore formulation [36] of the *pole pencil*, $S(\lambda_i)$, and of its kernel K_{λ_i}:

$$S(\lambda_i) = [\lambda_i I - A \mid -B] \quad ; \quad K_{\lambda_i} = \begin{bmatrix} N_{\lambda_i} \\ M_{\lambda_i} \end{bmatrix}$$

Then, any solution $v_i \in \Re^n$ can be written, for some vector k_i of appropriated dimension,

$$v_i = N_{\lambda_i} k_i, \quad \text{with } F v_i = M_{\lambda_i} k_i \tag{100}$$

If the eigenvector belongs to $\mathcal{K}er\ G$, it should satisfy:

$$G N_{\lambda_i} k_i = 0 \tag{101}$$

and if it belongs to $\mathcal{L}$ such that $\mathcal{L} \oplus \mathcal{K}er\ G = \Re^n$:

$$G N_{\lambda_i} k_i = e_i \tag{102}$$

with e_i the corresponding vector belonging to the canonical basis of $\Re^r$.

These formulations can be used directly for simple real eigenvalues. In the case of simple complex conjugate eigenvalues, $(\mu_i + j\sigma_i, \mu_i - j\sigma_i)$, the complex kernels $(K_{\mu_i + j\sigma_i}, K_{\mu_i - j\sigma_i})$ are replaced by the associated real kernels of the real pencils defined as follows. The closed-loop eigenvectors and controls associated to $\mu_i + j\sigma_i$ and $\mu_i - j\sigma_i$ (with $\sigma_i \neq 0$) are denoted: $\begin{bmatrix} v_i \\ w_i \end{bmatrix}, \begin{bmatrix} v_i^* \\ w_i^* \end{bmatrix}$. They are complex conjugate. The corresponding real directions, denoted

$\begin{bmatrix} v_{\mu i} \\ w_{\mu i} \end{bmatrix}, \begin{bmatrix} v_{\sigma i} \\ w_{\sigma i} \end{bmatrix}$ are defined as the real and imaginary parts of $\begin{bmatrix} v_i \\ w_i \end{bmatrix}$. If they are assigned to $\mathcal{K}er\ G$, they should satisfy the following equations:

$$\begin{bmatrix} \mu_i I - A & -\sigma_i I & -B & 0 \\ \sigma_i I & \mu_i I - A & 0 & -B \end{bmatrix} \begin{bmatrix} v_{\mu i} \\ v_{\sigma i} \\ w_{\mu i} \\ w_{\sigma i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Once all the eigenstructure has been selected, let $V = [V_1 \mid V_2]$ be the matrix of the desired real generalized eigenvectors, and $W = [W_1 \mid W_2]$ the associated control. The feedback gain matrix providing the desired eigenstructure assignment is:

$$F = WV^{-1} \tag{103}$$

Example (from [34]): Positive Invariance without global Stability

Consider the following data:

$$A = \begin{bmatrix} 0.4832 & 0.8807 & -0.7741 \\ -0.6135 & 0.6538 & -0.9626 \\ -0.2749 & 0.4899 & 0.9933 \end{bmatrix}, \quad B = \begin{bmatrix} -0.8360 & 0.4237 \\ 0.7469 & -0.2613 \\ -0.0378 & 0.2403 \end{bmatrix}$$

The open-loop system has unstable eigenvalues: $\begin{cases} 0.4481 + j0.9683 \\ 0.4481 - j0.9683 \\ 1.2341 \end{cases}$.

The state constraints are defined by

$$G = \begin{bmatrix} -0.5192 & 0.0189 & -0.3480 \\ -0.2779 & -0.7101 & -0.6953 \end{bmatrix} \quad ; \quad \omega = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$$

System (A, B, G) has an unstable zero $\lambda_1 = 1.6166$. This value is selected as a closed-loop eigenvalue, for which the associated eigenvector spans $\mathcal{K}er\ G$. The 2 other poles have been chosen as : $\begin{cases} \lambda_1' = 0.1 + j0.6 \\ \lambda_2' = 0.1 - j0.6 \end{cases}$. Then, we get

$$A_0 = \begin{bmatrix} 0.4618 & 1.0547 & -0.0583 \\ -0.4946 & 0.0478 & -1.4530 \\ -0.0875 & -0.3519 & 1.3070 \end{bmatrix}, V = \begin{bmatrix} -2.0143 & 0.1301 & -0.1828 \\ 0.6262 & -1.2049 & -0.1856 \\ -0.1656 & -0.2596 & 0.2626 \end{bmatrix}$$

$$\text{and } F = \begin{bmatrix} 0.4571 & -2.1552 & -0.2115 \\ 0.8517 & -3.8424 & 1.2722 \end{bmatrix}$$

On Fig 3, we can see the stable time evolution of x_1' and x_2', which are the coordinates of the state vector in an orthonormal basis of the space spanned by

the second and third columns of V. But the global unstability clearly stands out on Fig 4. Projection of the state vector on the first eigenvector, V_1, has a divergent evolution. In this example, global stability of the closed-loop system and positive invariance of $S(G, \omega)$ cannot be simultaneously achieved.

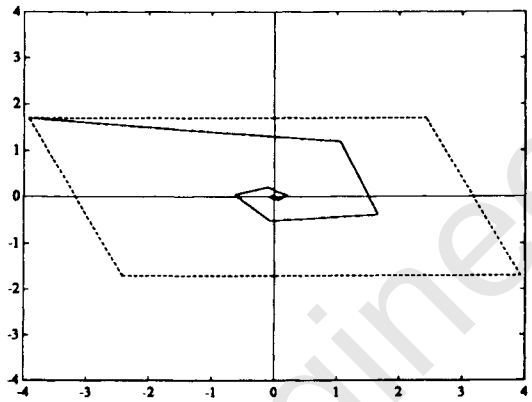


Figure 3: Projected trajectory on the plane (V_2, V_3)

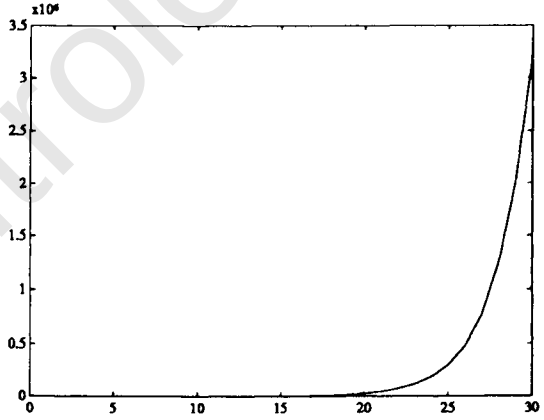


Figure 4: System trajectory projected on V_1

Example (from [34]) : Positive Invariance with global Stability

Now, consider the same system as before, subject to constraints defined by:

$$G = \begin{bmatrix} -0.6538 & 0.7741 & -0.9933 \\ 0.4899 & -0.9626 & -0.8360 \end{bmatrix} \; ; \; \omega = \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix}$$

The two figures (5, 6) show that positive invariance of $R(G,\omega)$ and global asymptotic stability are obtained when selecting

$$F = \begin{bmatrix} 0.5117 & -0.0986 & -1.5022 \\ -0.2111 & -0.2378 & -3.3025 \end{bmatrix}, \text{ which yields :}$$

$$A_0 = \begin{bmatrix} -0.0341 & 0.8624 & -0.9173 \\ -0.1761 & 0.6423 & -1.2219 \\ -0.3449 & 0.4365 & 0.2565 \end{bmatrix} \text{ and } V = \begin{bmatrix} -0.9129 & -0.8081 & -0.7202 \\ -0.0668 & -1.1403 & -0.4641 \\ -0.4580 & -0.3567 & 0.1123 \end{bmatrix}$$

Under this choice, the poles of the closed-loop system are:

$$\begin{cases} \lambda_1 = 0.6647 \quad (\text{ stable zero}) \\ \lambda_1' = 0.1 + j0.6, \\ \lambda_2' = 0.1 - j0.6 \end{cases}$$

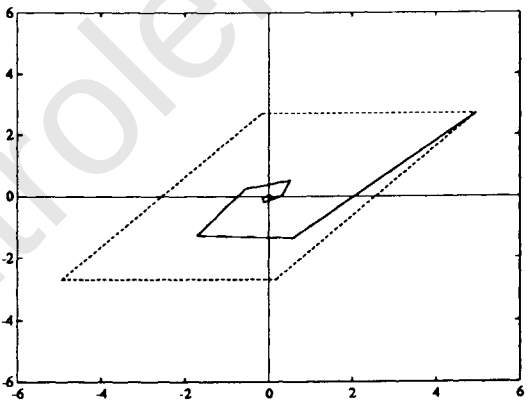


Figure 5: Projected trajectory on the plane (V_2, V_3)

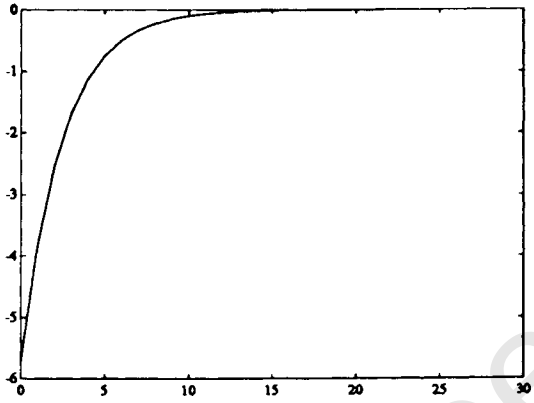


Figure 6: System trajectory projected on V_1

D. Modifications of The Eigenstructure Assignment Technique

1. The case of unstable zeros

As stated in Propositions IV.2 and IV.3, under the assumption rank $G = s \leq m$, $(A + BF)$-invariance of $\mathcal{K}er\ G$ and closed-loop stability cannot be simulataneously obtained if any invariant zero of the triplet (A,B,G) is unstable.

However, the positive invariance approach can also be used in that *non-minimum phase* case if the system is controllable, by imposing positive invariance of a symmetrical polytope Ωincluded in $S(G,\omega)$. Under an appropriate scaling of the row vectors of G, the positive vector ω can be replaced by 1_s. The set of admissible initial states for the resulting linear regulator is then restricted to Ω.

A technique for constructing such a set Ω is described in (Castelan et al. [17]) for the case of continuous-time linear systems. A similar technique can be used for discrete-time systems. The basic principles of this technique derive from the following Proposition :

Proposition IV.6
If the triplet (A, B, G) has unstable invariant zeros, the constrained regulation problem can be solved by restricting the initial states of the system to a positively invariant symmetrical polytope $S(\begin{bmatrix} Q_1 \\ G \end{bmatrix}, 1_n)$, *obtained by adding* $n - s$

independent row vectors to G.

The first step of the method consists of imposing a stable real Jordan form for the closed-loop system. The set of selected closed loop eigenvalues may include the stable zeros of (A, B, G), and the matrix of generalized real eigenvectors can be decomposed as $V = [V_1\ V_2]$. The selected generalized real eigenvectors V_1 and V_2 span two independent complementary subspaces with dimensions s and $n-s$. They are now related with matrix G through the following relation:

$$G[V1\ \ V_2] = [E\ \ I_s]. \tag{104}$$

Contrarily to the case of stable zeros, the row vectors of G cannot be selected as left generalized real eigenvectors of (A+BF). But the matrix of left generalized real eigenvectors can be decomposed as : $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$, with $Q_1 \in \Re^{s\times n}$, $Q_2 \in \Re^{(n-s)\times n}$. It satisfies :

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}(A+BF) = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}. \tag{105}$$

Under the controllability assumption, the eigenvalues of $\Lambda_2 \in \Re^{s\times s}$ and of $\Lambda_1 \in \Re^{(n-s)\times(n-s)}$ are all selected simple. They satisfy the spectral conditions $\mu_i + |\sigma_i| < 1$. Then,

$$\begin{bmatrix} |\Lambda_1| & 0 \\ 0 & |\Lambda_2| \end{bmatrix}\begin{bmatrix} 1_{n-s} \\ 1_s \end{bmatrix} < \begin{bmatrix} 1_{n-s} \\ 1_s \end{bmatrix}. \tag{106}$$

It is possible to construct a non null matrix, $M \in \Re^{s\times(n-s)}$ such that:

$$|M|1_{n-s} + |\Lambda_2|)1_s \leq 1_s \tag{107}$$

Under such an assignment, $S(Q_2, 1_s)$ is a positively invariant symmetrical polyhedron of the closed-loop system. From a matrix M satisfying (107), compute a matrix $E \in \Re^{s\times(n-s)}$ satisfying the equation:

$$E\Lambda_1 - \Lambda_2 E = M \tag{108}$$

This equation can be solved column by column by an assignment type equation: $e_j = (\lambda_{1j} I_s - \Lambda_2)^{-1} m_j$. Relation (108) implies:

$$\begin{bmatrix} I_{n-s} & 0 \\ E & I_s \end{bmatrix}\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 \\ M & \Lambda_2 \end{bmatrix}\begin{bmatrix} I_{n-s} & 0 \\ E & I_s \end{bmatrix} \tag{109}$$

The matrix of generalized real eigenvectors , $V = [V_1\ V_2]$, can be selected so as to satisfy (110).

$$G[V1\ \ V_2] = [E\ \ I_s]. \tag{110}$$

Then, the matrix of left generalized real eigenvectors satisfies: $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} G = [E \;\; I_s] \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$. Relations (105) and (109) yield:

$$\begin{bmatrix} Q_1 \\ G \end{bmatrix} (A + BF) = \begin{bmatrix} \Lambda_1 & 0 \\ M & \Lambda_2 \end{bmatrix} \begin{bmatrix} Q_1 \\ G \end{bmatrix}.$$

This relation, combined with relations (106) and (107) shows the positive invariance of the symmetrical polytope $S(\begin{bmatrix} Q_1 \\ G \end{bmatrix}, 1_n)$ for the closed-loop system.

2. Robustness improvement

In this paragraph, it is now assumed, for simplicity, that rank $G = s \leq m$ and that the triplet (A, B, G) does not have any unstable invariant zero, and that the closed-loop eigenvalues can all be selected simple and satisfying the spectral condition : $\mu_i + |\sigma_i| < 1$. Then, in the absence of uncertainties on the parameters of A and B, positive invariance of $S(G, 1_s)$ is obtained by the eigenstructure assignment technique of section IV.C. positive invariance that the pai To obtain a robust invariant control scheme, it is possible to construct a bounded positively invariant set, Ω, included in $S(G, 1_s)$. To satisfy constraints (60) all along the trajectory, the admissible initial states of the system will then be restricted to Ω. Clearly, a simple way to build such a candidate set Ω is to complete matrix G by $n - s$ independent row vectors, constituting a matrix G', to make up a non-singular matrix in $\Re^{n \times n}$, $R = \begin{bmatrix} G' \\ G \end{bmatrix}$.

The design problem can then be reduced to finding a gain matrix F for which the S.S.P. $S(R, e_n)$ is positively invariant in a robust way. Matrix $G' in \Re^{(n-s)\times n}$ cannot be freely chosen. But, by construction, a candidate matrix G' is available. The matrix of selected generalized real eigenvectors is $V = [V_1 \mid V_2]$. Then, G' can be constructed as the matrix of complementary left generalized real eigenvectors of $(A + BF)$. It satisfies:

$$RV = \begin{bmatrix} G' \\ G \end{bmatrix} [V_1 \mid V_2] = \begin{bmatrix} I_{n-s} & 0_{n-s*s} \\ 0_{s*n-s} & I_s \end{bmatrix} \tag{111}$$

And by the choice of the closed-loop eigenvalues, relation guarantees the positive invariance property of $S(R, 1_n)$ with respect to the closed-loop system (63).

Among the possible choices of a matrix V satisfying the design requirements, it is possible to select the candidate matrix which minimizes the condition number :

$$k(V) = \|V\|_2 \|V^{-1}\|_2.$$

As shown by Kautsky et al [38], the sensitivity of eigenvalues to structured perturbations decreases with $k(V)$. In the design technique described above, the degrees of freedom left to the designer are essentially in the assignment of left generalized real eigenvectors to eigenvalues. Thus, for reasonable size systems, this assignment problem can be solved by an implicit enumeration technique (Hennet et al., [29]).

V. INVARIANT REGULATION UNDER CONTROL CONSTRAINTS

A. The Positive Invariance Approach

This chapter analyzes the case of discrete-time linear multivariable systems with the same state-equation (59) as in chapter IV :

$$x_{k+1} = Ax_k + Bu_k \text{ for } k \in \mathcal{N}\nabla\exists \tag{112}$$

with $x_k \in \Re^n$, $u_k \in \Re^m$, $A \in \Re^{n*n}$, $B \in \Re^{n*m}$, $m \leq n$. But it is now the control vector u_k which is subject to symmetrical linear constraints, put under the generic form :

$$-1_m \leq Su_k \leq 1_m \text{ for } k \in \mathcal{N}, \quad \mathcal{S} \in \Re^{\Updownarrow \times \Updownarrow}, \quad \text{rank}(\mathcal{S}) = \Updownarrow. \tag{113}$$

The state of the system is supposed to be fully observed, and the investigated control law to regulate the system to the zero state is a constant state feedback:

$$u_k = Fx_k \text{ with } F \in \Re^{m*n} \tag{114}$$

In the state space, the input constraints then define the following polyhedral set:

$$S(SF, 1_m) = \{x \in \Re^n \ : \ -1_m \leq SFx \leq 1_m\} \tag{115}$$

Under such control constraints, the stabilizability issue has to be raised. The linear model (63) is valid only for states belonging to $S(SF, 1_m)$.

The approach by positive invariance (M.Vassilaki et al. [30]) consists of finding a state feedback matrix F and a domain in the state space, Ω satisfying the following requirements of *constrained invariant regulation:*

1. $\Omega \subseteq S(SF, 1_m)$
2. $(A + BF)(\Omega) \subseteq \Omega$
3. The eigenvalues $\lambda(A + BF)$ are stable.

For the consistency of the scheme, it is convenient to select as a candidate domain Ω, a symmetrical polyhedral domain $S(G, \omega)$ defined by :

$$S(G, \omega) = \{x \in \Re^n \ : \ -\omega \leq Gx \leq \omega\}. \tag{116}$$

with $G \in \Re^{g*n}$,$m \leq g$, $\omega \in \Re^g$ and $\omega_i > 0$ for $i = 1, ..., g$.

Two cases will be distinguished, depending on the choice of the candidate domain $S(G,\omega)$. In the first case, the domains $S(G,\omega)$ and $S(SF, 1_m)$ are supposed independently chosen. Under this assumption, the domain $S(G,\omega)$ may correspond to a domain of possible initial states or to a domain of state constraints, as in the LCRP (Linear Constrained Regulation Problem) framework studied by M.Vassilaki et al. [30], and by G.Bitsoris et al. [41], [40], [42], [43]. The second case, studied by J.C.Hennet et al. [32], [44], only supposes the existence of control constraints, and aims at maximizing the size of the domain $S(G,\omega)$, which is constructed from $S(SF, 1_m)$. Paragraph V.B analyzes the first case using a Linear Programming approach, and Paragraph V.C. solves the second case, using an eigenstructure assignment technique.

B. The Linear Programming Design Technique

To obtain a simple control scheme in which positive invariance of $S(G,\omega)$ implies robust stability of the closed-loop system (63), $x_{k+1} = (A + BF)x_k$, it suffices to consider the case when $S(G,\omega)$ is a polytope. This polytope can either be given or can be constructed from the domain of state constraints, if it is unbounded, using the results of sections III.D. The algorithm will thus be described for the case $G \in \Re^{n \times n}$, $\text{rank}(G) = n$, $\omega \in \Re^n_+$.

In this case, if it is possible to find a linear feedback control $u_k = F.x_k$ with $F \in \Re^{m*n}$, that guarantees :

- positive invariance of $S(G,\omega)$ with respect to (63),
- inclusion of $S(G,\omega)$ in $S(SF, 1_M)$,

then, this regulator solves the LCRP for any initial state of the system that belongs to $S(G,\omega)$.

The 2 requirements of this control scheme can be characterized by a set of linear equalities and inequalities obtained by application of Proposition II.3 and Lemma II.2 (Extended Farkas Lemma).

From Proposition II.3, positive invariance of $S(G,\omega)$ is equivalent to the existence of a matrix H with non-negative coefficients satisfying :

$$\begin{cases} H.G = G(A + B.F) \\ H.\omega \leq \omega \end{cases}$$

From Lemma II.2, it is easy to derive the following result.

Lemma V.1

If G has a full row-rank (rank (G) = s for $G \in \Re^{s \times n}$), a necessary and sufficient condition for inclusion of $S(G,\omega)$ in $S(F, 1_m)$ is the existence of a matrix

$D \in \Re^{m \times s}$ such that :

$$DG = F \tag{117}$$
$$|D|\omega \leq 1_m. \tag{118}$$

Proof

The proof is similar to the proof of Proposition II.3 in section II.H. From Lemma II.2, $S(G,\omega) \subset S(SF, 1_m)$ is equivalent to the existence of a matrix $\begin{bmatrix} D_1 & D_2 \\ D_3 & D_4 \end{bmatrix}$ satisfying :

$$\begin{bmatrix} D_1 & D_2 \\ D_3 & D_4 \end{bmatrix} \begin{bmatrix} G \\ -G \end{bmatrix} = \begin{bmatrix} F \\ -F \end{bmatrix} \tag{119}$$

$$\begin{bmatrix} |D_1| & |D_2| \\ |D_3| & |D_4| \end{bmatrix} \begin{bmatrix} \omega \\ \omega \end{bmatrix} \leq \begin{bmatrix} 1_m \\ 1_m \end{bmatrix} \tag{120}$$

Necessity of the condition of the Lemma is readily obtained from (119) and (120) by setting $D = D_1 - D_2$ and by using the fact that $D^+ \leq D_1$ and $D^- \leq D_2$. To show that the condition is also sufficient, it suffices to select $D_1 = D_3 = D^+$, and $D_2 = D_4 = D^-$.

□ To solve the LCRP, a LP algorithm can then be constructed by adding to the Linear Program (68), formulated in the case $s = n$, the inclusion relations (117), (118) . A linear representation of constraint (118) is obtained by replacing D by $D_1 - D_2$, with matrices D_1 and D_2 non-negative. The program takes the following form :

$$\begin{array}{lllll} \text{Minimize} & \epsilon & & & \\ \text{under} & (X+Y)\omega - \epsilon\omega & \leq & 0_n & \\ & (X-Y)G - GBF & = & PA & \\ \text{under} & (D_1+D_2)\omega - \epsilon 1_m & \leq & 0_n & \\ & (D_1-D_2)G - F & = & 0_{m \times n} & \\ & \epsilon,\ X,\ Y,\ D_1,\ D_2 & \geq & 0_{s \times s} & \end{array} \tag{121}$$

This formulation can be simplified by eliminating F, using the last equality.

If the optimal solution is such that $\epsilon^* < 1$, $S(G,\omega) \subset S(SF, 1_m)$ and the closed-loop system is asymptotically stable.

If the optimal solution of (121) is such that $\epsilon^* > 1$, then, the three requirements cannot be met together. The problem is then to understand why and to relax some requirements. The eigenstructure assignment approach is then more appropriate for facing this challenge.

To make sure that at the optimum, $X = H^+$, $Y = H^-$, $D_1 = D^+$, $D_2 = D^-$, it suffices to introduce in the objective function some very small positive

weights associated with all the components of matrices X, Y, D_1, D_2.

C. Approach by eigenstructure assignment

Let $P = [V_0 \mid V_I]$ be a matrix of generalized real eigenvectors of A. The $(n-r)$ columns of V_0 span the stable subspace of A in $\Re^n$, and the r columns of V_I span its unstable subspace. The associated real Jordan form is:

$$\begin{bmatrix} J_0 & 0 \\ 0 & J_I \end{bmatrix} = P^{-1}AP \tag{122}$$

Two different eigenstructure assignment schemes will be described for the cases $r \leq m$ and $r > m$.

1. The case $r \leq m$

The following proposition (J.C.Hennet et al. [44]) gives the conditions under which a perfect matching can be obtained between the invariant domain and the domain of constraints ($G = SF$, $\omega = 1_m$).

Proposition V.1
There exists a state feedback gain matrix F which meets requirements 1, 2 and 3 of constrained invariant regulation, with $\Omega = S(SF, 1_m)$, if and only if system (A,B) is stabilizable and the unstable subspace of A has dimension $r \leq m$.

Proof
necessity
Suppose that $S(SF, 1_m)$ is a positively invariant set of the closed-loop system:

$$x_{k+1} = (A + BF)x_k \tag{123}$$

Then, from Proposition II.3, there exists a matrix $H \in \Re^{m \times m}$ such that

$$HSF = SF(A + BF) \tag{124}$$

$$|H|1_m \leq 1_m \tag{125}$$

Relation (124) implies the closed loop invariance (in the sense of W.M.Wonham ([10])) of the subspace $\mathcal{K}er\ F$. The dimension of this uncontrolled subspace is greater or equal to $n - m$. This subspace cannot be included in the stable subspace of A, of dimension $n - r$ if the condition $r \leq m$ is not satisfied. This condition is therefore necessary to fulfill both requirements 2 and 3 with $\Omega = S(SF, 1_m)$.

sufficiency

This part of the proof is constructive. It describes the basic algorithm for obtaining the positive invariance of a given symmetrical polyhedron by eigenstructure assignment.

The stable subspace of A is kept unchanged by the chosen feedback law. Only the unstable eigenvalues need to be moved to stabilize the system. Then, the gain matrix F satisfies: $\text{span}(V_0) = \mathcal{K}er\ F$. The selected spectrum of the restriction of $(A + BF)$ to $(\frac{\Re^n}{\mathcal{K}er\ F})$ is a set of r symmetric complex numbers, $\lambda_i = \mu_i + j\sigma_i$ such that:

$$|\mu_i| + |\sigma_i| < 158 \tag{126}$$

The r new eigenvalues λ_i are selected distinct from the stable eigenvalues of A. A set of eigenvectors of (123) associated with these eigenvalues can be obtained by solving, for $i = 1, ..., r$, the equations:

- for $\lambda_i \in \Re$ $(\sigma_i = 0)$:

$$v_i = (\lambda_i I_n - A)^{-1} B w_i \tag{127}$$

$$S w_i = e_i \text{ with } e_i \in \Re^r \tag{128}$$

- and for $\sigma_i \neq 0$:

$$v_i = (\lambda_i I_n - A)^{-1} B w_i \tag{129}$$

$$v_{i+1} = (\lambda_i^* I_n - A)^{-1} B w_{i+1} \tag{130}$$

$$S\frac{w_i + w_{i+1}}{2} = e_i \tag{131}$$

$$S\frac{w_i - w_{i+1}}{2j} = e_{i+1} \tag{132}$$

Relation (127) expresses the fact that vector v_i belongs to the transmission subspace of λ_i. Under an appropriate choice of e_i, which is always possible (see e.g. N. Karcanias, B. Kouvaritakis [35]), $v_i \notin \text{range}\ (B)$. Then, the set of vectors v_i for $i = 1, ..., r$ is independent. It spans an (A,B) invariant subspace of $\Re^n$ which is complementary of range(V_0) in $\Re^n$. In general, a set of independent vectors (v_i) solutions of (127), (128) or of (129), (131) (130), (132) is obtained when choosing $(e_1, ..., e_r) = I_r$, the identity matrix. We then obtain the relation:

$$SW_S = \begin{bmatrix} 0_{(m-r)\times r} \\ I_r \end{bmatrix} \tag{133}$$

The generalized real eigenvectors associated with $(v_1, ..., v_r)$ are the columns of a matrix $V_S \in \Re^{n\times r}$. Their corresponding real inputs are the columns of $W_S \in \Re^{m\times r}$ computed by:

$$W_S = S^{-1} \begin{bmatrix} 0_{(m-r)\times r} \\ I_r \end{bmatrix}. \tag{134}$$

The state feedback gain matrix providing this assignment can be written:

$$F = WV^{-1} \text{ with } W = [0 \mid W_S]; \quad V = [V_0 \mid V_S] \tag{135}$$

This gain matrix has rank r. By construction, it satisfies:

$$SF = \begin{bmatrix} 0 \\ \mathcal{F} \end{bmatrix}, \text{ with } \begin{cases} \mathcal{F}V_S &= I_r \\ \mathcal{F}V_0 &= 0_{r\times(n-r)} \end{cases} \tag{136}$$

Under the feedback gain matrix of relation (135), The real Jordan form of the closed-loop system is:

$$J = \begin{bmatrix} J_0 & 0_{(n-r)*r} \\ 0_{r*(n-r)} & J_S \end{bmatrix} = V^{-1}(A + BF)V \tag{137}$$

By construction, from (135), (136), (137),

$$SF(A+BF)[V_0 \mid V_S] = H \text{ with } H = \begin{bmatrix} 0_{(m-r)\times(n-r)} & 0_{(m-r)\times r} \\ 0_{r\times(n-r)} & J_S \end{bmatrix}. \tag{138}$$

The matrix of left generalized real eigenvectors corresponding to matrix $V = [V_0 \mid V_S]$ is:

$$Q = \begin{bmatrix} D \\ \mathcal{F} \end{bmatrix} \tag{139}$$

with $D \in \Re^{(n-r)\times n}$ defined by $D = [I_{n-r} \mid 0_{(n-r)\times r}[V_0 \mid V_S]^{-1}$.
The row vectors of matrix Q are a set of left generalized real eigenvectors of $(A + BF)$. It satisfies $QV = VQ = I_n$. Right multiplication of the terms of (138) by matrix Q yields:

$$\mathcal{F}(A + BF) = J_S\mathcal{F}. \tag{140}$$

Furthermore, the structure of J_S and the spectral relations (58) imply that matrix $|J_S| - I_m$ is an M-matrix satisfying (A.Benzaouia, C.Burgat [23], G.Bitsoris [14]) :

$$(|J_S| - I_m)1_m \leq 0_m \tag{141}$$

Relations (140) and (141) are precisely the two basic conditions of positive invariance of $S(\mathcal{F}, 1_m)$ for the closed-loop system (63). And from (136), it is clear that $S(SF, 1_m) \equiv S(\mathcal{F}, 1_m)$.
□

2. The case $r > m$

From Proposition V.1, positive invariance of $S(SF, 1_m)$ cannot be obtained

with closed-loop stability if the number of inputs, m, is strictly smaller than the dimension of the unstable subspace of A, $\mathbf{r}$. But the three requirements of an admissible invariant control can still be met by constructing a positively invariant polyhedron strictly included in $S(SF, 1_m)$. This point will now be explained in a constructive way in the proof of the following Proposition.

Proposition V.2:
If the system is stabilizable and $r > m$, it is possible to construct a state feedback matrix $F \in \Re^{m\times n}$ and a matrix $\Phi \in \Re^{r\times n}$ such that $\Omega = S(\Phi, 1_r)$ satisfies the three requirements of constrained invariant regulation.

Proof :
All the unstable eigenvalues of A are controllable and can be shifted to the region of the complex plane included in the unit circle and bounded by constraints (58). The new eigenvalues are selected simple, distinct from each other and from the stable eigenvalues of A. As in the preceding section, the eigenstructure of the stable subspace of A is kept unchanged by the state feedback. The real Jordan form of $(A + BF)$ takes the form $\begin{bmatrix} J_0 & 0 \\ 0 & J_S \end{bmatrix}$.
$V = [V_0 \mid V_S]$ is the matrix of closed-loop generalized real eigenvectors, and $W = [0_{m*(n-r)} \mid W_S]$ is the matrix of corresponding "real inputs".
The matrix of left generalized real eigenvectors , $\bar{G} = V^{-1}$, can be decomposed as: $\bar{G} = \begin{bmatrix} G_0 \\ G \end{bmatrix}$. Matrix $G \in \Re^{r\times n}$ has rank r. It is associated with J_S, and generates the positively invariant symmetrical polyhedron $S(G, 1_r)$ for the closed-loop system (63).
By construction, the following relations are satisfied:

$$J_S G = G(A + BF) \tag{142}$$

$$(|J_S| - I_r)1_r \leq 0 \tag{143}$$

Now, partition matrices $W_S \in \Re^{m\times r}$ and $V_S \in \Re^{n\times r}$ as follows:

$$W_S = [W_1 \; W_2] \;, \quad \text{with } W_1 \in \Re^{m\times m} \;, \; W_2 \in \Re^{m\times(r-m)}$$

$$V_S = [V_1 \; V_2] \;, \quad \text{with } V_1 \in \Re^{n\times m} \;, \; V_2 \in \Re^{n\times(r-m)}$$

Matrix G is accordingly decomposed as: $G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}$ with $G_1 \in \Re^{m\times n}$, $G_2 \in \Re^{(r-m)\times n}$. The state feedback gain matrix satisfies:

$$F = [0 \; W_1 \; W_2][V_0 \; V_1 \; V_2]^{-1} = [W_1 \; W_2] \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \tag{144}$$

The associated real Jordan canonical form of $(A+BF)$ is : $J = \begin{bmatrix} J_0 & 0 & 0 \\ 0 & J_1 & 0 \\ 0 & 0 & J_2 \end{bmatrix}$,
where $J_0 \in \Re^{(n-r)\times(n-r)}$, $J_1 \in \Re^{m\times m}$, $J_2 \in \Re^{(r-m)\times(r-m)}$.

By convention, the blocks of J_1, J_2 can be ordered in the increasing order of $\gamma_i = |\mu_i| + |\sigma_i|$. The maximal value of γ_i is denoted γ_{max}. It is strictly smaller than 1 since it respects constraint (58).
Input directions of matrices $(W_1,\ W_2)$ are now selected so that the the closed-loop system (63) also admits as a positively invariant symmetrical polyhedron a polyhedron $S(\Phi, 1_r)$ included in $S(SF, 1_m)$. Define:

$$\Phi = \begin{bmatrix} SF \\ G_2 \end{bmatrix} = \begin{bmatrix} SW_1 & SW_2 \\ 0 & I_{r-m} \end{bmatrix} \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \tag{145}$$

By construction, the symmetrical polyhedron $S(\Phi, 1_r)$ is included in $S(SF, 1_m)$ and exactly fits into its facets. This polyhedron has $2m$ of its parallel facets generated by matrix SF.
The selected matrix W_1 of input vectors associated with J_1 satisfies $SW_1 = I_m$, so as to obtain:

$$SW_1 J_1 = J_1 SW_1. \tag{146}$$

The eigenvalues of J_S are all distinct and satisfy (58). Then, $(|J_1| - I_m)1_m < 0_m$, and it is always possible to construct a full rank matrix $K \in \Re^{m \times (r-m)}$ such that:

$$(|J_1| - I_m)1_m + |K|1_{r-m} \leq 0_r \tag{147}$$

For that, it suffices to choose the elements of K such that:

$$\sum_{j=1}^{r-m} |k_{lj}| \leq 1 - |\mu_l| - |\sigma_l|)\ , \quad \text{for } l = 1, \ldots, m. \tag{148}$$

The closed-loop dynamics in $(\frac{\Re^n}{Ker\ G})$ are described by : $\mathcal{H} = \begin{bmatrix} J_1 & K \\ 0 & J_2 \end{bmatrix}$.
The column-vectors of SW_2 can be computed from

$$SW_2 J_2 = J_1 SW_2 + K. \tag{149}$$

In particular, if its associated eigenvalue λ_{2j} is real, the jth column of SW_2, denoted y_j can be computed from the jth column of K, k_j by:

$$y_j = (\lambda_{2j} I_m - J_1)^{-1} k_j \text{ for } j = 1, ..., m. \tag{150}$$

From relations (146), (149), we obtain:

$$\begin{bmatrix} SW_1 & SW_2 \\ 0 & I_{r-m} \end{bmatrix} \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} = \begin{bmatrix} J_1 & K \\ 0 & J_2 \end{bmatrix} \begin{bmatrix} SW_1 & SW_2 \\ 0 & I_{r-m} \end{bmatrix} \tag{151}$$

The column vectors of matrices V_1 and V_2 are computed from the column vectors of matrices W_1 and W_2 by equations (127). Then, apply the feedback gain matrix (144) to obtain the positive invariance of $S(\Phi, 1_r)$, which directly derives from the properties of matrix $\mathcal{H}$:

$$\Phi(A + BF) = \mathcal{H}\Phi \tag{152}$$

$$(|\Phi| - I_r)1_r \leq 0_r. \tag{153}$$

□

3. The domain of admissible initial states

Under the schemes of this section, the domain of states for which the linear state feedback $u_k = Fx_k$ satisfies the constraints, is the symmetrical polyhedron $S(\Phi, 1_r)$, with $\Phi = \mathcal{F}$ in the case $r \leq m$. To obtain a robust scheme relatively to structured perturbations on matrices A and B, it suffices to bound the domain not only in the directions of the range of V_S, but also in the range of V_0 (J.C.Hennet,E.B.Castelan [44]). In particular, if all the eigenvalues of J_0 are simple and satisfy the spectral condition (58), a positively invariant bounded and admissible polytope of the closed-loop system (63) is:

$$S(Q, 1_n) = \{x \in \Re^n \ : \ -1_n \leq Qx \leq 1_n\} \text{ with } Q = \begin{bmatrix} D \\ \Phi \end{bmatrix}, \qquad (154)$$

where the rows of D are the left generalized real eigenvectors of $(A + BF)$ associated with J_0. The vector of bounds related to D, 1_{n-r}, is arbitrarily chosen. It can be multiplied by any positive number without breaking up the positive invariance property, at least for the unperturbed controlled system. The volume of the polytope $S(Q, 1_n)$ is a good measure of the quality of the eigenstructure assignment from the two following viewpoints [44]:

- Maximization of the size of the domain of admissible initial states.
  As mentionned before, this size is arbitrary large in the range of V_0, but it has to be maximized in the other directions.

- Robustness of the assignment.
  For vectors of given norms, the maximal volume is obtained for vectors which are as close as possible to orthogonality. For a set of eigenvectors, this property precisely characterizes the minimal sensitivity of eigenvalues relatively to structured perturbations (J.Kautsky et al. [38]).

The volume of $S(Q, 1_n)$ is proportional to the absolute value of the determinant of $V = [V_0 \mid V_S]$. In the proposed assignment scheme, the set of desired eigenvalues of J_S is supposed to be given. The only degrees of freedom are in the selection and assignment of the r eigenvalues, λ_i to the selected input directions. The best possible assignment can be obtained by implicit enumeration, with $\mathrm{Det}(V)$ as the function to be minimized.

4. Example (from [44])

Consider a third order system with dynamic matrix and input matrix:

$$A = \begin{bmatrix} 0.125 & -1.375 & 0.375 \\ -2.500 & -0.500 & 2.500 \\ 0.625 & 1.125 & -0.125 \end{bmatrix}, \quad B = \begin{bmatrix} 5.00 & -1.00 \\ 1.00 & 2.00 \\ 1.00 & 0.00 \end{bmatrix}$$

The open-loop system has two unstable eigenvalues, $(-3.0\ ,\ 2.0\ )$, and one stable, 0.5. The input vector is subject to constraints:

$$-1_2 \leq Su_k \leq 1_2 \text{ with } S = \begin{bmatrix} 1.00 & -1.00 \\ 2.00 & 1.00 \end{bmatrix}.$$

In this example, $r = m = 2$. The selected control law leaves the stable pole, $\lambda_1 = 0.5$, unchanged, and moves the two unstable eigenvalues of A to the pair of complex conjugate eigenvalues $0.4 \pm 0.4j$. These eigenvalues satisfy relation (58). If we select $\begin{cases} \lambda_2 = 0.4 - 0.4j \\ \lambda_3 = 0.4 + 0.4j \end{cases}$, and S^{-1} as the matrix of real input vectors associated with these last two eigenvalues, the matrix of closed-loop generalized real eigenvectors becomes:

$$V = \begin{bmatrix} 0.707 & -3.131 & 2.359 \\ 0.0 & 0.993 & 0.087 \\ 0.707 & -2.359 & 1.832 \end{bmatrix} \text{ under } F = \begin{bmatrix} 0.511 & 0.733 & -0.511 \\ 0.658 & -0.160 & -0.658 \end{bmatrix}.$$

The matrix of the closed-loop system takes the form: $\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.4 & -0.4 \\ 0 & 0.4 & 0.4 \end{bmatrix}$.

The basic eigenstructure assignment technique of section V.C.1 guarantees positive invariance of $S(SF, 1_2)$ for the perfectly deterministic system (123). Any point of this domain is asymptotically driven to the zero state by the control law $u_k = Fx_k$ without ever violating the control constraints. Now, if the system parameters are slightly uncertain, the requirements of constrained invariant regulation will be satisfied in a robust way by the same control law if the domain of admissible initial states is restricted to $S(Q_{eta}, 1_3)$ with :

$$Q_\eta = \begin{bmatrix} \eta D \\ SF \end{bmatrix} \text{ and } Q_1 = \begin{bmatrix} D \\ SF \end{bmatrix} = V^{-1},$$

Then, $D \simeq [-4.84 \quad -0.40 \quad 6.26]$, and the positive weighting term η can be selected small if the system parameters are almost perfectly known. Fig. 7 represents the admissible domain of initial states and an admissible trajectory, in projection on the plane orthogonal to the first column vector of V.

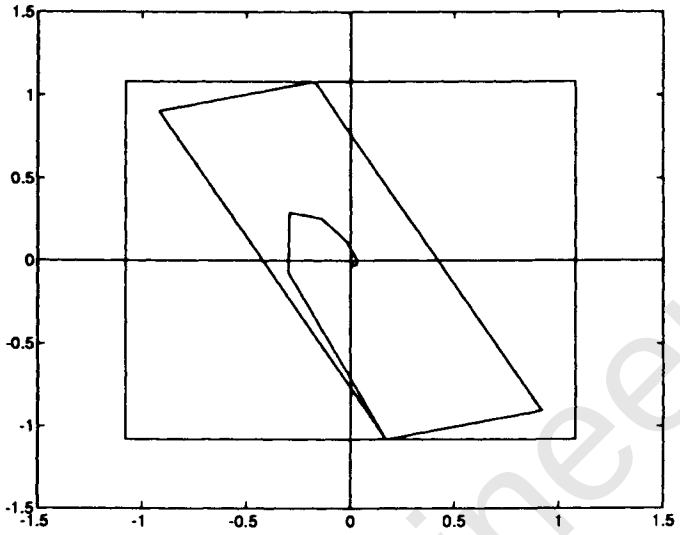


Figure 7: Projected positively invariant domain

VI. CONCLUSION

The positive invariance approach has been described as a valuable alternative to optimal control for rigorously solving regulation problems for systems subject to constraints on their state, output or control vector. Its main advantage is the obtention of closed-form solutions which are of the state-feedback type and can be easily computed using standard tools of linear algebra and software packages such as Matlab.

The homothesis property of positively invariant domain can also be used to recompute the control law from the knowledge of the current state of the system. Such a recursive computation - not studied in this paper - can improve the system dynamics and can be integrated in adaptive versions of the proposed algorithms.

The study has been presented in the discrete-time framework. This does not mean that its theory and algorithms do not apply to continuous-time systems. On the contrary, all the results of this paper have their counterpart in the continuous-time framework. From the obtention of positive invariance conditions for continuous-time linear systems by Bitsoris and Vassilaki [45], [46] and by Castelan, Hennet [17], [37], algorithms have been constructed for solving constrained control problems in the continuous-time framework. As for the discrete-time case, the two main design techniques are based on Linear

Programming and on Eigenstructure Assignment.

Positive invariance of closed domains also provides an interesting intermediate step between the more theoretical geometric approach which analyzes the invariance properties of subspaces, and the more practical devices elaborated by control engineers to respect practical constraints. In this respect, practical applicability of the proposed control schemes has been noticeably increased by some robustness considerations. In particular, the simple rule that consists of always bounding (in a proper way) the domain of admissible initial states is a guarantee to obtain a closed-loop system behavior robust to small perturbations or (and) uncertainties.

References

[1] E.G.Gilbert : "Linear Control Systems With Pointwise-in-Time Constraints : What Do We Do About Them ?", *1992 ACC, Chicago* , vol. 4, p.2565.

[2] J.M.Dion,L.Dugard,N.M.Tri : "Multivariable Adaptive Control with Input-Output Constraints" *CDC Los Angeles 1987*, pp. 1233-1238.

[3] C.E. Garcia, D.M. Prett and M. Morari : "Model Predictive Control and Practice : A Survey ", *Automatica* vol. 25, pp.335-348, 1989.

[4] J.P.Béziat -J.C.Hennet : " Generalized Minimum Variance Control of Constrained Multivariable Systems", *Intl Journal of Modelling and Simulation*, vol. 9, No.3, 1989, pp.79-84.

[5] J-P. Gauthier, G. Bornard : "Commande Multivariable en Pre'sence de Contraintes de Type Inégalité ",*RAIRO*, vol. 17, No. 13, 1983, pp. 205-222.

[6] J.C. Hennet, M. Vassilaki : "Modélisation et Gestion de Systèmes Multidimensionnels de Production avec Stockages", *RAIRO-APII*, vol. 21, No.3, 1987, pp. 3-16.

[7] P.J. Campo, M. Morari and C.N. Nett : "Multivariable Anti-windup and Bumpless Transfer : A General Theory ", *ACC 1989*, Pittsburgh, pp.1706-1711.

[8] M.Sznaier, Z.Benzaid " Robust control of systems under mixed time/frequency domain constraints via convex optimization" *IEEE CDC, Tucson* , 1992, pp. 2617-2622.

[9] M. Sznaier : " A Set Induced Norm Approach to the Robust Control of Constrained Systems ", *SIAM J. Control and Optimization*, vol. 31, No.3, 1993, pp.733-746.

[10] W.M.Wonham : *Linear multivariable control - A geometric approach.* Springer-Verlag 1985.

[11] H.Nijmeijer : "Controlled invariance for affine control systems", *Int. J. Control*, vol 34 No 4 (1981) pp.825-833.

[12] J.Cheganças, C.Burgat : "Régulateur P-invariant avec Contraintes sur la Commande", *AFCET Congress, Toulouse, France*, 1985, pp.193-203.

[13] C.Burgat - A.Benzaouia: "Stability properties of positively invariant linear discrete time systems ", *Journal of Mathematical Analysis and Applications*, vol. 143,No 2, (1989), pp. 587,596.

[14] G.Bitsoris : "Positively invariant polyhedral sets of discrete-time linear systems", *Int. Journal of Control*, vol 47 (1988), pp. 1713-26.

[15] G.Bitsoris : "On the positive invariance of polyhedral sets for discrete-time systems", *Systems and Control Letters*, 11 (1988),pp.243-248.

[16] J.C. Hennet and J.P. Beziat, "Invariant Regulators for a class of constrained linear systems", *Automatica,*, vol. 27, No.3, pp.549-554, 1991 ; also in *IFAC, Tallinn*, vol. 2, pp.299-304, 1990.

[17] E.B. Castelan and J.C. Hennet, "Eigenstructure assignment for state-constrained linear continuous-time systems" , *Automatica*, vol.28, No.3, pp. 605-611, 1992 ; also in *IFAC Symposium , Zurich*, 1991.

[18] A. Schrijver, *Theory of linear and integer programming.* John Wiley and Sons, 1987.

[19] A. Berman, R.J. Plemmons : *Non Negative Matrices in the Mathematical Sciences* Academic Press, 1979.

[20] J.C.Hennet : "Une extension du Lemme de Farkas et son application au problème de régulation linéaire sous contraintes." *Comptes-Rendus de l'Académie des Sciences*, t.308, Série I, pp.415-419, 1989.

[21] J.Chéganças : "Sur le concept d'invariance positive appliqué à l'étude de la commande contrainte des systèmes dynamiques ", *Doctorate of University Paul Sabatier, Toulouse, France*, 1985.

[22] S.Tarbouriech, C.Burgat, "Note on stability properties of some discrete-time systems." *to appear in J. Math Anal and Applications*, 1994, also note LAAS-CNRS No 93259, 1993.

[23] A. Benzaouia and C. Burgat, *Regulator problem for linear discrete-time systems with non symmetrical constrained control*, Int. J. Control, vol. 48, pp.2441-56, 1988.

[24] A. Brøndsted : *An introduction to convex polytopes.* Springer Verlag (1983)

[25] R.E.Kalman - J.E.Bertram : "Control systems analysis and design via the second method of Lyapunov", *Trans. A.S.M.E, D 82* (1960), pp.394-400.

[26] I.Glazman, Y.Liubitch : *Analyse linéaire dans les espaces de dimension finie.* Editions MIR, Moscou (1974).

[27] H.N.Rosenbrock : *A method of investigating stability.* IFAC proceedings, 1963, pp 590-594.

[28] J.C.Hennet - J.B.Lasserre : "Construction of Positively Invariant Polytopes for Stable Linear Systems *IFAC 12th World Congress, Sydney, 1993*, vol.9, pp.285-288.

[29] J.C. Hennet and E.B.Castelan : "Robust invariant controllers for constrained linear systems", *1992 American Control Conference, Chicago* Vol.2, pp.993-997, 1992.

[30] M.Vassilaki - J.C.Hennet - G.Bitsoris : "Feedback control of linear discrete-time systems under state and control constraints," *Int. Journal of Control*, vol 47 (1988), pp. 1727-35.

[31] A.G.J.Mac Farlane and N.Karcanias : "Poles and zeros of linear multivariable systems: a survey of the algebraic, geometric and complex-variable theory", *Int. J. Control*, vol.24 No 1 (1976) pp.33-74.

[32] J.C.Hennet - E.B.Castelan : "A unified framework for linear constrained regulation problems" in *Mathematics of the Analysis and Design of Process Control, North-Holland, 1992*, pp.337-346, also in *IMACS-MCTS*, Lille, vol.2, pp. 295-300, 1991.

[33] B.Kouvaritakis - A.G.J.Mac Farlane : "Geometric approach to analysis and synthesis of system zeros: parts 1 and 2", *Int. J. Control*, vol.23 No 2 (1976), pp.149-181.

[34] J.C.Hennet - E.B.Castelan : "Invariance and Stability by state-feedback for constrained linear systems" *ECC 1991, Grenoble*, vol.1, pp.367-373.

[35] N.Karcanias, B. Kouvaritakis : "The output zeroing problem and its relationship to the invariant zero structure : a matrix pencil approach", *Int. J. Control*, vol.30, No 3, pp.395-415, 1979.

[36] B.C.Moore : "On the flexibility offered by state feedback in multivariable systems beyond closed-loop eigenvalue assignment", *IEEE Trans on Automatic Control*, October 1976, pp.689-692.

[37] E.B.Castelan and J.C. Hennet : "On Invariant Polyhedra of Continuous-time Linear Systems", *IEEE Trans. Automatic Control*, vol. 38, No 11, pp. 1680-1685, 1993.

[38] J.Kautsky, N.K.Nichols, P.Van Dooren 1985 : "Robust pole assignment in linear state feedback", *Int. J. Control* vol.41, No.5, pp.1129-1155.

[39] G.Bitsoris, M.Vassilaki : "The Linear Constrained Regulation Problem for Discrete-time Systems", *11th IFAC World Congress, Tallinn, 1990*, vol.2, pp. 287-292.

[40] G.Bitsoris, M.Vassilaki : "Optimization approach to the linear constrained regulation problem for discrete-time systems", *Intl. J. Systems Sci.*, 1991, vol.22, No.10, pp.1953-1960.

[41] G.Bitsoris, M.Vassilaki : "The Linear Constrained Regulation Problem for Discrete-time Systems", *11th IFAC World Congress, Tallinn, 1990*, vol.2, pp. 287-292.

[42] E.Gravalou, G.Bitsoris : "Constrained Regulation of Linear Systems : An Algorithm", *30th IEEE-CDC, Brighton, 1991*, pp.1744-1747.

[43] G.Bitsoris, M.Vassilaki : "Design Techniques of Linear Constrained Discrete-Time Control Systems", in *Advances in Control and Dynamic Systems, C.T.Leondes Editor, Academic Press*, vol.56, 1993.

[44] J.C. Hennet and E.B.Castelan, "Constrained Control of Unstable Multivariable Linear Systems" *1993 European Control Conference , Groningen, Netherlands*, Vol.4, pp.2039-2043, 1993.

[45] M.Vassilaki and G. Bitsoris "Constrained Regulation of Linear Continous-time Dynamical Systems" *Systems and Control Letters, vol. 13*, 247-252 (1989).

[46] G.Bitsoris : "Existence of positively invariant polyhedral sets for continuous-time linear systems", *Control Theory and Advanced Technology*, vol 7, No.3 (1991), pp. 407-427.

This Page Intentionally Left Blank

Digital Control with H_∞ Optimality Criteria

Hannu T. Toivonen

Process Control Laboratory
Department of Chemical Engineering
Åbo Akademi University
20500 Turku/Åbo, Finland

I. INTRODUCTION

It is well known that the H_∞ control problem [1, 2] plays a central role in a wide range of robust and worst-case control issues. In particular, it is related to the robust control of systems with unstructured frequency-domain uncertainties. The H_∞ problem is today well established, and efficient state-space solution procedures exist for both continuous-time and discrete-time systems [3, 4].

In control system implementations it is common to apply digital control to plants which operate in continuous time. The controller then consists of a sampler, a discrete control law which processes the sampled output, and a hold element, which takes the discrete control sequence to a continuous-time signal. The question then arises how the sampler and hold elements affect the robustness properties of the closed-loop system. Clearly, the hybrid discrete/continuous-time nature of the closed-loop system is not correctly taken into account by standard continuous and discrete analysis and synthesis techniques. Therefore, it has been common in the literature to apply various approximative methods to the problem of designing digital controllers in the context of robust controller synthesis. One approach is to make the design in continuous time using well-established techniques,

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

and to discretize the controller obtained in this way, for example using bilinear transformation [5]. This approach is clearly limited to cases when the sampling frequency is high compared to the design bandwidth [5]. An alternative approximation technique is to discretize the continuous-time plant, and to apply standard discrete synthesis methods to the discrete process model [5, 6]. The main difficulty with this approach is to perform the discretization in such a way that various assumptions on the continuous-time plant concerning model uncertainties and performance specifications are preserved in the discretization process.

The limitations of standard continuous and discrete design methods in the treatment of sampled-data control systems have recently led to the development of a robust control theory for sampled-data control systems. In this line of work, Francis and Georgiou [7] and Chen and Francis [8] have studied the L_p-stability of sampled-data systems. Chen and Francis [9], Kabamba and Hara [10], and Sivashankar and Khargonekar [11, 12] have given methods for computing the L_2-induced norm of sampled-data control systems, and Leung *et al.* [13] have studied the performance of sampled-data control systems for bandlimited disturbances. An important problem in a robust control theory for sampled-data systems is the generalization of the H_∞ control problem to the sampled-data case [14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. The sampled-data H_∞ control problem consists of the minimization of the L_2-induced norm of the closed-loop system. This problem has been shown to be relevant to the robustness of sampled-data control systems, and it provides a basis for dealing with both unstructured and structured plant uncertainties in digital control [24, 25, 26]. A theory for the L_∞-induced performance in sampled-data systems has also been developed [27, 28, 29, 25].

In this contribution, we focus on the H_∞ sampled-data control problem. The main nonstandard feature of the sampled-data H_∞ problem is the fact that the closed-loop system is a hybrid discrete/continuous system, part of which (the plant) evolves in continuous time, while another part (the controller) evolves in discrete time. The induced norm that is minimized is based on the continuous-time input and output signals. In the recent literature, a number of solution approaches to the sampled-data H_∞ control problem have emerged. One is based on the 'lifting technique' [14, 15, 30, 22], in which the sampled-data system is represented as a discrete system with a finite-dimensional state, but with inputs and outputs which take values in infinite-dimensional spaces. This is due to the fact that the inputs and outputs of the system between the sampling instants $\{kh\}$ take values in the infinite-dimensional space $L_2[0, h]$. The sampled-data problem can then be solved via a standard finite-dimensional discrete H_∞ problem. A second approach solves the problem using methods based on

dynamic game theory [31, 32, 11, 23, 19]. In this approach the game-theory solutions of standard continuous and discrete H_∞ problems [33] are generalized to the sampled-data problem. The problem is then formulated as a hybrid discrete/continuous dynamic game, in which one player (the disturbance) acts in continuous time, whereas the other player (the controller) is restricted to act in discrete time. The sampled-data H_∞ problem has also been solved using a Hamiltonian optimal control approach [16, 17, 20, 21], similar to the time-domain solution of the H_∞ problem presented in [34]. A common feature of the solution methods for the sampled-data H_∞ problem is that all procedures lead to a problem representation in terms of an equivalent finite-dimensional discrete H_∞ problem. The solution procedures have been derived independently, and the connection between the resulting discrete problem representations is therefore not easy to see. Later it has been shown, however, that the discretization can be treated in a unified framework [35], which clarifies the connections between the various solution procedures.

The purpose of this contribution is to provide a tutorial presentation of various approaches for solving the sampled-data H_∞ control problem. The relevance of the problem to the robustness issue in sampled-data control is also discussed. The material is organized as follows. In Section II, the problem is stated. The solution of the sampled-data H_∞ problem using the lifting technique formalism and game theory approach, respectively, is discussed in Sections III and IV, respectively. In Section V a worst-case sampling approach to the problem is presented, which clarifies the connection between various solution approaches. Finally, some special topics are discussed. In Section VI the game-theoretic approach is applied to solve a dual-rate sampled-data H_∞ control problem, in which the sampling and hold elements operate at different rates. In Section VII the design of optimal sampling prefilters is considered in the context of sampled-data H_∞-optimal control. Finally, the relevance of the sampled-data H_∞ problem to the robustness issue is discussed in Section VIII.

Although most of the material covered here is available in the literature and the discussion is mainly tutorial in nature, some of the issues dealt with are new. In particular, the solution of the dual-rate sampled-data H_∞ control problem described in Section VI does not appear to have been presented before. In Section VIII, previously presented necessary and sufficient conditions for robust stability of sampled-data control systems are generalized to nonlinear time-invariant uncertainties.

II. PROBLEM FORMULATION

We study a linear finite-dimensional continuous-time plant described by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B_1 w(t) + B_2 u(t) \ , \ x(0) = 0 \\ z(t) &= C_1 x(t) + D_{12} u(t) \\ y(t) &= C_2 x(t) \end{aligned} \tag{1}$$

where $x(t) \in R^n$ is the state vector, $w(t) \in R^{m_1}$ is the disturbance, $u(t) \in R^{m_2}$ is the control signal, and $z(t) \in R^{p_1}$ and $y(t) \in R^{p_2}$ are the controlled and measured outputs. The system (1) is assumed to include all weighting filters as well as anti-aliasing filters. For simplicity and clarity, the system matrices are assumed time-invariant unless otherwise stated. It is, however, straightforward to generalize the results to time-varying systems.

The measurements are assumed to be available at discrete sampling instants $\{kh\}$, and may be corrupted by noise, i.e.,

$$\hat{y}_k = y(kh) + \hat{D}_{21} \hat{v}_k \tag{2}$$

where $h > 0$ is the sampling time and $\hat{v}_k \in R^{m_3}$ is a discrete measurement disturbance. Here, a circumflex ($\hat{}$) has been used to indicate signals and matrices which relate to the discrete-time part of the control system.

Part of the discussion becomes significantly simpler if some special assumptions on the system matrices are made. We list these simplifying assumptions here for future reference:

(A1) D_{12} has full column rank,

(A2) $\hat{D}_{21}$ has full row rank.

The control signal is assumed to be generated digitally using a hold device. Usually a zero-order hold is assumed, i.e. the control signal is taken piecewise constant between the sampling instants,

$$u(t) = \hat{u}_k \ , \ t \in [kh, kh + h). \tag{3}$$

We consider discrete controllers of the form

$$\hat{u} = \mathcal{K} \hat{y}, \tag{4}$$

where $\mathcal{K}$ is a causal discrete operator. In analogy with standard H_∞ control problems, a worst-case performance measure induced by L_2/l_2-disturbances is considered. Here the performance measure is taken as the induced norm of the closed-loop system defined as

$$J(\mathcal{K}) := \sup \Big\{ \frac{\|z\|_{L_2}}{[\|w\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2]^{1/2}} \Big\}, \tag{5}$$

where the supremum is taken over $(w, \hat{v}) \neq (0,0)$ in $L_2 \oplus l_2$. Both infinite-horizon and finite-time problems will be considered. Depending on the context, the signals are then taken to belong to the spaces $L_2[0,\infty)$ and $l_2(0,\infty)$ in the infinite-horizon case, and to $L_2[0,Nh]$ and $l_2(0,N)$ in the finite-time case.

In some studies on the sampled-data H_∞ problem zero measurement noise has been assumed. Then (5) reduces to the L_2-induced norm of the closed-loop system. As the results readily generalize to the case with measurement noise, we choose in this contribution to include it. The case with zero measurement noise can be recovered by setting $\hat{D}_{21} = 0$. Note, however, that the simplifying assumption (A2) does not hold in this case, and some of the formulas which are given take a different form when the assumption is relaxed.

The sampled-data H_∞ problem consists of the problem of minimizing $J(\mathcal{K})$. More precisely, we consider the problem of finding a digital controller such that

$$J(\mathcal{K}) < \gamma \tag{6}$$

holds for a specified positive constant γ. In analogy with other H_∞ problems, the performance measure (5) can then be minimized by checking whether (6) has a solution for successively smaller values of γ, a procedure known in the literature as 'γ-iteration'.

Note that (6) is equivalent to the inequality

$$\|z\|_{L_2}^2 - \gamma^2\left[\|w\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2\right] < 0 \ , \text{ all } (w,\hat{v}) \neq (0,0). \tag{7}$$

The equivalence of (6) and (7) will be used extensively. The condition (7) is particularly useful, since the expression on the left-hand side is quadratic in the variables. This makes it possible to apply linear-quadratic control and game theory methods to check whether (7) holds for a particular control law, and to solve the problem of finding a controller which achieves the performance bound (7), when such a controller exists.

III. DISCRETE SYSTEM REPRESENTATION

It is natural to represent the sampled-data control system defined by Eqs. (1)–(3) as a discrete system with the system state defined at the sampling instants $\{kh\}$ [15, 22]. Consider the evolution of the system state $x(t)$ at the sampling instants kh. Then

$$x(kh+h) = e^{Ah}x(kh) + \int_0^h e^{A(h-\lambda)}B_1 w(kh+\lambda)d\lambda$$

$$+ \int_0^h e^{A(h-\lambda)} B_2 d\lambda \hat{u}_k \ , \ x(0) = 0 \tag{8a}$$

$$z(kh+\tau) = C_1 e^{A\tau} x(kh) + \int_0^\tau C_1 e^{A(\tau-\lambda)} B_1 w(kh+\lambda) d\lambda + \Big(\int_0^\tau C_1 e^{A(\tau-\lambda)} B_2 d\lambda + D_{12} \Big) \hat{u}_k \ , \ \tau \in [0, h) \tag{8b}$$

$$\hat{y}_k = C_2 x(kh) + \hat{D}_{21} \hat{v}_k \ , \ k = 0, 1, \ldots . \tag{8c}$$

This equation describes the sampled-data system as a discrete system with finite-dimensional state vector $x(kh)$, but with inputs and outputs taking values in the infinite-dimensional spaces $L_2^{m_1}[0, h)$ and $L_2^{p_1}[0, h)$, respectively. In order to obtain a compact discrete state-space representation of (8), introduce the lifted discrete signals $\hat{z} := (\hat{z}_0, \hat{z}_1, \ldots)$ and $\hat{w} := (\hat{w}_0, \hat{w}_1, \ldots)$, belonging to the space $l_2^{L_2[0,h)}$ of sequences with elements in $L_2[0, h)$, and defined according to

$$\hat{z}_k(\tau) = z(kh+\tau) \ , \ \hat{w}_k(\tau) = w(kh+\tau) \ , \ \tau \in [0, h) \ , \ k = 0, 1, \ldots . \tag{9}$$

By construction, the lifting operation $L_2 \to l_2^{L_2[0,h)}$ defined by Eq. (9) is norm preserving [15, 22],

$$\|\hat{z}\|_{l_2} = \|z\|_{L_2} \ , \ \|\hat{w}\|_{l_2} = \|w\|_{L_2}. \tag{10}$$

The system (8) can then be written compactly as

$$\hat{x}_{k+1} = \hat{A}\hat{x}_k + \hat{B}_1 \hat{w}_k + \hat{B}_2 \hat{u}_k \ , \ \hat{x}_0 = 0 \tag{11a}$$

$$\hat{z}_k = \hat{C}_1 \hat{x}_k + \hat{D}_{11} \hat{w}_k + \hat{D}_{12} \hat{u}_k \tag{11b}$$

$$\hat{y}_k = \hat{C}_2 \hat{x}_k + \hat{D}_{21} \hat{v}_k \tag{11c}$$

where $\hat{x}_k := x(kh)$, and $\hat{A}$, $\hat{B}_1$, $\hat{B}_2$, $\hat{C}_1$, $\hat{D}_{11}$, $\hat{D}_{12}$, $\hat{C}_2$, and $\hat{D}_{21}$ are operators on the respective spaces, i.e.,

$$\begin{aligned}
\hat{A} &: R^n \to R^n \\
\hat{B}_1 &: L_2^{m_1}[0, h) \to R^n \\
\hat{B}_2 &: R^{m_2} \to R^n \\
\hat{C}_1 &: R^n \to L_2^{p_1}[0, h) \\
\hat{D}_{11} &: L_2^{m_1}[0, h) \to L_2^{p_1}[0, h) \\
\hat{D}_{12} &: R^{m_2} \to L_2^{p_1}[0, h) \\
\hat{C}_2 &: R^n \to R^{p_2} \\
\hat{D}_{21} &: R^{m_3} \to R^{p_2}
\end{aligned}$$

defined as

$$\begin{aligned}
\hat{A} &= e^{Ah} \\
\hat{B}_1\hat{w} &= \int_0^h e^{A(h-\lambda)}B_1\hat{w}(\lambda)d\lambda \\
\hat{B}_2 &= \int_0^h e^{A(h-\lambda)}B_2 d\lambda \\
(\hat{C}_1\hat{x})(\tau) &= C_1 e^{A\tau}\hat{x} \\
(\hat{D}_{11}\hat{w})(\tau) &= \int_0^\tau C_1 e^{A(\tau-\lambda)}B_1\hat{w}(\lambda)d\lambda \\
(\hat{D}_{12}\hat{u})(\tau) &= \Big(\int_0^\tau C_1 e^{A(\tau-\lambda)}B_2 d\lambda + D_{12}\Big)\hat{u} \\
\hat{C}_2 &= C_2.
\end{aligned} \tag{12}$$

The system described by Eq. (8), or the lifted representation (11), is a hybrid discrete/continuous system, part of which (the plant) evolves in continuous time, while another part (the controller) evolves in discrete time. For such systems, Chen and Francis [8] have introduced of useful stability concept in terms of hybrid stability, defined as internal stability when the signals are taken to be in L_2 and l_2, respectively. Their analysis shows thats the sampled-data system described by Eqs. (1)–(3) is stabilizable by a feedback law (4) if and only if the pair $(\hat{A}, \hat{B}_2)$ is stabilizable and the pair $(\hat{C}_2, \hat{A})$ is detectable in discrete time. For any discrete controller (4), the closed-loop system is stable if and only if the corresponding discrete system $(\hat{A}, \hat{B}_2, \hat{C}_2)$ is stable.

By Eq. (10), the performance measure (5) equals the l_2-induced norm of the discrete system (11), i.e.,

$$J(\mathcal{K}) = \sup\Big\{\frac{\|\hat{z}\|_{l_2}}{[\|\hat{w}\|_{l_2}^2 + \|\hat{v}\|_{l_2}^2]^{1/2}}\Big\} \tag{13}$$

where the supremum is taken over $(\hat{w}, \hat{v}) \neq (0,0)$ in $l_2^{L_2[0,h)} \oplus l_2^{m_3}$. The control problem defined by the sampled-data system (1)–(3) and the performance bound (6) can thus be characterized in terms of a discrete H_∞ control problem defined by the system (11) and the performance measure (13). Its solution is, however, complicated by the fact that the system (11) has infinite-dimensional input and output spaces.

Note that the operators $\hat{B}_1$, $\hat{C}_1$ and $\hat{D}_{12}$ have finite ranks and they may therefore be represented by finite-dimensional matrices, whereas the operator $\hat{D}_{11}$ has infinite rank. In [22], the operator $\hat{D}_{11}$ was approximated by finite-rank operators obtained by expanding $\hat{w}_k$ and $\hat{z}_k$ in orthonormal bases in $L_2[0,h)$. Since $\hat{D}_{11}$ is compact, it can be approximated to any desired

accuracy in this way. Hence the H_∞ sampled-data control problem can also be solved to any specified accuracy in terms of finite-dimensional discrete H_∞ control problems [22]. A related approach, in which the continuous-time signals are expanded in terms of sequences of step functions, has been applied in [36, 37] to the discretization of continuous-time controllers.

An exact solution to the H_∞ sampled-data control problem, which does not involve any approximations, was given by Bamieh *et al.* [14] and Bamieh and Pearson [15]. They applied a loop-shifting procedure to the lifted system described by Eq. (11), which removes the operator $\hat{D}_{11}$ and preserves closed-loop stability and norm. In this way the H_∞ sampled-data problem could be solved exactly as a finite-dimensional discrete H_∞ control problem.

Note that a necessary condition for the performance bound (6) to hold is $\|\hat{D}_{11}\| < \gamma$. A feedback connection can then be introduced around the plant (Figure 1) to define new input and output signals $\acute{w}_k$ and $\acute{r}_k$ according to

$$\begin{bmatrix} \acute{r}_k \\ \gamma\hat{w}_k \end{bmatrix} = \Theta \begin{bmatrix} \gamma\acute{w}_k \\ \hat{z}_k \end{bmatrix} \tag{14}$$

where $\Theta : L_2[0,h) \oplus L_2[0,h) \to L_2[0,h) \oplus L_2[0,h)$ is the unitary operator given by

$$\Theta = \begin{bmatrix} -\gamma^{-1}\hat{D}_{11} & (I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*)^{1/2} \\ (I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11})^{1/2} & \gamma^{-1}\hat{D}_{11}^* \end{bmatrix} \tag{15}$$

where D_{11}^* denotes the adjoint operator.

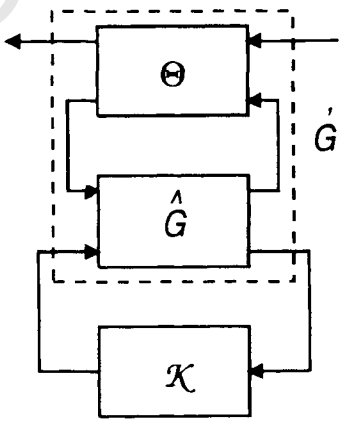


Fig. 1. The feedback connection (11), (14).

Substituting the new variables $\acute{w}_k$ and $\acute{r}_k$ from Eq. (14) into the system equation (11) gives the state-space realization

$$\begin{aligned} \hat{x}_{k+1} &= \acute{A}\hat{x}_k + \acute{B}_2\hat{u}_k + \acute{B}_1\left(I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11}\right)^{-1/2}\acute{w}_k \\ \acute{r}_k &= \left(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*\right)^{-1/2}\left[\hat{C}_1\hat{x}_k + \hat{D}_{12}\hat{u}_k\right] \\ \hat{y}_k &= \hat{C}_2\hat{x}_k + \hat{D}_{21}\hat{v}_k \end{aligned} \tag{16}$$

where

$$\begin{aligned} \acute{A} &= \hat{A} + \gamma^{-2}\hat{B}_1\hat{D}_{11}^*\left(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*\right)^{-1}\hat{C}_1 \\ \acute{B}_2 &= \hat{B}_2 + \gamma^{-2}\hat{B}_1\hat{D}_{11}^*\left(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*\right)^{-1}\hat{D}_{12}. \end{aligned} \tag{17}$$

We then have the following result [15].

Lemma 1. *Consider the lifted system described by Eq. (11). Assume that* $\|\hat{D}_{11}\| < \gamma$. *Then for any controller* $\hat{u} = \mathcal{K}\hat{y}$ *the following are equivalent:*
(i) $\mathcal{K}$ *stabilizes the system (11) and* $J(\mathcal{K}) < \gamma$,
(ii) $\mathcal{K}$ *stabilizes the system (16) and*

$$\sup\left\{\frac{\|\acute{r}\|_{l_2}}{\left[\|\acute{w}\|_{l_2}^2 + \|\hat{v}\|_{l_2}^2\right]^{1/2}}\right\} < \gamma. \tag{18}$$

Proof: Note that the transformation $\Theta : L_2[0,h) \oplus L_2[0,h) \to L_2[0,h) \oplus L_2[0,h)$ is unitary. Hence

$$\|\acute{r}_k\|_{L_2[0,h)}^2 + \gamma^2\|\hat{w}_k\|_{L_2[0,h)}^2 = \gamma^2\|\acute{w}_k\|_{L_2[0,h)}^2 + \|\hat{z}_k\|_{L_2[0,h)}^2$$

or

$$\|\acute{r}_k\|_{L_2[0,h)}^2 - \gamma^2\|\acute{w}_k\|_{L_2[0,h)}^2 = \|\hat{z}_k\|_{L_2[0,h)}^2 - \gamma^2\|\hat{w}_k\|_{L_2[0,h)}^2,$$

and the norm result follows. The stability part of the lemma is harder to show. In [15], it was shown using an operator-valued version of the Redheffer lemma. Here we will present an alternative proof in Section V (Theorem 5) in connection with a worst-case sampling approach. □

In the state-space representation (16) the operator $\hat{D}_{11}$ is removed, but the input and output signals $\acute{w}_k$ and $\acute{r}_k$, taking values in $L_2[0,h)$, have still infinite dimensions. It is, however, straightforward to obtain an equivalent finite-dimensional system characterization using the fact that all operators in (16) have finite ranks. An application of Lemma C.1 in Appendix C results in the following theorem.

Theorem 1. *Consider the lifted system described by Eq. (11). Assume that* $\|\hat{D}_{11}\| < \gamma$. *Define the finite-dimensional system*

$$\begin{aligned} \hat{x}_{k+1} &= \acute{A}\hat{x}_k + \acute{B}_2\hat{u}_k + \acute{B}_1\hat{v}_k \\ \acute{z}_k &= \acute{C}_1\hat{x}_k + \acute{D}_{12}\hat{u}_k \\ \hat{y}_k &= \hat{C}_2\hat{x}_k + \hat{D}_{21}\hat{v}_k \end{aligned} \tag{19}$$

where $\acute{A}$ *and* $\acute{B}_2$ *are given by Eq. (17), and* $\acute{B}_1, \acute{C}_1$ *and* $\acute{D}_{12}$ *are defined by*

$$\acute{B}_1\acute{B}_1' = \hat{B}_1(I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11})^{-1}\hat{B}_1^* \tag{20}$$

$$\begin{bmatrix} \acute{C}_1' \\ \acute{D}_{12}' \end{bmatrix}[\acute{C}_1 \quad \acute{D}_{12}] = \begin{bmatrix} \hat{C}_1^* \\ \hat{D}_{12}^* \end{bmatrix}(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*)^{-1}[\hat{C}_1 \quad \hat{D}_{12}].$$

Then for any controller $\hat{u} = \mathcal{K}\hat{y}$ *the following are equivalent:*

(i) $\mathcal{K}$ *stabilizes the system (11) and* $J(\mathcal{K}) < \gamma$,

(ii) $\mathcal{K}$ *stabilizes the system (19) and*

$$\sup\left\{\frac{\|\acute{z}\|_{l_2}}{[\|\acute{v}\|_{l_2}^2 + \|\hat{v}\|_{l_2}^2]^{1/2}}\right\} < \gamma. \tag{21}$$

In (19), all signals are finite-dimensional. Explicit formulae for the matrices $\acute{A}, \acute{B}_1, \acute{B}_2, \acute{C}_1$ and $\acute{D}_{12}$ are given in [15]. These expression will not be repeated here, but instead alternative formulae will be presented in Section V (Lemma 9) in connection with a worst-case sampling approach.

A possible inconvenience with the result in Theorem 1 is the fact that the system matrices $\acute{A}$ and $\acute{B}_2$ in (19) are different from the system matrices of the discrete representation (11). The set of stabilizing controllers of the discrete system (19) is therefore in general different from the set of controllers which stabilize the sampled-data system (11). It is, however, possible to define a second norm-preserving variable transformation, which takes the system (19) into a finite-dimensional discrete system with the same system matrices $\hat{A}$ and $\hat{B}_2$ as the discrete representation (11), cf. Hayakawa *et al.* [26].

The problem of finding a sampled-data controller (4) such that (21) holds is a standard finite-dimensional discrete H_∞ control problem. This problem can be solved using standard discrete techniques for the discrete H_∞ control problem [4, 33], and it will not be discussed here.

Theorem 1 gives a complete solution of the H_∞ sampled-data control problem defined in Section II in terms of a finite-dimensional discrete H_∞ control problem. Note that the result holds for the finite-time case as well, and that it can easily be modified to time-varying system. It is, however, not easy to apply the result to more general situations, such as problems where the dynamics of the sampler and/or the hold function are part of the design problem, or cases when the periods of the sampler and the hold function are different. It is therefore also well motivated to study other methods to the H_∞ sampled-data control problem.

IV. A DYNAMIC GAME SOLUTION

An alternative approach to the discretization procedure of Section III is to solve the sampled-data control problem directly using a time-domain

approach to the mixed discrete/continuous H_∞ control problem defined in Section II. This approach is related to existing time-domain solution methods for standard continuous and discrete H_∞ control problems [3, 4, 33, 34]. These methods are related to dynamic games [33], and in this framework the sampled-data H_∞ control problem can be considered as a hybrid discrete/continuous dynamic game. The approach has been applied in [23] to a time-varying, finite-horizon problem, and by Sun *et al.* [19] to the stationary problem. In a related study, Sivashankar and Khargonekar [11, 12] have applied a similar approach to compute the induced norm of sampled-data systems. The dynamic game procedure is also well suited for treating sampled-data control problems with free hold functions [18, 31, 32, 33].

It is convenient to develop the procedure first for a finite-time control problem over the time interval $[0, Nh]$, in which the induced norm $J(\mathcal{K})$, Eq. (5), is taken to be induced by signals in $L_2[0, Nh]$ and $l_2(0, N)$. In analogy with Eq. (7), introduce the finite-time cost function

$$L_{[0,Nh]}(\hat{u}, w, \hat{v}) := \|z\|^2_{L_2[0,Nh]} - \gamma^2 \left[\|w\|^2_{L_2[0,Nh]} + \|\hat{v}\|^2_{l_2(0,N)}\right]. \tag{22}$$

The performance bound $J(\mathcal{K}) < \gamma$ holds on $[0, Nh]$ if and only if the inequality $L_{[0,Nh]}(\hat{u}, w, \hat{v}) < 0$ holds for all $(w, \hat{v}) \neq (0, 0)$ in $L_2[0, Nh] \oplus l_2(0, N)$. The problem of finding a control sequence which achieves this is a two-player discrete/continuous dynamic game, where the u-player, restricted to using piecewise constant signals according to Eq. (3), tries to ensure that $L_{[0,Nh]}(\hat{u}, w, \hat{v}) < 0$ for all $(w, \hat{v}) \neq (0, 0)$. For its solution, standard methods of dynamic game theory [33] can be employed [23].

The solution to the sampled-data H_∞ problem is obtained in several stages. First, the solution to a sampled-data H_∞ problem with complete state information is obtained. This leads to an H_∞ filtering problem with sampled measurements. A combination of the control and filtering results then gives the solution to the original control problem.

A. THE STATE FEEDBACK CONTROL PROBLEM

In this subsection, the solution of the control problem defined above is obtained for the case when complete information of the state is available to the controller. The sampled-data system described by Eqs. (1) and (3) can be represented as a hybrid discrete/continuous system described by the

state-space equations

$$\begin{aligned}\begin{bmatrix}\dot{x}(t)\\ \dot{u}(t)\end{bmatrix} &= A_e\begin{bmatrix}x(t)\\ u(t)\end{bmatrix} + B_{e1}w(t)\ ,\ x(0^-)=0\\ z(t) &= C_{e1}\begin{bmatrix}x(t)\\ u(t)\end{bmatrix}\ ,\ t\in[kh,kh+h)\\ \begin{bmatrix}x(kh)\\ u(kh)\end{bmatrix} &= \begin{bmatrix}I & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}x(kh^-)\\ u(kh^-)\end{bmatrix} + \begin{bmatrix}0\\ I\end{bmatrix}\hat{u}_k\ ,\ k=0,1,\dots,N-1\end{aligned} \tag{23}$$

where

$$A_e := \begin{bmatrix}A & B_2\\ 0 & 0\end{bmatrix}\ ,\ B_{e1} := \begin{bmatrix}B_1\\ 0\end{bmatrix}\ ,\ C_{e1} := [C_1 \quad D_{12}]. \tag{24}$$

Lemma 2. *Consider the system (23). Assume that assumption (A1) holds. Assume that there is a bounded symmetric positive semidefinite matrix*

$$S(t) := \begin{bmatrix}S_{11} & S_{12}\\ S_{12}' & S_{22}\end{bmatrix}(t)\ ,\ t\in[0,Nh) \tag{25}$$

which satisfies the following Riccati differential equation with jumps associated with (23),

$$\begin{aligned}-\dot{S}(t) &= A_e'S(t) + S(t)A_e + \gamma^{-2}S(t)B_{e1}B_{e1}'S(t) + C_{e1}'C_{e1},\ t\neq kh\\ S(kh^-) &= \begin{bmatrix}S_{11}(kh^-) & 0\\ 0 & 0\end{bmatrix},\\ S_{11}(kh^-) &= S_{11}(kh) - S_{12}(kh)S_{22}^{-1}(kh)S_{12}'(kh),\\ S(Nh^-) &= 0,\ k=N-1,\dots,1,0.\end{aligned} \tag{26}$$

Then

$$\int_0^{Nh}(z'z-\gamma^2w'w)dt = \sum_{k=0}^{N-1}(\hat{u}_k-\hat{u}_k^o)'S_{22}(kh)(\hat{u}_k-\hat{u}_k^o) - \gamma^2\|w-w^o\|_2^2$$

where

$$\hat{u}_k^o = -S_{22}^{-1}(kh)S_{12}'(kh)x(kh) \tag{27a}$$

$$w^o(t) = \gamma^{-2}B_1'[S_{11}(t)x(t) + S_{12}(t)u(t)]. \tag{27b}$$

Proof: From (23) and (26) we have

$$\int_{kh}^{kh+h}(z'z-\gamma^2 w'w)dt=\int_{kh}^{kh+h}\left[z'z-\gamma^2 w'w+\frac{d}{dt}\left([x'\ \ u']S\begin{bmatrix}x\\u\end{bmatrix}\right)\right]dt$$
$$-x'(kh+h)S_{11}(kh+h^-)x(kh+h)+[x'(kh)\ \ \hat{u}_k']S(kh)\begin{bmatrix}x(kh)\\\hat{u}_k\end{bmatrix}.$$

Here

$$\int_{kh}^{kh+h}\left[z'z-\gamma^2 w'w+\frac{d}{dt}\left([x'\ \ u']S\begin{bmatrix}x\\u\end{bmatrix}\right)\right]dt$$
$$=-\gamma^2\int_{kh}^{kh+h}(w-w^o)'(w-w^o)dt$$

and

$$[x'(kh)\ \ \hat{u}_k']S(kh)\begin{bmatrix}x(kh)\\\hat{u}_k\end{bmatrix}=\hat{u}_k'S_{22}(kh)\hat{u}_k+2\hat{u}_k'S_{12}'(kh)x(kh)$$
$$+x'(kh)S_{11}(kh)x(kh)$$
$$=(\hat{u}_k-\hat{u}_k^o)'S_{22}(kh)(\hat{u}_k-\hat{u}_k^o)+x'(kh)S_{11}(kh^-)x(kh).$$

The result of the lemma then follows by summing over k and observing that $x(0)=0$ by assumption. □

Remark. For simplicity, Lemma 2 has been stated for the case when the matrices $S_{22}(kh)$ are invertible. Assumption (A1) is a sufficient condition for this to hold. The assumption can, however, be relaxed [23], and when $S_{22}(kh)$ is singular, its inverse in (26) and (27*a*) is replaced by the pseudoinverse [23].

From Lemma 2 we have the following state feedback result.

Theorem 2. *Consider the sampled-data system described by Eqs. (1) and (3). Assume that assumption (A1) holds. There exists a state feedback law* $\hat{u}=K(\{x(kh)\})$ *such that the performance bound*

$$\|z\|^2_{L_2[0,Nh]}-\gamma^2\|w\|^2_{L_2[0,Nh]}<0\ ,\quad \textit{all } w\neq 0 \tag{28}$$

holds for the closed loop if and only if the Riccati equation with jumps (26) has a bounded symmetric positive semidefinite solution on $[0,Nh)$. *In that case, the performance bound (28) is achieved by the linear discrete causal state feedback law (27a).*

Proof: The result can be shown via standard arguments in H_∞ control theory, cf. [33, 23, 19]. For the sake of completeness, a proof is presented in Appendix A. □

The sampled-data H_∞ control problem can now be solved via an H_∞ filtering problem with sampled measurements. For this purpose, introduce the variables $v(t) := w(t) - w^o(t)$ and $\hat{r}_k := S_{22}^{1/2}(kh)(\hat{u}_k - \hat{u}_k^o)$, i.e.

$$v(t) := w(t) - \gamma^{-2} B_1' [S_{11}(t)x(t) + S_{12}(t)\hat{u}_k],\ t \in [kh, kh+h),\ k = 0, 1, \ldots \tag{29}$$

and

$$\hat{r}_k := S_{22}^{1/2}(kh)[S_{22}^{-1}(kh)S_{12}'(kh)x(kh) + \hat{u}_k]\ ,\ k = 0, 1, \ldots, N-1. \tag{30}$$

The system (1)–(3) is then described by

$$\begin{aligned} \dot{x}(t) &= \bar{A}(t)x(t) + \bar{B}_1 v(t) + \bar{B}_2(t)\hat{u}_k\ ,\ t \neq kh\ ,\ x(0) = 0 \\ \hat{r}_k &= \bar{C}_{1,k}x(kh) + \bar{D}_{12,k}\hat{u}_k \\ \hat{y}_k &= \hat{C}_2 x(kh) + \hat{D}_{21}\hat{v}_k\ ,\ k = 0, 1, \ldots, N-1 \end{aligned} \tag{31}$$

where

$$\begin{aligned} \bar{A}(t) &:= A + \gamma^{-2} B_1 B_1' S_{11}(t) \\ \bar{B}_1 &:= B_1 \\ \bar{B}_2(t) &:= B_2 + \gamma^{-2} B_1 B_1' S_{12}(t) \\ \bar{C}_{1,k} &:= S_{22}^{-1/2}(kh) S_{12}'(kh) \\ \bar{D}_{12,k} &:= S_{22}^{1/2}(kh). \end{aligned} \tag{32}$$

From Lemma 2 we have,

$$\|\hat{r}\|_{l_2}^2 - \gamma^2 \|v\|_{L_2}^2 = \|z\|_{L_2}^2 - \gamma^2 \|w\|_{L_2}^2. \tag{33}$$

The sampled-data H_∞ control problem can now be characterized as follows.

Lemma 3. *Consider the sampled-data system described by Eqs. (1)–(3). Assume that assumption (A1) holds. Then for any discrete causal controller $\hat{u} = \mathcal{K}\hat{y}$ the following statements are equivalent:*

(i) $J(\mathcal{K}) < \gamma$ *on* $[0, Nh]$,

(ii) the Riccati differential equation with jumps (26) has a bounded solution on $[0, Nh)$, *and*

$$\|\hat{r}\|_{l_2}^2 < \gamma^2 \left[\|v\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2\right] \tag{34}$$

holds for all $(v, \hat{v}) \neq (0, 0)$ *for the system (31).*

Moreover, assuming that the Riccati differential equation with jumps (26) has a bounded solution on $[0, Nh)$, *there exists a discrete causal controller which achieves the performance bound (34) if and only if there exists*

an estimator $\hat{r}_e$, which depends causally on $\hat{y}$ and $\hat{u}$, and which achieves the performance bound

$$\|\hat{r} - \hat{r}_e\|_{l_2}^2 < \gamma^2 \left[\|v\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2\right], \text{ all } (v, \hat{v}) \neq (0, 0). \tag{35}$$

Proof: The equivalence of (i) and (ii) follows from the fact that a necessary condition for (i) is the existence of a state feedback which achieves the performance bound $J(\mathcal{K}) < \gamma$, Theorem 2, and the identity (33). Note that since a finite-horizon problem is considered, the stability issue is not involved. In order to prove the second part of the theorem, assume first that the controller $\hat{u} = \mathcal{K}\hat{y}$ achieves the bound (34). Form the causal estimator

$$\begin{aligned}
\dot{\hat{x}}(t) &= \bar{A}(t)\hat{x}(t) + \bar{B}_2(t)\left[\hat{u}_k - (\mathcal{K}\hat{y})_k\right], \ t \in [kh, kh + h), \ \hat{x}(0) = 0 \\
\hat{r}_{e,k} &= \bar{C}_{1,k}\hat{x}(kh) + \bar{D}_{12,k}\left[\hat{u}_k - (\mathcal{K}\hat{y})_k\right], \ k = 0, 1, \ldots, N - 1.
\end{aligned}$$

For the input $\hat{u} = \mathcal{K}\hat{y}$, we have $\hat{r}_e \equiv 0$, and the estimator achieves the performance bound (35) since (34) holds. It is easy to see that the estimation error is independent of $\hat{u}$, and hence the above estimator achieves the performance bound (35) for all inputs $\hat{u}$. Conversely, assume that the performance bound (35) is achieved for some causal estimator $\hat{r}_e$. As $\bar{D}_{12,k}$ has full row rank, the control signal $\hat{u}$ can be chosen so that $\hat{r}_e \equiv 0$. Then $\hat{r} - \hat{r}_e \equiv \hat{r}$, and it follows that the controller constructed in this way achieves the performance bound (34). □

By Lemma 3, the sampled-data H_∞ control problem is reduced to an H_∞-optimal estimation problem. The solution of the estimation problem with sampled measurements that is involved will given in the next subsection.

B. H_∞ OPTIMAL ESTIMATION WITH SAMPLED MEASUREMENTS

The estimation results needed for the sampled-data H_∞ problem are given in Lemma 4 and Lemma 5 below.

Lemma 4. *Consider the system*

$$\begin{aligned}
\dot{x}(t) &= \bar{A}(t)x(t) + \bar{B}_1(t)v(t), \ x(0) = 0 \\
\hat{r}_k &= \bar{C}_{1,k}x(kh) \\
\hat{y}_k &= \hat{C}_2 x(kh) + \hat{D}_{21}\hat{v}_k \ , \ t \in [0, Nh].
\end{aligned} \tag{36}$$

Assume that assumption (A2) holds. There exists a discrete causal filter $\hat{r}_e = \mathcal{F}\hat{y}$ such that

$$\|\hat{r} - \hat{r}_e\|_{l_2}^2 < \gamma^2 \left[\|v\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2\right] \tag{37}$$

holds for all $(v, \hat{v}) \neq (0,0)$, *if and only if there exists a solution to the Riccati equation with jumps,*

$$\begin{aligned} \dot{N}(t) &= \bar{A}(t)N(t) + N(t)\bar{A}(t)' + \bar{B}_1(t)\bar{B}_1'(t)\ ,\ t \neq kh \\ N(kh) &= N(kh^-)\Sigma_k^{-1} \\ \Sigma_k &= I + [\hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2 - \gamma^{-2}\bar{C}_{1,k}'\bar{C}_{1,k}]N(kh^-) \\ N(0^-) &= 0\ ,\ k = 0,1,\ldots,N, \end{aligned} \tag{38}$$

such that the matrices Σ_k, $k = 0,1,\ldots,N$, *have only positive eigenvalues. In this case, the estimator*

$$\begin{aligned} \dot{\hat{x}}(t) &= \bar{A}(t)\hat{x}(t),\ t \in [kh, kh+h)\ ,\ \hat{x}(0^-) = 0 \\ \hat{x}(kh) &= \hat{x}(kh^-) \\ &\quad + N(kh^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2N(kh^-)\hat{C}_2']^{-1}\big(y_k - \hat{C}_2\hat{x}(kh^-)\big) \\ \hat{r}_{e,k} &= \bar{C}_{1,k}\hat{x}(kh)\ ,\ k = 0,1,\ldots,N \end{aligned} \tag{39}$$

achieves the performance bound (37).

Proof: See Appendix B. □

It is also of interest to consider the open-loop H_∞-optimal estimator for estimating the controlled state $z(t)$ based on the sampled measurements $\hat{y}$. We have the following H_∞ estimation result.

Lemma 5. *Consider the system described by Eqs. (1) and (2) on* $t \in [0, Nh]$. *There exists a causal filter* $\mathcal{F}: l_2 \to L_2$ *such that the estimator* $z_e = \mathcal{F}\hat{y}$ *achieves*

$$\|z - z_e\|_{L_2}^2 < \gamma^2\left[\|w\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2\right] \tag{40}$$

for all $(w, \hat{v}) \neq (0,0)$, *if and only if there exists a bounded symmetric positive semidefinite solution* $Q(\cdot)$ *to the Riccati differential equation with jumps*

$$\begin{aligned} \dot{Q} &= AQ + QA' + \gamma^{-2}QC_1'C_1Q + B_1B_1'\ ,\ t \neq kh \\ Q(kh) &= Q(kh^-) - Q(kh^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2Q(kh^-)\hat{C}_2']^{-1}\hat{C}_2Q(kh^-) \\ Q(0^-) &= 0\ ,\ k = 0,1,\ldots,N. \end{aligned} \tag{41}$$

In this case, the estimator

$$\begin{aligned} \dot{\hat{x}}(t) &= A(t)\hat{x}(t),\ t \neq kh,\ \hat{x}(0^-) = 0 \\ \hat{x}(kh) &= \hat{x}(kh^-) \\ &\quad + Q(kh^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2Q(kh^-)\hat{C}_2']^{-1}\big(y_k - \hat{C}_2\hat{x}(kh^-)\big) \\ z_e(t) &= C_1\hat{x}(t)\ ,\ k = 0,1,\ldots,N \end{aligned} \tag{42}$$

achieves the performance bound (40).

Proof: The result can be proved in the same way as Lemma 4. See also reference [18]. □

C. SOLUTION OF THE SAMPLED-DATA H_∞ CONTROL PROBLEM

The results of Lemma 3 and Lemma 4 can now be combined to obtain a solution to the H_∞ control problem as follows.

Lemma 6. *Consider the sampled-data system described by Eqs. (1)–(3). Assume that assumptions (A1) and (A2) hold. There exists a discrete causal control law $\hat{u} = \mathcal{K}\hat{y}$ which achieves the performance bound $J(\mathcal{K}) < \gamma$ on $[0, Nh]$ if and only if*

(i) the Riccati equation with jumps (26) has a bounded positive semidefinite solution $S(t)$ on $[0, Nh)$, and

(ii) for $\bar{A}(t)$, $\bar{B}_1$, $\bar{B}_2(t)$ and $\bar{C}_{1,k}$ defined by Eq. (32), there exists a bounded positive semidefinite matrix $N(t)$ which satisfies

$$\begin{aligned} \dot{N}(t) &= \bar{A}(t)N(t) + N(t)\bar{A}(t)' + \bar{B}_1\bar{B}_1', \ t \in [kh, kh+h) \\ N(kh) &= N(kh^-)\Sigma_k^{-1} \\ \Sigma_k &= I + [\hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2 - \gamma^{-2}\bar{C}_{1,k}'\bar{C}_{1,k}]N(kh^-) \\ N(0^-) &= 0, \end{aligned} \tag{43}$$

such that the matrices Σ_k, $k = 0, 1, \ldots, N-1$, have only positive eigenvalues.

Moreover, when conditions (i) and (ii) are satisfied, the performance bound $J(\mathcal{K}) < \gamma$ is achieved by the controller

$$\begin{aligned} \dot{\hat{x}}(t) &= \bar{A}(t)\hat{x}(t) + \bar{B}_2(t)\hat{u}_k \ , \ t \in [kh, kh+h) \ , \ \hat{x}(0^-) = 0 \\ \hat{x}(kh) &= \hat{x}(kh^-) \\ &\quad + N(kh^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2N(kh^-)\hat{C}_2']^{-1}\big(y_k - \hat{C}_2\hat{x}(kh^-)\big) \\ \hat{u}_k &= -S_{22}^{-1}(kh)S_{12}'(kh)\hat{x}(kh) \ , \ k = 0, 1, \ldots, N-1. \end{aligned} \tag{44}$$

Note that the controller (44) sets $\hat{r}_e \equiv 0$, where $\hat{r}_e$ achieves the performance bound (35) according to Lemma 4. Hence it follows from the proof of Lemma 3 that the controller achieves the performance bound as stated.

The sampled-data H_∞-optimal controller can also be characterized in terms of the state-feedback solution of Theorem 2 and the H_∞-optimal open-loop estimator of Lemma 5.

Theorem 3. *Consider the sampled-data system described by Eqs. (1)–(3). Assume that assumptions (A1) and (A2) hold. There exists a discrete*

causal control law $\hat{u} = \mathcal{K}\hat{y}$ which achieves the performance bound $J(\mathcal{K}) < \gamma$ on $[0, Nh]$ if and only if the following conditions are satisfied:

(i) The Riccati differential equation with jumps (26) has a bounded positive semidefinite solution $S(t)$ on $[0, Nh)$,

(ii) the Riccati differential equation with jumps (41) has a bounded positive semidefinite solution $Q(t)$ on $[0, Nh)$, and

(iii) $\rho(S_{11}(t)Q(t)) < \gamma^2$ for all $t \in [0, Nh)$.

Moreover, when conditions (i)–(iii) hold, a controller which achieves the performance bound $J(\mathcal{K}) < \gamma$ is given by Eq. (44), where

$$N(t) = Q(t)(I - \gamma^{-2}S_{11}(t)Q(t))^{-1}. \tag{45}$$

Proof: By Lemma 6, a discrete controller which achieves the performance bound $J(\mathcal{K}) < \gamma$ exists if and only if the Riccati equations (26) and (43) have bounded solutions. When (43) has a solution, it is straightforward to verify that $Q(t) := N(t)(I + \gamma^{-2}S_{11}(t)N(t))^{-1}$ is a solution to (41). Moreover, *(iii)* follows from $S_{11}(t)Q(t) = \gamma^2 S_{11}(t)N(t)(\gamma^2 I + S_{11}(t)N(t))^{-1}$ and the positive semidefiniteness of the matrices $S_{11}(t)$ and $N(t)$. Conversely, if (41) has a solution such that *(iii)* holds, then $N(t)$ defined by (45) is positive semidefinite and satisfies (43). □

For computational purposes, it is useful to note that the Riccati equations with jumps (26) and (41) can be reduced to discrete Riccati equations as follows.

Lemma 7. *(a) Define the $2(n + m_2) \times 2(n + m_2)$ matrix*

$$\Pi(t) := \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}(t) := \exp\left(- \begin{bmatrix} A_e & \gamma^{-2}B_{e1}B_{e1}' \\ -C_{e1}'C_{e1} & -A_e' \end{bmatrix} t \right) \tag{46}$$

where the matrices A_e, B_{e1} and C_{e1} are defined by Eq. (24). Then the solution of the Riccati differential equation with jumps (26) is given by

$$\begin{aligned} S(kh^-) &= A_d'S(kh)A_d - A_d'S(kh)B_d\left(B_d'S(kh)B_d\right)^{-1}B_d'S(kh)A_d, \\ S(kh - t) &= [\Pi_{21}(t) + \Pi_{22}(t)S(kh^-)][\Pi_{11}(t) + \Pi_{12}(t)S(kh^-)]^{-1}, t \in (0, h] \\ S(Nh^-) &= 0\ ,\ k = N, \ldots, 1, 0 \end{aligned} \tag{47}$$

where

$$A_d := \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}\ ,\ B_d := \begin{bmatrix} 0 \\ I \end{bmatrix}. \tag{48}$$

(b) Define the $2n \times 2n$ matrix

$$\Lambda(t) := \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}(t) := \exp\left(\begin{bmatrix} -A' & -\gamma^{-2}C_1'C_1 \\ B_1B_1' & A \end{bmatrix} t \right). \tag{49}$$

Then the solution of the Riccati differential equation with jumps (41) is given by

$$\begin{aligned} Q(kh+t) &= [\Lambda_{21}(t) + \Lambda_{22}(t)Q(kh)][\Lambda_{11}(t) + \Lambda_{12}(t)Q(kh)]^{-1}\ ,\ t \in [0,h) \\ Q(kh) &= Q(kh^-) - Q(kh^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2 Q(kh^-)\hat{C}_2']^{-1}\hat{C}_2 Q(kh^-) \\ Q(0^-) &= 0\ ,\ k = 0,1,\ldots,N. \end{aligned} \tag{50}$$

Remark. In analogy with other H_∞ control problems, the infinite horizon case for time-invariant plants can be obtained as a limiting case from Theorem 3 as $N \to \infty$. More precisely, assume that the pairs (A, B_1) and (C_1, A) of the continuous-time system (1) are stabilizable and detectable, respectively, and that the pairs $(\hat{A}, \hat{B}_2)$ and $(\hat{C}_2, \hat{A})$ of the discrete system (11) are stabilizable and detectable. Then the performance bound $J(\mathcal{K}) < \gamma$ holds on $[0, \infty)$ if and only if the discrete Riccati equations (47) and (50) have corresponding stationary stabilizing solutions such that condition (*iii*) of Theorem 3 holds for all t, see [19].

V. A WORST-CASE SAMPLING APPROACH

The procedures described in Sections III and IV lead to different finite-dimensional discrete characterizations of the H_∞-optimal sampled-data control problem, cf. Theorem 1, and Theorem 3 and Lemma 7, respectively. This indicates that the discretization comprises an essential step in the solution methods. It is therefore well motivated to study the discretization process which is involved in more detail. In this section we study the discretization step in terms of a worst-case sampling approach [35]. The approach clarifies the connection between the discrete representations of the loop-shifting result of Theorem 1 and the characterizations described in Section IV.

Consider the lifted discrete system representation (11). In analogy with the standard finite-dimensional discrete H_∞ problem, a necessary condition for the performance bound $J(\mathcal{K}) < \gamma$ is $\|\hat{D}_{11}\| < \gamma$. Equivalently, it is required that for any bounded $\hat{x}_k$, $\hat{u}_k$, the quadratic function

$$\begin{aligned} L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k) &:= \|\hat{z}_k\|_2^2 - \gamma^2 \|\hat{w}_k\|_2^2 \\ &= \int_0^h (\hat{z}_k'\hat{z}_k - \gamma^2 \hat{w}_k'\hat{w}_k)\,dt \end{aligned} \tag{51}$$

has a bounded supremum in $\hat{w}_k$. In this case, the maximum of (51) is achieved by a unique $\hat{w}_k^o$ in $L_2[0,h)$. Depending on whether the maximizing strategy is represented in open or closed loop form, we obtain two different expressions for $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$.

Introduce the symmetric positive semidefinite matrix

$$P(t) := \begin{bmatrix} P_{11} & P_{12} \\ P_{12}' & P_{22} \end{bmatrix}(t) \tag{52}$$

which satisfies the following Riccati differential equation (cf. Eq (26))

$$\begin{aligned} -\dot{P}(t) &= A_e' P(t) + P(t) A_e + \gamma^{-2} P(t) B_{e1} B_{e1}' P(t) + C_{e1}' C_{e1}, \\ P(h) &= 0, \ t \in [0, h] \end{aligned} \tag{53}$$

where the matrices A_e, B_{e1} and C_{e1} are defined by Eq. (24).

Lemma 8. *Consider $\hat{z}_k \in L_2[0,h)$ defined by Eq. (11b). Then we have the following characterizations of the quadratic function $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$:*

(a) $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$ has a bounded supremum in $\hat{w}_k$ if and only if the Riccati differential equation (53) has a bounded solution on $[0,h]$. In this case, the maximum of $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$ is achieved by the optimal closed-loop strategy

$$\hat{w}_k^o(t) := \gamma^{-2} B_{e1}' P(t) \begin{bmatrix} x(kh+t) \\ \hat{u}_k \end{bmatrix}, \ t \in [0,h). \tag{54}$$

Moreover, (51) can be expressed as

$$\|\hat{z}_k\|_2^2 - \gamma^2 \|\hat{w}_k\|_2^2 = [\hat{x}_k' \quad \hat{u}_k'] P(0) \begin{bmatrix} \hat{x}_k \\ \hat{u}_k \end{bmatrix} - \gamma^2 \int_0^h \tilde{w}_k'(t) \tilde{w}_k(t) dt \tag{55}$$

where $\tilde{w}_k(t)$ is the pointwise deviation from the optimal closed-loop strategy, i.e.,

$$\tilde{w}_k(t) := \hat{w}_k(t) - \gamma^{-2} B_{e1}' P(t) \begin{bmatrix} x(kh+t) \\ \hat{u}_k \end{bmatrix}, \ t \in [0,h). \tag{56}$$

(b) $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$ has a bounded supremum in $\hat{w}_k$ if and only if $\|\hat{D}_{11}\| < \gamma$. In this case, the maximum of $L_k(\hat{w}_k, \hat{x}_k, \hat{u}_k)$ is achieved by the optimal open-loop strategy

$$\hat{w}_k^o := \gamma^{-2} (I - \gamma^{-2} \hat{D}_{11}^* \hat{D}_{11})^{-1} \hat{D}_{11}^* [\hat{C}_1 \quad \hat{D}_{12}] \begin{bmatrix} \hat{x}_k \\ \hat{u}_k \end{bmatrix} \tag{58}$$

where $\hat{D}_{11}^*$ denotes the adjoint operator. Moreover, (51) can be expressed as

$$\begin{aligned} \|\hat{z}_k\|_2^2 - \gamma^2 \|\hat{w}_k\|_2^2 = [\hat{x}_k' \ \hat{u}_k'] \begin{bmatrix} \hat{C}_1^* \\ \hat{D}_{12}^* \end{bmatrix} (I - \gamma^{-2} \hat{D}_{11} \hat{D}_{11}^*)^{-1} [\hat{C}_1 \ \hat{D}_{12}] \begin{bmatrix} \hat{x}_k \\ \hat{u}_k \end{bmatrix} \\ - \gamma^2 < \hat{w}_k - \hat{w}_k^o, (I - \gamma^{-2} \hat{D}_{11}^* \hat{D}_{11})(\hat{w}_k - \hat{w}_k^o) > \end{aligned} \tag{59}$$

where $< \cdot, \cdot >$ *denotes the inner product defined on* $L_2[0, h)$.

Proof: (a) See the proof of Lemma 2. *(b)* The result follows from standard linear operator theory and a completion of squares argument [35]. □

Lemma 8 provides two ways of introducing a variable transformation, which takes the H_∞ sampled-data control problem to a finite-dimensional discrete H_∞ problem. From Lemma 8(a) we have the following characterization of $J(\mathcal{K})$ in terms of a finite-dimensional discrete system.

Theorem 4. *Consider the sampled-data system described by Eq. (11). Define the finite-dimensional discrete system*

$$\begin{aligned} \hat{x}_{k+1} &= \tilde{A}\hat{x}_k + \tilde{B}_2\hat{u}_k + \tilde{B}_1\tilde{v}_k, \ \hat{x}_0 = 0 \\ \tilde{z}_k &= \tilde{C}_1\hat{x}_k + \tilde{D}_{12}\hat{u}_k \\ \hat{y}_k &= \hat{C}_2\hat{x}_k + \hat{D}_{21}\hat{v}_k \end{aligned} \tag{60}$$

where

$$\tilde{A} = \Phi(h,0) \ , \ \tilde{B}_2 = \int_0^h \Phi(h,\lambda)[B_2 + \gamma^{-2}B_1B_1'P_{12}(\lambda)]d\lambda$$

$$\tilde{B}_1\tilde{B}_1' = \int_0^h \Phi(h,\lambda)B_1B_1'\Phi'(h,\lambda)d\lambda$$

$$\begin{bmatrix} \tilde{C}_1' \\ \tilde{D}_{12}' \end{bmatrix} [\tilde{C}_1 \quad \tilde{D}_{12}] = P(0)$$

and $\Phi(\cdot,\cdot)$ *is the state transition matrix associated with* $A+\gamma^{-2}B_1B_1'P_{11}(t)$. *Then for any control law* $\hat{u} = \mathcal{K}\hat{y}$ *the following are equivalent:*

(i) $\mathcal{K}$ *stabilizes the system (11) and* $J(\mathcal{K}) < \gamma$,

(ii) $\mathcal{K}$ *stabilizes the system (60) and*

$$\sup\left\{\frac{\|\tilde{z}\|_{l_2}}{[\|\tilde{v}\|_{l_2}^2 + \|\hat{v}\|_{l_2}^2]^{1/2}}\right\} < \gamma. \tag{61}$$

Proof: Assume first that $\mathcal{K}$ internally stabilizes (11) and (60). From Lemma 8(a), introduce the variable $\tilde{w}_k \in L_2[0,h)$ defined by Eq. (56), and

$$\tilde{z}_k := \tilde{C}_1\hat{x}_k + \tilde{D}_{12}\hat{u}_k. \tag{62}$$

Then the sampled-data system is described by

$$\begin{aligned} \dot{x}(kh+t) &= [A + \gamma^{-2}B_1B_1'P_{11}(t)]x(kh+t) + [B_2 + \gamma^{-2}B_1B_1'P_{12}(t)]\hat{u}_k \\ &\quad + B_1\tilde{w}_k(t) \ , \ t \in [0,h) \ , \ x(0) = 0 \\ \tilde{z}_k &= \tilde{C}_1x(kh) + \tilde{D}_{12}\hat{u}_k \\ \hat{y}_k &= \hat{C}_2x(kh) + \hat{D}_{21}\hat{v}_k. \end{aligned} \tag{63}$$

By Lemma 8(a), we have $\|\tilde{z}\|_2^2 - \gamma^2\|\tilde{w}\|_2^2 = \|\hat{z}\|_2^2 - \gamma^2\|\hat{w}\|_2^2$. Integrating (63) over the sampling interval gives for the state

$$x(kh+h) = \tilde{A}x(kh) + \tilde{B}_2\hat{u}_k + \int_0^h \Phi(h,\tau)B_1\tilde{w}_k(\tau)d\tau.$$

The norm result of the theorem then follows from Lemma C.1(b) in Appendix C, see also [26, 23].

In order to prove the stability part of the theorem, assume first that $\mathcal{K}$ internally stabilizes (11), and consider the system (11) with input $\hat{w}^o$ defined by (54). Let $\bar{x}_k$ denote the combined state of the system and the controller at time kh. Then $J(\mathcal{K}) < \gamma$ implies the existence of $M_1 > 0$ such that for any initial time k_0h and initial state $\bar{x}_{k_0}$,

$$\|\hat{z}\|_{l_2(k_0,\infty)}^2 - \gamma^2\|\hat{w}^o\|_{l_2(k_0,\infty)}^2 \le M_1^2\|\bar{x}_{k_0}\|^2.$$

From (54), (55) and the continuity of $P(t)$ it also follows that there exists $M_2 > 0$ such that

$$\|\hat{w}^o\|_{l_2(k_0,\infty)}^2 \le M_2^2\left[\|\hat{z}\|_{l_2(k_0,\infty)}^2 - \gamma^2\|\hat{w}^o\|_{l_2(k_0,\infty)}^2\right].$$

Since $\mathcal{K}$ stabilizes the system (11), there exists $\delta > 0$ such that

$$\|\bar{x}\|_{l_2(k_0,\infty)} \le \delta\left[\|\hat{w}^o\|_{l_2(k_0,\infty)} + \|\bar{x}_{k_0}\|\right].$$

Hence it follows that

$$\|\bar{x}\|_{l_2(k_0,\infty)} \le \delta(M_1M_2+1)\|\bar{x}_{k_0}\|.$$

As the inequality holds for any initial state $\bar{x}_{k_0}$, the result implies that the closed-loop system (11) with input $\hat{w}^o$ is internally stable, or equivalently, that the controller $\mathcal{K}$ stabilizes the system (60). The reverse result can be shown analogously. □

A similar result is obtained from Lemma 8(b).

Theorem 5. *Consider the sampled-data system described by Eq. (11). Define the finite-dimensional discrete system*

$$\begin{aligned} \hat{x}_{k+1} &= \acute{A}\hat{x}_k + \acute{B}_2\hat{u}_k + \acute{B}_1\hat{v}_k, \quad \hat{x}_0 = 0 \\ \acute{z}_k &= \acute{C}_1\hat{x}_k + \acute{D}_{12}\hat{u}_k \\ \hat{y}_k &= \hat{C}_2\hat{x}_k + \hat{D}_{21}\hat{v}_k \end{aligned} \tag{64}$$

where

$$\begin{aligned}
\acute{A} &= \hat{A} + \gamma^{-2}\hat{B}_1\hat{D}_{11}^*(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*)^{-1}\hat{C}_1 \\
\acute{B}_2 &= \hat{B}_2 + \gamma^{-2}\hat{B}_1\hat{D}_{11}^*(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*)^{-1}\hat{D}_{12} \\
\acute{B}_1\acute{B}_1' &= \hat{B}_1(I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11})^{-1}\hat{B}_1^* \\
\begin{bmatrix} \acute{C}_1' \\ \acute{D}_{12}' \end{bmatrix}[\acute{C}_1 \quad \acute{D}_{12}] &= \begin{bmatrix} \hat{C}_1^* \\ \hat{D}_{12}^* \end{bmatrix}(I - \gamma^{-2}\hat{D}_{11}\hat{D}_{11}^*)^{-1}[\hat{C}_1 \quad \hat{D}_{12}].
\end{aligned}$$

Then for any control law $\hat{u} = \mathcal{K}\hat{y}$ the following are equivalent:
(i) $\mathcal{K}$ stabilizes the system (11) and $J(\mathcal{K}) < \gamma$,
(ii) $\mathcal{K}$ stabilizes the system (64) and

$$\sup\left\{\frac{\|\acute{z}\|_{l_2}}{[\|\acute{v}\|_{l_2}^2 + \|\hat{v}\|_{l_2}^2]^{1/2}}\right\} < \gamma. \tag{65}$$

Proof: Assume first that $\mathcal{K}$ stabilizes both (11) and (64). From Lemma 8(b), introduce the variables

$$\begin{aligned}
\acute{w}_k &:= (I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11})^{1/2}(\hat{w}_k - \hat{w}_k^o) \\
\acute{z}_k &:= \acute{C}_1\hat{x}_k + \acute{D}_{12}\hat{u}_k.
\end{aligned} \tag{66}$$

Then $\|\acute{z}\|_2^2 - \gamma^2\|\acute{w}\|_2^2 = \|\hat{z}\|_2^2 - \gamma^2\|\hat{w}\|_2^2$, and

$$\hat{x}_{k+1} = \acute{A}\hat{x}_k + \acute{B}_2\hat{u}_k + \hat{B}_1(I - \gamma^{-2}\hat{D}_{11}^*\hat{D}_{11})^{-1/2}\acute{w}_k.$$

The norm result of the theorem then follows by applying Lemma C.1(b) in Appendix C.

The stability part of the theorem can be shown in the same way as in Theorem 4. □

It is interesting to compare the discrete characterizations of Theorems 4 and 5 with other discrete representations of the H_∞ sampled-data control problem. Theorem 4 is similar to the characterizations obtained by game-theory type methods [11, 12, 23], cf. Section IV. The difference is, that in the game-theory based methods, the Riccati differential equation (53) over a single sampling interval is imbedded as part of the mixed discrete/continuous Riccati equation (26) over the whole control time. The characterization in Theorem 5 is identical to the one obtained via a loop-shifting approach in [15], cf. Theorem 1.

It is clear from the construction of the discrete systems of Theorems 4 and 5 that they are closely connected. Bamieh and Pearson [15] give explicit

formulae for the matrices in Eq. (64). Here we give alternative formulae, derived for the system matrices in (60). The formulae are based on the optimal control characterization of Lemma 8(a). This approach results in very simple formulae, requiring only one matrix exponential.

Recall the $2(n+m_2)\times 2(n+m_2)$ matrix exponential $\Pi(t)$ defined by Eq. (46). Note that it follows from the structure of the matrices A_e and B_{e1} that Π_{11} and Π_{12} can be partitioned accordingly as

$$\Pi_{11}=\begin{bmatrix}\Pi_{1111} & \Pi_{1112}\\ 0 & I\end{bmatrix},\quad \Pi_{12}=\begin{bmatrix}\Pi_{1211} & 0\\ 0 & 0\end{bmatrix}. \tag{67}$$

Lemma 9. *The discrete systems of Theorems 4 and 5 are identical in the sense that*

$$\begin{gathered}\tilde{A}=\acute{A}\ ,\ \tilde{B}_1\tilde{B}_1'=\acute{B}_1\acute{B}_1'\ ,\ \tilde{B}_2=\acute{B}_2\ ,\\ \begin{bmatrix}\tilde{C}_1'\\ \tilde{D}_{12}'\end{bmatrix}[\tilde{C}_1\quad \tilde{D}_{12}]=\begin{bmatrix}\acute{C}_1'\\ \acute{D}_{12}'\end{bmatrix}[\acute{C}_1\quad \acute{D}_{12}].\end{gathered} \tag{68}$$

The system matrices are given by

$$\begin{gathered}\tilde{A}=\Pi_{1111}(h)^{-1},\tilde{B}_2=-\tilde{A}\Pi_{1112}(h),\\ \tilde{B}_1\tilde{B}_1'=-\gamma^2\tilde{A}\Pi_{1211}(h)\ ,\\ \begin{bmatrix}\tilde{C}_1'\\ \tilde{D}_{12}'\end{bmatrix}[\tilde{C}_1\quad \tilde{D}_{12}]=\Pi_{21}(h)\Pi_{11}(h)^{-1}\end{gathered} \tag{69}$$

where $\Pi(h)$ is the exponential matrix defined by Eq. (46) and the partition (67).

Proof: The first part of the lemma follows from the construction of the discrete systems (60) and (64). For the second part, we need to evaluate the matrices in Theorem 4 in terms of the matrix $\Pi(h)$. First note that the matrix $P(t)$ of Eq. (53) can be expressed in terms of the matrix $\Pi(\cdot)$ as

$$P(t)=\Pi_{21}(h-t)\Pi_{11}(h-t)^{-1}\ ,\ t\in[0,h]. \tag{70}$$

The matrix $\Pi(h)$ is the transition matrix from time h to time 0 of the system

$$\begin{bmatrix}\dot{p}_1(t)\\ \dot{p}_2(t)\end{bmatrix}=\begin{bmatrix}A_e & \gamma^{-2}B_{e1}B_{e1}'\\ -C_{e1}'C_{e1} & -A_e'\end{bmatrix}\begin{bmatrix}p_1(t)\\ p_2(t)\end{bmatrix}. \tag{71}$$

Introduce the variable transformation

$$\begin{bmatrix}r(t)\\ s(t)\end{bmatrix}:=\begin{bmatrix}I & 0\\ -P(t) & I\end{bmatrix}\begin{bmatrix}p_1(t)\\ p_2(t)\end{bmatrix}. \tag{72}$$

Then we have from Eqs. (71) and (53),

$$\begin{bmatrix} \dot{r}(t) \\ \dot{s}(t) \end{bmatrix} = \begin{bmatrix} A_e + \gamma^{-2} B_{e1} B_{e1}' P(t) & \gamma^{-2} B_{e1} B_{e1}' \\ 0 & -(A_e + \gamma^{-2} B_{e1} B_{e1}' P(t))' \end{bmatrix} \begin{bmatrix} r(t) \\ s(t) \end{bmatrix}.$$

Using the upper triangular structure, it is easy to see that

$$\begin{bmatrix} r(0) \\ s(0) \end{bmatrix} = \begin{bmatrix} \Phi_e(0,h) & -\Phi_e(0,h) W_e \\ 0 & \Phi_e'(0,h) \end{bmatrix} \begin{bmatrix} r(h) \\ s(h) \end{bmatrix} \tag{73}$$

where $\Phi_e(\cdot,\cdot)$ is the transition matrix associated with $A_e + \gamma^{-2} B_{e1} B_{e1}' P(t)$, and W_e is the controllability Gramian,

$$W_e := \gamma^{-2} \int_0^h \Phi_e(h,\lambda) B_{e1} B_{e1}' \Phi_e'(h,\lambda) d\lambda. \tag{74}$$

From Eqs. (70), (71) and (72) we have

$$\begin{aligned} \begin{bmatrix} r(0) \\ s(0) \end{bmatrix} &= \begin{bmatrix} I & 0 \\ -P(0) & I \end{bmatrix} \Pi(h) \begin{bmatrix} I & 0 \\ P(h) & I \end{bmatrix} \begin{bmatrix} r(h) \\ s(h) \end{bmatrix} \\ &= \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ 0 & \Pi_{22} - \Pi_{21} \Pi_{11}^{-1} \Pi_{12} \end{bmatrix} (h) \begin{bmatrix} r(h) \\ s(h) \end{bmatrix}. \end{aligned} \tag{75}$$

By Eqs. (73) and (75),

$$\Phi_e(h,0) = \Pi_{11}(h)^{-1}, \quad W_e = -\Phi_e(h,0) \Pi_{12}(h). \tag{76}$$

Invoking the structure of the matrices A_e and B_{e1}, we have

$$A_e + \gamma^{-2} B_{e1} B_{e1}' P(t) = \begin{bmatrix} A + \gamma^{-2} B_1 B_1' P_{11}(t) & B_2 + \gamma^{-2} B_1 B_1' P_{12}(t) \\ 0 & 0 \end{bmatrix}.$$

It follows that the associated state transition matrix $\Phi_e(h,0)$ and Gramian matrix W_e can be expressed in terms of the matrices $\tilde{A}$, $\tilde{B}_1$ and $\tilde{B}_2$ defined in Theorem 4 as

$$\Phi_e(h,0) = \begin{bmatrix} \tilde{A} & \tilde{B}_2 \\ 0 & I \end{bmatrix}, \quad W_e = \begin{bmatrix} \tilde{B}_1 \tilde{B}_1' & 0 \\ 0 & 0 \end{bmatrix}. \tag{77}$$

The formulas to be proved follow from Eqs. (77), (76), (67), and (70). □

VI. A DUAL-RATE CONTROL PROBLEM

An important generalization of the standard sampled-data control system described in Section II is the dual-rate control problem, where the sampling and hold elements operate at different rates. The rate by which the

measured signal can be sampled may for example be limited, whereas the hold function can operate at a faster rate. A more complex situation still is the multirate control problem, where various outputs are sampled at different rates, and various control inputs are generated by hold elements which operate at different rates. Provided the sampling and control rates are synchronized, i.e. they are rationally related, the multirate control system is periodic, and the sampled-data dual-rate and multirate control problems can be solved by a generalization of the lifting technique described in Section III, in which the system is described in terms of an equivalent lifted discrete system [38, 39]. A possible drawback of this approach is, however, that the period of the multirate system may be long, which results in a lifted system description having high order.

In this section we study the H_∞-optimal control of dual-rate sampled-data systems, in which the sampling and hold elements operate at different rates. Instead of using the lifting technique, we apply the game-theory based approach of Section IV. The solution obtained in this way does not require that the dual-rate system be periodic. Hence the rates of the sampler and hold elements may be asynchronously related. The H_2-optimal controller for such asynchronous dual-rate systems has been obtained in [40], but the H_∞-optimal control problem does not appear to have been studied for this case before.

In this section we consider the finite-dimensional linear plant

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B_1 w(t) + B_2 u(t) \ , \ x(0) = 0 \\ z(t) &= C_1 x(t) + D_{12} u(t) \ , \ t \in [0, T] \\ \hat{y}_i &= \hat{C}_2 x(is) + \hat{D}_{21} \hat{v}_i \ , \ i = 0, 1, \ldots, [T/s] \end{aligned} \tag{78}$$

where $T > 0$ is the control horizon. The output in (78) is sampled at time instants $\{is\}$, where $s > 0$. It is assumed that the control signal $u(t)$ is generated by a zero-order hold device according to Eq. (3), with the period h. The sampler and hold device are assumed to have different periods, i.e. $s \neq h$. In the dual-rate sampled-data control problem, we consider discrete causal control laws $\{\hat{u}_k\} = \mathcal{K}(\{\hat{y}_i\})$, such that $\hat{u}_k$ is a function of only past measurements $\{\hat{y}_i : is \leq kh\}$.

The dual-rate H_∞ control problem consists of finding a discrete causal controller such that the worst-case performance bound (6), $J(\mathcal{K}) < \gamma$, is achieved for a given positive constant γ. The problem can be solved using the approach of Section IV and taking into account the dual rate nature of the system. By Lemma 3, the dual-rate H_∞ problem has a solution if and only if the Riccati differential equation with jumps (26) has a bounded solution on $[0, T]$, and there exists a causal estimator $\{\hat{r}_{e,k}\} = \mathcal{F}(\{\hat{y}_i\})$

for the system (31), which achieves the performance bound (35). The existence of such an estimator is given by the following lemma, which is a straightforward modification of Lemma 4 to the dual-rate problem.

Lemma 10. *Consider the system*

$$\begin{aligned} \dot{x}(t) &= \bar{A}(t)x(t) + \bar{B}_1(t)v(t), \ x(0) = 0 \\ \hat{r}_k &= \bar{C}_{1,k}x(kh) \\ \hat{y}_i &= \hat{C}_2 x(ih) + \hat{D}_{21}\hat{v}_i \ , \ t \in [0,T] \end{aligned} \tag{79}$$

Assume that assumption (A2) holds. There exists a discrete causal filter $\{\hat{r}_{e,k}\} = \mathcal{F}(\{\hat{y}_i\})$ *such that*

$$\|\hat{r} - \hat{r}_e\|_{l_2}^2 < \gamma^2[\|v\|_{L_2}^2 + \|\hat{v}\|_{l_2}^2] \tag{80}$$

holds for all $(v, \hat{v}) \neq (0,0)$, *if and only if there exists a bounded symmetric positive semidefinite matrix* $N(t)$ *which satisfies the Riccati differential equation with jumps,*

$$\begin{aligned} \dot{N}(t) &= \bar{A}(t)N(t) + N(t)\bar{A}'(t) + \bar{B}_1(t)\bar{B}_1'(t) \ , \ t \neq kh, is \\ N(is) &= N(is^-)[I + \hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2 N(is^-)]^{-1} \ , \ \text{if } is \neq kh \\ N(kh) &= N(kh^-)\Sigma_k^{-1} \\ \Sigma_k &= I - \gamma^{-2}\bar{C}_{1,k}'\bar{C}_{1,k}N(kh^-) \ , \ \text{if } kh \neq is \\ \Sigma_k &= I + [\hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2 - \gamma^{-2}\bar{C}_{1,k}'\bar{C}_{1,k}]N(kh^-) \ , \ \text{if } kh = is, \\ N(0^-) &= 0 \ , \ t \in [0,T], \end{aligned} \tag{81}$$

such that the matrices Σ_k, $k = 0, 1, \ldots, [T/h]$, *have only positive eigenvalues. In this case, the estimator*

$$\begin{aligned} \dot{\hat{x}}(t) &= \bar{A}(t)\hat{x}(t) \ , \ t \neq is \\ \hat{x}(is) &= \hat{x}(is^-) \\ &\quad + N(is^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2 N(is^-)\hat{C}_2']^{-1}\big(y_i - \hat{C}_2\hat{x}(is^-)\big) \\ \hat{r}_{e,k} &= \bar{C}_{1,k}\hat{x}(kh) \ , \ \hat{x}(0^-) = 0 \end{aligned} \tag{82}$$

achieves the performance bound (80).

It is useful to introduce the following relation between the closed-loop H_∞-optimal filtering result of Lemma 10 and the open-loop filtering result of Lemma 5 associated with the system (78).

Lemma 11. *Assume that the Riccati differential equation with jumps (26) has a bounded solution* $S(t)$ *on* $[0,T]$ *with boundary condition* $S(T) = 0$.

Then there exists a bounded symmetric positive semidefinite solution $N(t)$ to the Riccati differential equation with jumps (81), where $\bar{A}(t)$, $\bar{B}_1(t)$, and $\bar{C}_{1,k}$ are defined by Eq. (32), if and only if the following conditions hold:

(i) There exists a bounded positive semidefinite matrix $Q(t)$ which satisfies the Riccati differential equation with jumps,

$$\begin{aligned} \dot{Q} &= AQ + QA' + \gamma^{-2}QC_1'C_1Q + B_1B_1' \ , \ t \in [is, is+s) \\ Q(is) &= Q(is^-) - Q(is^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2Q(is^-)\hat{C}_2']^{-1}\hat{C}_2Q(is^-) \\ Q(0^-) &= 0 \ , \ i = 0, 1, \ldots, [T/s] \end{aligned} \tag{83}$$

(ii) $\rho(S_{11}(t)Q(t)) < \gamma^2$ for all $t \in [0, T]$.

When these conditions are satisfied, the matrix $N(t)$ of Eq. (81) is given by

$$N(t) = Q(t)\big(I - \gamma^{-2}S_{11}(t)Q(t)\big)^{-1}. \tag{84}$$

We can now state the main result of this section, which gives the solution of the dual-rate sampled-data H_∞ control problem. Let $\Phi(\cdot,\cdot)$ denote the state transition matrix associated with the system matrix $\bar{A}(\cdot)$ of the system (79), and define

$$\Gamma(t, t') := \int_{t'}^{t} \Phi(t, \tau)\bar{B}_2(\tau)d\tau. \tag{85}$$

where $\bar{A}(t)$ and $\bar{B}_2(t)$ are defined by Eq. (32). Introduce also the definitions

$$k_p(t) := \max\{k : k \text{ integer}, \ kp < t\} \tag{86}$$

and

$$\underline{t}(t) := \max\{k_h(t)h, k_s(t)s\}. \tag{87}$$

We now have the following characterization of the solution to the dual rate H_∞-optimal control problem (cf. Theorem 3).

Theorem 6. *Consider the sampled-data system defined by Eq. (78) and the hold function (3). Assume that assumptions (A1) and (A2) hold. There exists a discrete causal controller $\hat{u} = \mathcal{K}\hat{y}$ which achieves the performance bound $J(\mathcal{K}) < \gamma$ on $[0, T]$ if and only if the following conditions are satisfied:*

(i) The Riccati differential equation with jumps (26) has a bounded positive semidefinite solution $S(t)$ on $[0, T]$ with boundary condition $S(T) = 0$,

(ii) the Riccati differential equation (83) has a bounded positive semidefinite solution $Q(t)$ on $[0, T]$, and

(iii) $\rho(S_{11}(t)Q(t)) < \gamma^2$ for all $t \in [0, T]$.

Moreover, when conditions (i)–(iii) are satisfied, a controller which achieves the performance bound $J(\mathcal{K}) < \gamma$ is given by

$$\begin{aligned}
\hat{x}(is^-) &= \Phi(is, \underline{t}(is))\hat{x}(\underline{t}(is)) + \Gamma(is, \underline{t}(is))\hat{u}_{k_h(is)} \ , \ \hat{x}(0) = 0 \\
\hat{x}(is) &= \hat{x}(is^-) \\
&\quad + N(is^-)\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2 N(is^-)\hat{C}_2']^{-1}(y_i - \hat{C}_2\hat{x}(is^-)) \\
\hat{x}(kh) &= \Phi(kh, \underline{t}(kh))\hat{x}(\underline{t}(kh)) + \Gamma(kh, \underline{t}(kh))\hat{u}_{k_h(kh)} \\
\hat{u}_k &= -S_{22}^{-1}(kh)S_{12}'(kh)\hat{x}(kh)
\end{aligned} \tag{88}$$

where $N(t)$ is given by Eq. (84).

Note that the Riccati equations with jumps (26) and (83) can be reduced to discrete Riccati equations according to Lemma 7. In analogy with Lemma 9, we can derive explicit formulas for the matrices $\Phi(\cdot,\cdot)$ and $\Gamma(\cdot,\cdot)$ which appear in the expression (88) for the H_∞-optimal dual-rate controller. For this purpose, consider the $2(n+m_2) \times 2(n+m_2)$ exponential matrix (46), and recall that the blocks Π_{11} and Π_{12} can be partitioned according to Eq. (67).

Lemma 12. *For $kh \le t' \le t \le kh + h$, the state transition matrix $\Phi(t,t')$ associated with the system matrix $\bar{A}(t)$ and the matrix $\Gamma(t,t')$ in Eq. (85) are given by*

$$\begin{aligned}
\Phi(t,t') &= \Pi_{1111}(\Delta t) + \Pi_{1211}(\Delta t)S_{11}(t') \\
\Gamma(t,t') &= \Pi_{1112}(\Delta t) + \Pi_{1211}(\Delta t)S_{12}(t')
\end{aligned} \tag{89}$$

where $\Delta t := t' - t$ and the matrices $S_{11}(\cdot)$ and $S_{12}(\cdot)$ are defined by Eq. (47) and the partition (25).

Proof: The proof is analogous with Lemma 9. □

VII. OPTIMAL SAMPLING PREFILTER FOR H_∞ CONTROL

In the above treatment, the sampling prefilter has been assumed given *a priori*, and is included as part of the system equations (1). It is, however, also of interest to study the problem of optimal sampling prefilter design in the context of H_∞ control. In linear quadratic gaussian control, for example, it is known that the optimal sampling prefilter consists of a Kalman filter for the continuous-time plant [6].

In this section we consider the sampled-data H_∞ control problem in the case when the selection of the sampling prefilter is included as part of

the design problem. It is convenient to write the system equations in the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B_1 w(t) + B_2 u(t) \ , \ x(0) = 0 \\ z(t) &= C_1 x(t) + D_{12} u(t) \\ y(t) &= C_2 x(t) + D_{21} w(t) \ , \ t \in [0, Nh] \end{aligned} \tag{90}$$

where the measurement disturbance, which is now a continuous-time signal, is included as part of the vector $w(t)$. We assume that assumption $(A1)$ holds and the matrices D_{21} and B_1 satisfy

$$D_{21} D_{21}' > 0 \tag{91a}$$

$$D_{21} B_1' = 0. \tag{91b}$$

In contrast to the discrete measurement case, Eq. (2), it is now necessary to assume that the matrix D_{21} has full row rank for the problem to be well posed. Assumption (91b) implies that the plant and measurement noise are independent. This assumption is not restrictive, but is made for convenience, as it results in significantly simpler formulas.

The controller now consists of a sampling prefilter to process the continuous output $y(t)$, and a discrete control law which acts on sampled outputs of the sampling prefilter. The control signal u is thus given by

$$u = \mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}y, \tag{92}$$

where $\mathcal{F} : L_2 \to L_2$ is a continuous-time causal sampling prefilter, $\mathcal{S} : L_2 \to l_2$ is the sampling operator defined by

$$(\mathcal{S}v)(k) = v(kh), \ h > 0, \tag{93}$$

$\mathcal{K} : l_2 \to l_2$ is a discrete causal transfer function, and $\mathcal{H} : l_2 \to L_2$ is the hold function (3),

$$(\mathcal{H}\hat{u})(t) = \hat{u}_k \ , t \in [kh, kh + h). \tag{94}$$

Note that the sampling operator $\mathcal{S}$ is an unbounded operator, whereas the hold function $\mathcal{H}$ is a bounded operator.

The performance bound is defined in analogy with Eqs. (5) and (6). The problem to be considered is thus to design a controller (92), such that

$$\sup\{\frac{\|z\|_{L_2}}{\|w\|_{L_2}}\} < \gamma, \tag{95}$$

where the supremum is taken over $w \neq 0$ in $L_2[0, Nh]$.

In [41], a sampling prefilter was designed in a way analogous to the approach in Section IV, in which the H_∞-optimal state-feedback controller (27a) of Theorem 2 is first obtained, and using Lemma 3, an H_∞-optimal sampling prefilter $\hat{r}_e = \mathcal{S}\mathcal{F}y$ $(L_2 \to l_2)$ is then constructed for the system (31). This procedure leads to a linear sampling prefilter which has the property that it is periodically time-varying, even in the infinite-horizon case when the plant is time-invariant. This is due to the fact that the estimated variable $\hat{r}_k$ in (31) is discrete, while the filter operates in continuous time. It is, however, possible to represent the controller (92) in such a way that the sampling prefilter $\mathcal{F}$ becomes time-invariant in the stationary, time-invariant case [42]. Here this result will be derived by constructing the controller in terms of the adjoint system, and transforming the final result back to the original system.

If there exists a controller of the form (92) which achieves the performance bound (95), then there exists a feedback law (92) with linear $\mathcal{F}$ and $\mathcal{K}$ which achieves (95) [41]. Moreover, the filter $\mathcal{F}$ is such that the operator $\mathcal{S}\mathcal{F}$ is bounded. We may therefore assume that the operators $\mathcal{F}$ and $\mathcal{K}$ are linear and $\mathcal{S}\mathcal{F}$ is bounded. Write the system (90) in operator notation as

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} \mathcal{G}_{11} & \mathcal{G}_{12} \\ \mathcal{G}_{21} & \mathcal{G}_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}. \tag{96}$$

The closed-loop system (90), (92) then becomes

$$z = \left[\mathcal{G}_{11} + \mathcal{G}_{12}(I - \mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}\mathcal{G}_{22})^{-1}\mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}\mathcal{G}_{21}\right] w. \tag{97}$$

The performance bound (95) is then equivalent to

$$\|\mathcal{G}_{11} + \mathcal{G}_{12}(I - \mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}\mathcal{G}_{22})^{-1}\mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}\mathcal{G}_{21}\| < \gamma, \tag{98}$$

or, in terms of the adjoint,

$$\|\mathcal{G}_{11}^* + \mathcal{G}_{21}^*(\mathcal{S}\mathcal{F})^*\mathcal{K}^*\mathcal{H}^*[I - \mathcal{G}_{22}^*(\mathcal{S}\mathcal{F})^*\mathcal{K}^*\mathcal{H}^*]^{-1}\mathcal{G}_{12}^*\| < \gamma. \tag{99}$$

It is straightforward to show that the adjoint system

$$\begin{bmatrix} q \\ s \end{bmatrix} = \begin{bmatrix} \mathcal{G}_{11}^* & \mathcal{G}_{21}^* \\ \mathcal{G}_{12}^* & \mathcal{G}_{22}^* \end{bmatrix} \begin{bmatrix} \nu \\ \eta \end{bmatrix} \tag{100}$$

has the state-space representation

$$\begin{aligned} -\dot{p} &= A'p + C_1'\nu + C_2'\eta\ ,\ p(Nh) = 0 \\ q &= B_1'p + D_{21}'\eta \\ s &= B_2'p + D_{12}'\nu\ ,\ t \in [0, Nh]. \end{aligned} \tag{101}$$

The feedback law

$$\eta = (\mathcal{SF})^* \mathcal{K}^* \mathcal{H}^* s \tag{102}$$

then gives

$$q = [\mathcal{G}_{11}^* + \mathcal{G}_{21}^* (\mathcal{SF})^* \mathcal{K}^* \mathcal{H}^* [I - \mathcal{G}_{22}^* (\mathcal{SF})^* \mathcal{K}^* \mathcal{H}^*]^{-1} \mathcal{G}_{12}^*] \nu. \tag{103}$$

Hence the problem of finding a controller (92) such that (98) holds is equivalent to the problem of finding an anticausal feedback (102) such that

$$\|q\|_2^2 - \gamma^2 \|\nu\|_2^2 < 0 \ , \ \text{all } \nu \neq 0, \tag{104}$$

holds for the adjoint system (101). This problem may be solved along standard lines. First, we have the following result.

Lemma 13. *Consider the adjoint system (101). Assume that there is a bounded symmetric positive semidefinite matrix $P(t)$ which satisfies the following Riccati differential equation associated with (101),*

$$\begin{aligned} \dot{P}(t) &= AP(t) + P(t)A' - P(t)C_2'(D_{21}D_{21}')^{-1}C_2P(t) \\ &\quad + \gamma^{-2}P(t)C_1'C_1P(t) + B_1B_1' \ , t \in [0, Nh], \\ P(0) &= 0. \end{aligned} \tag{105}$$

Then

$$\begin{aligned} \int_0^{Nh} (q'q - \gamma^2 \nu'\nu)dt = &\int_0^{Nh} (\eta - \eta^o)' D_{21} D_{21}' (\eta - \eta^o) dt \\ &- \gamma^2 \int_0^{Nh} (\nu - \nu^o)'(\nu - \nu^o) dt \end{aligned}$$

where

$$\eta^o(t) = -(D_{21}D_{21}')^{-1} C_2 P(t) p(t) \tag{106a}$$

$$\nu^o(t) = \gamma^{-2} C_1 P(t) p(t). \tag{106b}$$

Proof: The result is standard in dynamic game theory [33], and it can be shown in the same way as Lemma 2. □

From Lemma 13 the following state feedback result for the adjoint system (101) is obtained.

Lemma 14. *Consider the adjoint system (101). There exists a linear anticausal state feedback law $\eta = \mathcal{F}^* p$ such that (104) holds for the closed loop if and only if the Riccati equation (105) has a bounded symmetric positive*

semidefinite solution on $[0, Nh]$. *In that case, the state feedback law* (106*a*) *achieves the performance bound (104).*

Proof: This is a standard result in continuous-time H_∞ control theory [33, 34], and it can be proven in the same way as Theorem 2. □

Introduce the variable transformation

$$\tilde{q} := (D_{21}D_{21}')^{1/2}(\eta - \eta^o) \tag{107a}$$

$$\tilde{\nu} := \nu - \nu^o. \tag{107b}$$

Then the adjoint system (101) can be written

$$\begin{aligned} -\dot{p} &= (A + \gamma^{-2}P(t)C_1'C_1)'p + C_1'\tilde{\nu} + C_2'\eta \ , \ p(Nh) = 0 \\ \tilde{q} &= (D_{21}D_{21}')^{-1/2}C_2P(t)p(t) + (D_{21}D_{21}')^{1/2}\eta \\ s &= (B_2 + \gamma^{-2}P(t)C_1'D_{12})'p(t) + D_{12}'\tilde{\nu}. \end{aligned} \tag{108}$$

From Lemma 13, we have

$$\|\tilde{q}\|_2^2 - \gamma^2\|\tilde{\nu}\|_2^2 = \|q\|_2^2 - \gamma^2\|\nu\|_2^2.$$

Hence, the feedback law (102) achieves the performance bound (104) when applied to the adjoint system (101) if and only if the Riccati differential equation (105) has a bounded solution on $[0, Nh]$, and it achieves the performance bound

$$\|\tilde{q}\|_2^2 - \gamma^2\|\tilde{\nu}\|_2^2 < 0 \ , \ \text{all } \tilde{\nu} \neq 0, \tag{109}$$

when applied to (108). Now write (108) in operator form as

$$\begin{bmatrix} \tilde{q} \\ s \end{bmatrix} = \begin{bmatrix} \tilde{\mathcal{G}}_{11}^* & \tilde{\mathcal{G}}_{21}^* \\ \tilde{\mathcal{G}}_{12}^* & \tilde{\mathcal{G}}_{22}^* \end{bmatrix} \begin{bmatrix} \tilde{\nu} \\ \eta \end{bmatrix}. \tag{110}$$

With the feedback law (102), we have

$$\tilde{q} = \left[\tilde{\mathcal{G}}_{11}^* + \tilde{\mathcal{G}}_{21}^*(\mathcal{SF})^*\mathcal{K}^*\mathcal{H}^*[I - \tilde{\mathcal{G}}_{22}^*(\mathcal{SF})^*\mathcal{K}^*\mathcal{H}^*]^{-1}\tilde{\mathcal{G}}_{12}^*\right]\tilde{\nu}.$$

The performance bound (109) is thus equivalent to

$$\|\tilde{\mathcal{G}}_{11}^* + \tilde{\mathcal{G}}_{21}^*(\mathcal{SF})^*\mathcal{K}^*\mathcal{H}^*[I - \tilde{\mathcal{G}}_{22}^*(\mathcal{SF})^*\mathcal{K}^*\mathcal{H}^*]^{-1}\tilde{\mathcal{G}}_{12}^*\| < \gamma,$$

or

$$\|\tilde{\mathcal{G}}_{11} + \tilde{\mathcal{G}}_{12}(I - \mathcal{HKSF}\tilde{\mathcal{G}}_{22})^{-1}\mathcal{HKSF}\tilde{\mathcal{G}}_{21}\| < \gamma. \tag{111}$$

Now consider the adjoint of (110),

$$\begin{bmatrix} \tilde{z} \\ y \end{bmatrix} = \begin{bmatrix} \tilde{\mathcal{G}}_{11} & \tilde{\mathcal{G}}_{12} \\ \tilde{\mathcal{G}}_{21} & \tilde{\mathcal{G}}_{22} \end{bmatrix} \begin{bmatrix} \tilde{w} \\ u \end{bmatrix}. \tag{112}$$

Then, the problem of finding a feedback (102) such that the performance bound (104) holds for (101) is equivalent to finding a feedback (92),

$$u = \mathcal{H}\mathcal{K}\mathcal{S}\mathcal{F}y,$$

such that the performance bound (111) holds for the adjoint system (112), or equivalently,

$$\|\tilde{z}\|_2^2 - \gamma^2\|\tilde{w}\|_2^2 < 0 \ , \ \text{all } \tilde{w} \neq 0. \tag{113}$$

Now, the system (112) has the state space representation (cf. Eq.(108))

$$\begin{aligned}
\dot{\hat{x}}(t) &= \big(A + \gamma^{-2}P(t)C_1'C_1\big)\hat{x}(t) + P(t)C_2'(D_{21}D_{21}')^{-1/2}\tilde{w}(t) \\
&\quad + \big(B_2 + \gamma^{-2}P(t)C_1'D_{12}\big)u(t) \ , \ \hat{x}(0) = 0 \\
\tilde{z}(t) &= C_1\hat{x}(t) + D_{12}u(t) \\
y(t) &= C_2\hat{x}(t) + (D_{21}D_{21}')^{1/2}\tilde{w}(t).
\end{aligned} \tag{114}$$

Note that a knowledge of the output $y(t)$ implies that the disturbance $\tilde{w}(t)$, and hence the state $\hat{x}(t)$, is known, since

$$\tilde{w}(t) = (D_{21}D_{21}')^{-1/2}\big(y(t) - C_2\hat{x}(t)\big).$$

The problem of finding a controller (92) such that (113) holds for the system (114) thus reduces to a sampled-data H_∞ problem with complete state information for the system (114). This problem in turn can be solved according to Theorem 2.

To sum up, the sampled-data H_∞ problem with optimal sampling prefilter can be characterized according to the following lemma. Introduce the matrices

$$\begin{aligned}
\bar{A}_e(t) &:= \begin{bmatrix} A + \gamma^{-2}P(t)C_1'C_1 & B_2 + \gamma^{-2}P(t)C_1'D_{12} \\ 0 & 0 \end{bmatrix}, \\
\bar{B}_{e1}(t) &:= \begin{bmatrix} P(t)C_2'(D_{21}D_{21}')^{-1/2} \\ 0 \end{bmatrix}, \ \bar{C}_{e1} := [C_1 \quad D_{12}].
\end{aligned} \tag{115}$$

Lemma 15. *Consider the system (90). There exists a causal controller (92) which achieves the performance bound (95) if and only if the following conditions are satisfied:*

(i) there is a bounded solution $P(t)$ on $[0, Nh]$ to the Riccati equation (105),

(ii) there is a bounded matrix

$$\bar{S}(t) := \begin{bmatrix} \bar{S}_{11} & \bar{S}_{12} \\ \bar{S}_{12}' & \bar{S}_{22} \end{bmatrix}(t) \ , \ t \in [0, Nh), \tag{116}$$

which satisfies the following Riccati differential equation with jumps,

$$\begin{aligned}
-\dot{\bar{S}}(t) &= \bar{A}_e'(t)\bar{S}(t) + \bar{S}(t)\bar{A}_e(t) + \gamma^{-2}\bar{S}(t)\bar{B}_{e1}(t)\bar{B}_{e1}'(t)\bar{S}(t) \\
&\quad + \bar{C}_{e1}'\bar{C}_{e1}, \ t \neq kh, \\
\bar{S}(kh^-) &= \begin{bmatrix} \bar{S}_{11}(kh^-) & 0 \\ 0 & 0 \end{bmatrix}, \\
\bar{S}_{11}(kh^-) &= \bar{S}_{11}(kh) - \bar{S}_{12}(kh)\bar{S}_{22}^{-1}(kh)\bar{S}_{12}'(kh), \\
\bar{S}(Nh^-) &= 0 \ , \ k = N-1, \ldots, 1, 0.
\end{aligned} \tag{117}$$

Moreover, when conditions (i) and (ii) are satisfied, the performance bound (95) is achieved by the controller

$$\begin{aligned}
\dot{\hat{x}}(t) &= \left(A + \gamma^{-2}P(t)C_1'C_1\right)\hat{x}(t) + \left(B_2 + \gamma^{-2}P(t)C_1'D_{12}\right)u(t) \\
&\quad + P(t)C_2'(D_{21}D_{21}')^{-1}\left(y(t) - C_2\hat{x}(t)\right) \ , \ \hat{x}(0) = 0 \\
\hat{u}_k &= -\bar{S}_{22}^{-1}(kh)\bar{S}_{12}'(kh)\hat{x}(kh) \ , \ k = 0, 1, \ldots, N-1 \\
u(t) &= \hat{u}_k \ , \ t \in [kh, kh+h).
\end{aligned} \tag{118}$$

Note that the controller (118) has the form (92). In analogy with the sampled-data control problem described in Section IV, the controller (118) can be expressed in terms of the sampled-data state-feedback problem solution for the system (90) (Theorem 2) and the continuous-time filter of Lemma 14. The following theorem is analogous with Theorem 3.

Theorem 7. *Consider the system (90). A control law (92) which achieves the performance bound (95) exists if and only if the following conditions are satisfied:*

(i) The Riccati differential equation with jumps (26) has a bounded solution $S(t)$ on $[0, Nh)$,

(ii) the Riccati differential equation (105) has a bounded solution $P(t)$ on $[0, Nh)$, and

(iii) $\rho(S_{11}(t)P(t)) < \gamma^2$ for all $t \in [0, Nh)$.

Moreover, when the conditions (i)–(iii) hold, a controller which achieves the performance bound (95) is given by Eq. (118), where

$$\bar{S}(t) = S(t)\left(I - \gamma^{-2}\begin{bmatrix} P(t) & 0 \\ 0 & 0 \end{bmatrix} S(t)\right)^{-1}.$$

In analogy with Theorem 3, the infinite-horizon, stationary case may be obtained as a limiting case from Theorem 7 and the corresponding stationary stabilizing solutions of the Riccati equations (47) and (105), assuming stabilizability of the pairs $(\hat{A}, \hat{B}_2)$ and (A, B_1) in (11) and (90), respectively, and detectability of the pairs (C_2, A) and (C_1, A) in (90), cf. [42].

VIII. ROBUST STABILITY OF SAMPLED-DATA SYSTEMS

In analogy with the standard continuous-time and discrete-time H_∞ problems, the sampled-data H_∞ problem studied in the previous sections is closely related to the robust stability issue. The robust stability problem in sampled-data control has been studied in a number of papers. Chen and Francis [24] obtained a sufficient stability condition for additive plant uncertainties using a small gain argument. Sivashankar and Khargonekar [25] obtained sufficient and necessary stability conditions for the class of linear and time-varying uncertainties. Sufficient and necessary conditions for linear time-invariant uncertainties have been obtained for SISO system by Bai [43] and Dullerud and Glover [44], and in the general case by Dullerud and Glover 45].

In this section sufficient and necessary conditions for robust stability of sampled-data systems are given in terms of an H_∞-type performance bound. The conditions given in [25] are shown to be sufficient and necessary also in the case when the uncertainty is assumed to belong to the set of causal time-invariant (nonlinear) operators.

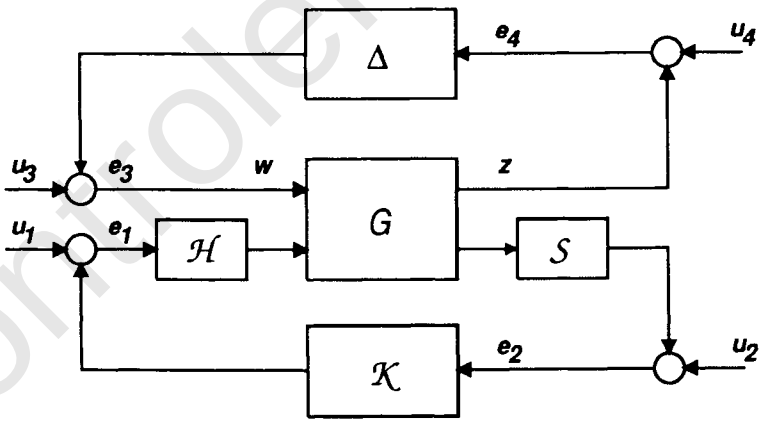


Fig. 2. Feedback connection for robust stability analysis.

Consider the sampled-data feedback system with plant uncertainty shown in Figure 2. A wide class of plant uncertainties can be rearranged into the form shown in the figure [25]. Here G denotes the nominal time-

invariant continuous-time plant assumed to be described by Eq. (1), $\mathcal{S}$ is the sampler described by Eq. (93), $\mathcal{K}$ is the discrete controller, Eq. (4), and $\mathcal{H}$ is the hold function described by Eq. (94). The operator $\Delta : L_2 \to L_2$ represents a plant uncertainty assumed to belong to some specified set. The feedback system in Figure 2 is defined to be robustly stable in the L_2-norm if the nominal system is internally asymptotically stable and the operator taking $(u_1, u_2, u_3, u_4) \in l_2 \oplus l_2 \oplus L_2 \oplus L_2$ to $(e_1, e_2, e_3, e_4) \in l_2 \oplus l_2 \oplus L_2 \oplus L_2$ is bounded for all Δ in the specified uncertainty set [25].

Here we will present sufficient and necessary conditions for robust stability of sampled-data systems for two different uncertainty sets in terms of a related H_∞-type performance bound. The first class consists of bounded, linear h-periodic causal operators. For this uncertainty set sufficient and necessary conditions were obtained in [25]. The second class to be considered is the class of bounded, nonlinear time-invariant operators. The sufficient and necessary robust stability conditions presented here for this uncertainty set is new.

Introduce the notation T_{zw} for the closed-loop operator $L_2 \to L_2$ taking w to z in Figure 2,

$$T_{zw} := G_{11} + G_{12}\mathcal{H}\mathcal{K}(I - \mathcal{S}G_{22}\mathcal{H}\mathcal{K})^{-1}\mathcal{S}G_{21}. \tag{119}$$

Theorem 8. *Consider the sampled-data system in Figure 2. The following statements are equivalent:*

(a) The closed-loop system in Figure 2 is robustly stable for all linear, h-periodic and causal $\Delta : L_2 \to L_2$ such that $\|\Delta\| < 1$,

(b) the closed-loop system in Figure 2 is robustly stable for all causal, time-invariant $\Delta : L_2 \to L_2$ such that $\|\Delta\| < 1$,

(c) the nominal closed-loop system ($\Delta = 0$) in Figure 2 is internally asymptotically stable and $\|T_{zw}\| \leq 1$.

Proof: $(a) \Leftrightarrow (c)$: This result proved in [25].
$(c) \Rightarrow (b)$: The result follows in a straightforward way from the small gain theorem, cf. [25]. The signals e_1, e_2, e_3 and e_4 in Figure 2 are given by

$$\begin{aligned} e_1 &= \mathcal{K}e_2 + u_1 \\ e_2 &= \mathcal{S}G_{22}\mathcal{H}e_1 + \mathcal{S}G_{21}e_3 + u_2 \\ e_3 &= \Delta e_4 + u_3 \\ e_4 &= G_{12}\mathcal{H}e_1 + G_{11}e_3 + u_4. \end{aligned} \tag{120}$$

Solving for e_1–e_4 in terms of u_1–u_4 gives

$$\begin{aligned} e_1 = {} & M_1\big(I + \mathcal{K}\mathcal{S}G_{21}(I - \Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1\big)u_1 \\ & + M_1\mathcal{K}\big(I + \mathcal{S}G_{21}(I - \Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1\mathcal{K}\big)u_2 \end{aligned}$$

$$
\begin{aligned}
&+ M_1\mathcal{K}\mathcal{S}G_{21}(I-\Delta T_{zw})^{-1}u_3 + M_1\mathcal{K}\mathcal{S}G_{21}(I-\Delta T_{zw})^{-1}\Delta u_4\\
e_2 &= M_2\mathcal{S}\big(G_{22}\mathcal{H} + G_{21}(I-\Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1\big)u_1\\
&\quad + M_2\big(I + \mathcal{S}G_{21}(I-\Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1\mathcal{K}\big)u_2 \qquad (121)\\
&\quad + M_2\mathcal{S}G_{21}(I-\Delta T_{zw})^{-1}u_3 + M_2\mathcal{S}G_{21}(I-\Delta T_{zw})^{-1}\Delta u_4\\
e_3 &= (I-\Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1u_1 + (I-\Delta T_{zw})^{-1}\Delta G_{12}\mathcal{H}M_1\mathcal{K}u_2\\
&\quad + (I-\Delta T_{zw})^{-1}u_3 + (I-\Delta T_{zw})^{-1}\Delta u_4\\
e_4 &= (I-T_{zw}\Delta)^{-1}G_{12}\mathcal{H}M_1u_1 + (I-T_{zw}\Delta)^{-1}G_{12}\mathcal{H}M_1\mathcal{K}u_2\\
&\quad + (I-T_{zw}\Delta)^{-1}T_{zw}u_3 + (I-T_{zw}\Delta)^{-1}u_4.
\end{aligned}
$$

where $M_1 := (I-\mathcal{K}\mathcal{S}G_{22}\mathcal{H})^{-1}$ and $M_2 := (I-\mathcal{S}G_{22}\mathcal{H}\mathcal{K})^{-1}$. By internal stability of the nominal system, the operators M_1 and M_2 exist and are bounded. Likewise, internal stability of the nominal system implies that the operators $M_1\mathcal{K}$, $M_1\mathcal{K}\mathcal{S}G_{21}$, $M_2\mathcal{S}G_{22}\mathcal{H}$, $M_2\mathcal{S}G_{21}$, $G_{12}\mathcal{H}M_1$, and $G_{12}\mathcal{H}M_1\mathcal{K}$ in (121) are bounded. Moreover, $\|T_{zw}\| \le 1$ and $\|\Delta\| < 1$ imply that $(I-\Delta T_{zw})^{-1}$ exists and is bounded. Hence the robust stability result (b) follows from (c).

$(b) \Rightarrow (c)$: Internal stability of the nominal system is obviously a necessary condition for the robust stability property (b). In order to prove that the norm bound in (c) is also a necessary condition, it will be assumed that $\|T_{zw}\| > 1$, and a causal nonlinear time-invariant Δ with $\|\Delta\| < 1$ will then be constructed such that the system in Figure 2 is unstable.

Our construction will be somewhat similar to one used by Poolla and Ting [46] in a study of robust stabilization with nonlinear time-varying controllers. Introduce the truncation operator Π_t, defined as the orthogonal projection from $L_2[0,\infty)$ onto $L_2[0,t]$. Let $L_{2e}[0,\infty)$ denote the extension of $L_2[0,\infty)$ [47], such that for $w \in L_{2e}[0,\infty)$, $\Pi_t w \in L_2[0,t]$ for all $t > 0$. Assuming that $\|T_{zw}\| > 1$, there exist $w \in L_{2e}[0,\infty)$ and $\epsilon > 0$ such that

$$\lim_{t\to\infty} \|\Pi_t w\| = \infty, \qquad (122)$$

and

$$\lim_{t\to\infty} \frac{\|\Pi_t T_{zw} w\|}{\|\Pi_t w\|} > 1+\epsilon. \qquad (123)$$

In order to show this, one can use the fact that a similar result has been shown to hold in the l_2-norm for discrete time-varying systems [46]. The continuous-time result (122), (123) then follows by approximating the L_2-signals by sequences of step functions and the operator T_{zw} by a corresponding discrete operator, and using the fact that the space of piecewise constant functions is dense in L_2. By Eq. (123), there exists $t_0 > 0$ such that

$$\frac{\|\Pi_t T_{zw} w\|}{\|\Pi_t w\|} > 1+\epsilon, \text{ all } t > t_0. \qquad (124)$$

Define the signal

$$\tilde{w} := \begin{cases} 0, & t \le t_0 \\ w, & t > t_0 \end{cases} \tag{125}$$

and let $y := T_{zw}w$. For any $v \in L_{2e}[0,\infty)$, introduce

$$p(v) := \inf\{t : v(t) \neq 0\}, \tag{126}$$

and

$$s(v) := p(v) - p(y), \tag{127}$$

and define

$$t(v) := \inf\{t \ge p(v) : \|v(t)\| < \|y(t - s(v))\|\}. \tag{128}$$

Now define the operator Δ according to

$$\Delta(v) := \Pi_{t(v)}(\mathcal{T}_{s(v)}\tilde{w}) \tag{129}$$

where $\mathcal{T}_s$ is the translation operator; $(\mathcal{T}_s\tilde{w})(t) := \tilde{w}(t-s)$. By construction, the operator Δ is nonlinear, strictly causal and time-invariant, and $\|\Delta\| \le 1/(1+\epsilon) < 1$. Moreover, we have $(I - \Delta T_{zw})w = w - \Delta y = w - \tilde{w} = \Pi_{t_0}w$. Hence $(I - \Delta T_{zw})^{-1}\Pi_{t_0}w = w$, and since $\|\Pi_{t_0}w\| < \infty$, while w by assumption has unbounded norm in $L_2[0,\infty)$, cf. Eq. (122), it follows that the operator $(I - \Delta T_{zw})^{-1}$ is unbounded. Hence, in view of Eq. (121), the system in Figure 2 does not satisfy the robust stability property (b). □

Remark. Note that by the small gain theorem, the norm condition in Theorem 8(c) is sufficient conditions for robust stability for general norm-bounded uncertainties with $\|\Delta\| < 1$. This has been used to obtain a sufficient condition for robust stability for various uncertainty classes, cf. for example [24, 26, 48]. Theorem 8 shows that the norm condition (c) is also a necessary condition in the classes of linear h-periodic and nonlinear time-invariant uncertainties, respectively. It appears hard to make the result of the theorem more strict. For linear time-invariant uncertainties, the condition (c) is in general conservative. For this case, sufficient and necessary conditions for robust stability have been derived in [43, 44, 45].

IX. CONCLUSIONS

For sampled-data systems, a natural generalization of the standard H_∞ control problem consists of the minimization of the L_2-induced norm of the closed loop. This problem is related to the robustness and worst-case performance of sampled-data control systems. The sampled-data H_∞ control problem has been completely solved for a range of situations, including various types of sampling and hold elements, different sampling rates, etc.

In this contribution an overwiew of various approaches to solve the sampled-data control has been given. We have also shown (Section V) that various discretization methods which preserve the induced norm can be treated in a unified way, thus indicating that there are connections between different solution methods. Some special topics have also been discussed, including dual-rate control and optimal sampling prefilter design. For the sake of clarity, it has been assumed throughout that the hold element is a fixed zero-order hold function. It is, however, straightforward to generalize the procedures to generalized hold functions [20, 21, 26], or the case when the hold function is not given *a priori*, but is part of the design problem [18, 20, 21, 32]. Another possible generalization of the approaches discussed here is the multirate control of sampled-data systems. The multirate sampled-data control problems can be solved via the lifting technique approach described in Section III [38, 39]. It would also be interesting to apply the approach described in Section VI to more general multirate control problems.

APPENDIX A

Proof of Theorem 2: [*Sufficiency:*] Assume that the Riccati equation (26) has a bounded solution on $[0, Nh)$. Let $\hat{u}_k = \hat{u}_k^o, k = 0, 1, \ldots, N-1$. Then $w - w^o$ is given by

$$\dot{x}(t) = Ax(t) - B_2 S_{22}^{-1}(kh) S_{12}'(kh) x(kh) + B_1 w(t), \quad x(0) = 0$$

$$(w - w^o)(t) = -\gamma^{-2} B_1' [S_{11}(t)x(t) - S_{12}(t) S_{22}^{-1}(kh) S_{12}'(kh) x(kh)] + w(t),$$

$$t \in [kh, kh+1) \ , \ k = 0, 1, \ldots, N-1.$$

It is easy to see that there exists $c > 0$ such that $\|w - w^o\| \geq c\|w\|$ [23]. Hence we have from Lemma 2,

$$\|z\|_{L_2[0,Nh]}^2 - \gamma^2 \|w\|_{L_2[0,Nh]}^2 = -\gamma^2 \|w - w^o\|_{L_2[0,Nh]}^2 \leq -\gamma^2 c^2 \|w\|_{L_2[0,Nh]}^2$$

$$< 0 \ , \text{ if } w \neq 0.$$

The state feedback law (27*a*) thus achieves the performance bound (28). [*Necessity:*] In order to prove that the Riccati equation (26) also provides a necessary condition for the performance bound (28) to be achieved by state feedback, assume that there exists a state feedback law $\hat{u} = \mathcal{K}(\{x(kh)\})$ such that the inequality (28) holds. It will be shown that the Riccati equation with jumps (26) has a bounded solution on $[0, Nh)$. Consider on the sampling interval $[kh - h, kh)$ the system

$$\begin{bmatrix} \dot{p}_1(t) \\ \dot{p}_2(t) \end{bmatrix} = \begin{bmatrix} A_e & \gamma^{-2} B_{e1} B_{e1}' \\ -C_{e1}' C_{e1} & -A_e' \end{bmatrix} \begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix}, \quad p_2(kh^-) = S(kh^-) p_1(kh^-). \tag{A.1}$$

Introduce the state transition matrix associated with (A.1),

$$\begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix} = \begin{bmatrix} \Phi_{11}(t, kh^-) \\ \Phi_{21}(t, kh^-) \end{bmatrix} p_1(kh^-) \ , \ \begin{bmatrix} \Phi_{11}(kh^-, kh^-) \\ \Phi_{21}(kh^-, kh^-) \end{bmatrix} = \begin{bmatrix} I \\ S(kh^-) \end{bmatrix}.$$

If the transition matrix $\Phi_{11}(t, kh^-)$ is nonsingular on $[kh - h, kh)$, then the matrix $S(t) := \Phi_{21}(t, kh^-)\Phi_{11}^{-1}(t, kh^-)$ is a solution to the Riccati differential equation (26) on $[kh - h, kh)$, and $p_2(t) = S(t)p_1(t)$. To prove the required result, it is thus sufficient to show that the transition matrix $\Phi_{11}(t, kh^-)$ is nonsingular on $[kh - h, kh)$ for $k = N, N-1, \ldots, 1$.

Suppose on the contrary, that for some $k^* \in \{1, \ldots, N\}$, $t^* \in [k^*h - h, k^*h)$ is the largest time on $[0, Nh)$ such that $\Phi_{11}(t^*, k^*h^-)$ is singular. Then there is a nontrivial $p_1(k^*h^-)$ such that

$$p_1(t^*) = \Phi_{11}(t^*, k^*h^-)p_1(k^*h^-) = 0.$$

Then $p_2(t) \not\equiv 0$ on $[t^*, k^*h)$, since if $p_2(t^*) = 0$, it would follow from (A.1) that $p_1(t) = p_2(t) \equiv 0$ on $[t^*, k^*h)$, which contradicts our assumption on p_1. Define the signal

$$\bar{w}(t) := \begin{cases} 0 & , \ 0 \le t < t^* \\ \gamma^{-2} B_{e1}' p_2(t) & , \ t^* \le t < k^*h \\ w^o(t) & , \ k^*h \le t \le Nh. \end{cases} \tag{A.2}$$

The signal $\bar{w}$ is not identically zero on $[t^*, k^*h)$, since if $B_{e1}'p_2(t) \equiv 0$ on $[t^*, k^*h)$, then by (A.1), $\dot{p}_1 = A_e p_1$, and since $p_1(t^*) = 0$ by assumption, it would follow that $p_1(t) = 0$ on $[t^*, k^*h)$, which again leads to a contradiction. Hence $\bar{w}(t) \not\equiv 0$ on $[t^*, k^*h)$.

Now suppose $\hat{u}$ is the control signal generated by the feedback law $\hat{u} = \mathcal{K}(\{x(kh)\})$, when $w = \bar{w}$. Then $x(t) = 0$ for $t \in [0, t^*]$, and $\hat{u}_k = 0$ for $k < k^*$, because $x(0) = 0$ and $\bar{w}(t) = 0$, $t < t^*$. With $\bar{w}(t)$ given by (A.2), p_1 and $[x', u']'$ in (25) satisfy the same differential equation on $[t^*, k^*h)$. Since $p_1(t^*) = [x'(t^*), u'(t^*)]'$, it follows that $p_1(t) = [x'(t), u'(t)]'$ on $[t^*, k^*h)$. Direct calculation gives

$$\frac{d}{dt}\left([x', u']'p_2\right) = -z'z + \gamma^2 \bar{w}'\bar{w}.$$

Integrating from t^* to k^*h, and using the relation $p_2(t) = S(t)p_1(t)$ and the boundary value for $S(k^*h^-)$ in (26) gives

$$\begin{aligned} -\int_{t^*}^{k^*h} \left[z'z - \gamma^2 \bar{w}'\bar{w}\right] dt &= \int_{t^*}^{k^*h} \frac{d}{dt}\left([x', u']'p_2\right) dt \\ &= x'(k^*h) S_{11}(k^*h^-) x(k^*h) \end{aligned}$$

and we obtain

$$\begin{aligned}\|z\|^2_{L_2[0,Nh]} - \gamma^2\|\bar{w}\|^2_{L_2[0,Nh]} &= \int_{t^*}^{k^*h} \left[z'z - \gamma^2\bar{w}'\bar{w}\right]dt \\ &\quad + \int_{k^*h}^{Nh} \left[z'z - \gamma^2\bar{w}'\bar{w}\right]dt \\ &= \int_{t^*}^{k^*h} \left[z'z - \gamma^2\bar{w}'\bar{w}\right]dt + \sum_{k=k^*}^{N-1} \left(\hat{u}_k - \hat{u}_k^o\right)' S_{22}(kh)\left(\hat{u}_k - \hat{u}_k^o\right) \\ &\quad + x'(k^*h)S_{11}(k^*h^-)x(k^*h) \\ &= \sum_{k=k^*}^{N-1} \left(\hat{u}_k - \hat{u}_k^o\right)' S_{22}(kh)\left(\hat{u}_k - \hat{u}_k^o\right) \;\geq\; 0.\end{aligned}$$

Here we have used the identity in Lemma 2 to express the integral in the interval $[k^*h, Nh]$. Since $\bar{w} \neq 0$, this contradicts the assumption that there exists a feedback law which achieves the inequality (28). □

APPENDIX B

Proof of Lemma 4: The result can be proved by taking the estimation problem into an equivalent standard discrete H_∞ estimation problem. Let $\bar{\Phi}(\cdot,\cdot)$ denote the state transition matrix associated with the system matrix $\bar{A}(\cdot)$. Then we have from Eq. (36),

$$x(kh+h) = \bar{F}_k x(kh) + \int_{kh}^{kh+h} \bar{\Phi}(kh+h,\tau)\bar{B}_1(\tau)v(\tau)d\tau \tag{B.1}$$

where $F_k := \bar{\Phi}(kh+h, kh)$. Introduce the finite-dimensional discrete system

$$\begin{aligned} x(kh+h) &= \bar{F}_k x(kh) + \hat{B}_k\hat{w}_k,\;\; x(0) = 0 \\ \hat{r}_k &= \bar{C}_{1,k}x(kh) \\ \hat{y}_k &= \hat{C}_2 x(kh) + \hat{D}_{21}\hat{v}_k,\;\; k = 0,1,\ldots,N \end{aligned} \tag{B.2}$$

where $\hat{B}_k$ is defined according to

$$\hat{B}_k\hat{B}_k' = \int_{kh}^{kh+h} \bar{\Phi}(kh+h,\lambda)\bar{B}_1(\lambda)\bar{B}_1'(\lambda)\bar{\Phi}'(kh+h,\lambda)d\lambda. \tag{B.3}$$

By Lemma C.1(b), the discrete filter $\hat{r}_e = \mathcal{F}\hat{y}$ achieves the performance bound (37) for the system (36) if and only if it achieves the performance bound

$$\|\hat{r} - \hat{r}_e\|^2_{l_2} < \gamma^2\left[\|\hat{w}\|^2_{l_2} + \|\hat{v}\|^2_{l_2}\right],\;\text{all}\;(\hat{w},\hat{v}) \neq (0,0), \tag{B.4}$$

for the discrete system (B.2). The problem of finding a discrete filter which achieves the bound (B.4) is a standard discrete H_∞ filtering problem, the solution of which is known in the literature. By the results on discrete H_∞ estimation described in the literature [33, 49], there exists a discrete filter $\hat{r}_e = \mathcal{F}\hat{y}$ which achieves the performance bound (B.4) for the system (B.2) if and only if there exists a solution to the discrete Riccati equation

$$\begin{aligned} N_{k+1} &= \bar{F}_k N_k \Sigma_k^{-1} \bar{F}_k' + \hat{B}_k \hat{B}_k', \\ \Sigma_k &= I + [\hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2 - \gamma^{-2}\bar{C}_{1,k}'\bar{C}_{1,k}]N_k, \\ N_0 &= 0, \ k = 0, 1, \ldots, N, \end{aligned} \tag{B.5}$$

such that

$$\gamma^2 I - \bar{C}_{1,k}[I + N_k\hat{C}_2'(\hat{D}_{21}\hat{D}_{21}')^{-1}\hat{C}_2]^{-1}N_k\bar{C}_{1,k}' > 0 \ , \ k = 0, 1, \ldots, N,$$

or equivalently, that the matrices Σ_k, $k = 0, 1, \ldots, N$, have only positive eigenvalues. When the Riccati equation (B.5) has a solution, a filter which achieves the performance bound (B.4) is given by

$$\begin{aligned} \hat{x}(kh + h^-) &= \bar{F}_k\hat{x}(kh), \ \hat{x}(0) = 0 \\ \hat{x}(kh) &= x(kh^-) \\ &\quad + N_k\hat{C}_2'[\hat{D}_{21}\hat{D}_{21}' + \hat{C}_2N_k\hat{C}_2']^{-1}(\hat{y}_k - \hat{C}_2\hat{x}(kh^-)) \\ \hat{r}_{e,k} &= \bar{C}_{1,k}\hat{x}(kh) \ , \ k = 0, 1, \ldots, N. \end{aligned} \tag{B.6}$$

The proof is completed by observing that (B.5), (B.6) can be written in the form (38), (39), with $N(kh^-) := N_k$. □

APPENDIX C

The following lemma is used in Theorems 1, 4 and 5. The result has been used in a number of papers on sampled-data control, and various proofs it can be found in [15, 22, 23, 26, 50].

Here we consider linear operators $\Psi: R^n \to L_2[0, h]$ and $\Gamma: L_2[0, h] \to R^n$. Let $\mathcal{R}(\Psi) := \Psi R^n$ denote the range space of Ψ. Then the output space $L_2[0, h]$ of Ψ can be decomposed as $L_2[0, h] = \mathcal{R}(\Psi) \oplus \mathcal{R}(\Psi)^\perp$. With respect to this decomposition, Ψ has the representation

$$\Psi = \begin{bmatrix} \bar{\Psi} \\ 0 \end{bmatrix} \ : \ R^n \to \mathcal{R}(\Psi) \oplus \mathcal{R}(\Psi)^\perp. \tag{C.1}$$

Similarly, let $\mathcal{N}(\Gamma) := \{w \in L_2[0, h] : \Gamma w = 0\}$ denote the null space of Γ. Then the initial space $L_2[0, h]$ of Γ can be decomposed as $L_2[0, h] =$

$\mathcal{N}(\Gamma)^{\perp} \oplus \mathcal{N}(\Gamma)$. With respect to this decomposition, Γ has the representation

$$\Gamma = [\bar{\Gamma} \quad 0] \; : \; \mathcal{N}(\Gamma)^{\perp} \oplus \mathcal{N}(\Gamma) \to R^n. \tag{C.2}$$

Since the spaces $\mathcal{R}(\Psi)$ and $\mathcal{N}(\Gamma)^{\perp}$ are finite-dimensional, the operators $\bar{\Psi}$ and $\bar{\Gamma}$ in the decompositions (C.1) and (C.2) have finite-dimensional matrix representations.

Lemma C.1. *(a) Consider the linear operator $\Psi: R^n \to L_2[0,h]$. The linear operator $\bar{\Psi}$ in the decomposition (C.1) has a finite-dimensional matrix representation M defined according to*

$$M'M = \Psi^*\Psi \tag{C.3}$$

where Ψ^* denotes the adjoint operator.

Moreover, if Ψ is given by

$$(\Psi x)(\tau) = C(\tau)\Phi(\tau,0)x \;, \quad \tau \in [0,h], \tag{C.4}$$

where $\Phi(\cdot,\cdot)$ is a state transition matrix, and the matrix $C(\tau)$ is a piecewise continuous bounded function of τ, then

$$M'M = \int_0^h \Phi'(\lambda,0)C'(\lambda)C(\lambda)\Phi(\lambda,0)d\lambda. \tag{C.5}$$

(b) Consider the linear operator $\Gamma: L_2[0,h] \to R^n$. The linear operator $\bar{\Gamma}$ in the decomposition (C.2) has a finite-dimensional matrix representation N defined according to

$$NN' = \Gamma\Gamma^*. \tag{C.6}$$

Moreover, if Γ is given by

$$\Gamma w = \int_0^h \Phi(h,\tau)B(\tau)w(\tau)d\tau, \tag{C.7}$$

where $\Phi(\cdot,\cdot)$ is a state transition matrix, and the matrix $B(\tau)$ is a piecewise continuous bounded function of τ, then

$$NN' = \int_0^h \Phi(h,\lambda)B(\lambda)B'(\lambda)\Phi'(h,\lambda)d\lambda. \tag{C.8}$$

Proof: (a) The operator $\Psi^*\Psi: R^n \to R^n$ has an $n \times n$ matrix representation, and hence there exists a matrix M which satisfies (C.3). The matrix M is

a matrix representation of the operator $\bar{\Psi}$ if and only if for any x_1, x_2 in R^n,

$$< \Psi x_1, \Psi x_2 > = x_1' M' M x_2.$$

But this follows from the definition of M, since

$$< \Psi x_1, \Psi x_2 > = < x_1, \Psi^* \Psi x_2 > = x_1' M' M x_2.$$

The second claim follows from the fact that the adjoint is given by

$$\Psi^* w = \int_0^h \Phi'(\tau, 0) C'(\tau) w(\tau) d\tau.$$

(*b*) The linear operator Γ has a representation (C.2) if and only if the adjoint operator $\Gamma^* : R^n \to L_2[0, h]$ has a representation

$$\Gamma^* = \begin{bmatrix} \bar{\Gamma}^* \\ 0 \end{bmatrix} : R^n \to \mathcal{N}(\Gamma)^\perp \oplus \mathcal{N}(\Gamma).$$

The result then follows by applying (*a*) to the adjoint Γ^*. □

ACKNOWLEDGMENT

This work was in part supported by the Academy of Finland.

REFERENCES

1. G. Zames, "Feedback and Optimal Sensitivity: Model Reference Transformations, Multiplicative Seminorms, and Approximate Inverses," *IEEE Trans. Automat. Control* **26**, 301–320 (1981).
2. B. Francis, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.
3. J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-Space Solutions to Standard H_2 and H_∞ Control Problems," *IEEE Trans. Automat. Control* **34**, 831–847 (1989).
4. P. A. Iglesias and K. Glover, "State-Space Approach to Discrete-Time $\mathcal{H}_\infty$ Control," *Int. J. Control* **54**, 1031–1073 (1991).
5. R. Y. Chiang and M. G. Safonov, *Robust Control Toolbox*, The MathWorks, Inc., 1988.
6. K. J. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design*, Prentice Hall, 1984.
7. B. Francis and T. T. Georgiou, "Stability Theory for Linear Time-Invariant Plants with Periodic Digital Controllers," *IEEE Trans. Automat. Control* **33**, 820–832 (1988).

8. T. Chen and B. Francis, "Input-Output Stability of Sampled-Data Systems," *IEEE Trans. Automat. Control* **36**, 50–58 (1991).
9. T. Chen and B. Francis, "On the L_2 Induced Norm of a Sampled-Data System," *Systems and Control Letters* **15**, 211–219 (1990).
10. P. T. Kabamba and S. Hara, "On Computing the Induced Norm of Sampled Data Systems," *Proceedings of the American Control Conference*, San Diego, CA, pp. 319–320 (1990).
11. N. Sivashankar and P. P. Khargonekar, "Worst Case Performance Analysis of Linear Systems with Jumps with Applications to Sampled-Data Systems," *Proc. of the American Control Conference*, Chicago, IL, pp. 692–696 (1992).
12. N. Sivashankar and P. P. Khargonekar, "Induced Norms for Sampled-Data Systems," *Automatica* **28**, 1267–1272 (1992).
13. G. M. H. Leung, T. P. Perry, and B. A. Francis, "Performance Analysis of Sampled-Data Control Systems," *Automatica* **27**, 699–704 (1991).
14. B. Bamieh, J. B. Pearson, B. A. Francis and A. Tannenbaum, "A Lifting Technique for Linear Periodic Systems with Applications to Sampled-Data Control," *Systems & Control Letters* **17**, 79–88 (1991).
15. B. Bamieh and J. B. Pearson, "A General Framework for Linear Periodic Systems with Application to $\mathcal{H}^\infty$ Sampled-Data Control," *IEEE Trans. Automatic Control* **37**, 418–435 (1992).
16. P. T. Kabamba and S. Hara, "Worst-Case Analysis and Design of Sampled-Data Control Systems," *Proceedings of the 29th IEEE Conference on Decision and Control*, Honolulu, Hawaii, pp. 202–203 (1990).
17. P. T. Kabamba and S. Hara, "Worst-Case Analysis and Design of Sampled-Data Control Systems," *IEEE Trans. Autom. Control* **38**, 1337–1357 (1993).
18. W. Sun, K. M. Nagpal, and P. P. Khargonekar, "$\mathcal{H}_\infty$ Control and Filtering for Sampled-Data Systems," *IEEE Trans. Automat. Control* **38**, 1162–1175 (1993). (Also in *Proceedings of the American Control Conference*, Boston, MA, 1652–1657 (1991).)
19. W. Sun, K. Nagpal, P. P. Khargonekar, and K. R. Poolla, "Digital Control Systems: $\mathcal{H}_\infty$ Controller Design with a Zero-Order Hold Function," *Proceedings of the 31st IEEE Conference on Decision and Control*, Tucson, Arizona, pp. 475–480 (1992).
20. G. Tadmor, "Optimal H_∞ Sampled-Data Control in Continuous-Time Systems," *Proceedings of the American Control Conference*, Boston, MA, pp. 1658–1663 (1991).
21. G. Tadmor, "H_∞ Optimal Sampled-Data Control in Continuous Time Systems," *Int. J. Control* **56**, 99–141 (1992).
22. H. T. Toivonen, "Sampled-Data Control of Continuous-Time Systems with an H_∞ Optimality Criterion," *Automatica* **28**, 45–54 (1992).

23. H. T. Toivonen, "Sampled-Data H_∞ Optimal Control of Time-Varying Systems," *Automatica* **28**, 823–826 (1992).
24. T. Chen and B. A. Francis, "Sampled-Data Optimal Design and Robust Stabilization," *Proceedings of the American Control Conference*, Boston, MA, pp. 2704–2709 (1991).
25. N. Sivashankar and P. P. Khargonekar, "Robust Stability and Performance Analysis of Sampled-Data Systems," *IEEE Trans. Automat. Control* **38**, 58–69 (1993). (Also in *Proceedings of the 30th IEEE Conference on Decision and Control*, Brighton, England, pp. 881–886 (1991).)
26. Y. Hayakawa, Y. Yamamoto, and S. Hara, "H_∞ Type Problem for Sampled-Data Control Systems – A Solution via Minimum Energy Characterization," *Proceedings of the 31st IEEE Conference on Decision and Control*, Tucson, Arizona, pp. 463–468 (1992).
27. B. Bamieh, M. A. Dahleh, and J. B. Pearson, "Minimization of the $\mathcal{L}^\infty$-Induced Norm for Sampled-Data Systems," *IEEE Trans. Automatic Control* **AC-38**, 717–732 (1992).
28. G. Dullerud and B. A. Francis, "$\mathcal{L}_\infty$ Analysis and Design of Sampled-Data Systems," *IEEE Trans. Automat. Control* **AC-37**, 436–446 (1992).
29. M. H. Khammash, "Necessary and Sufficient Conditions for the Robustness of Time-Varying Systems with Applications to Sampled-Data Systems," *IEEE Trans. Automat. Control* **AC-38**, 49–57 (1993).
30. Y. Yamamoto, "New Approach to Sampled-Data Control Systems – A Function Space Method," *Proceedings of the 29th IEEE Conference on Decision and Control*, pp. 1882–1887 (1990).
31. T. Başar, "Optimal H^∞ Designs Under Sampled State Measurements," *Systems & Control Letters* **16**, 399–410 (1991).
32. P. Bernhard, "Application of the Min–Max Certainty Equivalence Principle to the Sampled Data Output Feedback H^∞ Control Problem," *Systems & Control Letters* **16**, 229–234 (1991).
33. T. Başar and P. Bernhard, *H_∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhauser, 1991.
34. G. Tadmor, "Worst-Case Design in the Time Domain: The Maximum Principle and the Standard H_∞ Problem," *Math. Control Signals Systems* **3**, pp. 301–324 (1990).
35. H. T. Toivonen, "Worst-Case Sampling for Sampled-Data H_∞ Design," *Proceedings of the 32rd IEEE Conference on Decision and Control*, San Antonio, TX, pp. 337–342 (1993).
36. J. P. Keller and B. D. O. Anderson, "H_∞ Optimal Controller Discretization," *Int. J. Robust and Nonlinear Control* **1** 125–137 (1990).
37. J. P. Keller and B. D. O. Anderson, "A New Approach to the Discretization of Continuous-Time Controllers," *IEEE Trans. Automat.*

Control **37**, 214–223 (1992).

38. P. Voulgaris and B. Bamieh, "Optimal $\mathcal{H}^\infty$ and $\mathcal{H}^2$ Control of Hybrid Multirate Systems," *Systems & Control Letters* **20**, 249–261 (1993).
39. T. Chen and Li Qiu, "$\mathcal{H}_\infty$ Design of General Multirate Sampled-Data Control Systems," *Proceedings of the 32rd IEEE Conference on Decision and Control*, San Antonio, TX, pp. 315–320 (1993).
40. P. Voulgaris and B. Bamieh, "Control of Asynchronous Sampled-Data Systems," *Proceedings of the 32rd IEEE Conference on Decision and Control*, San Antonio, TX, pp. 785–786 (1993).
41. H. T. Toivonen, "Sampling Prefilters with an H_∞ Criterion," *Proceedings of the American Control Conference*, Chicago, IL, pp. 697–701 (1992).
42. W. Sun, K. Nagpal, and P. P. Khargonekar, "Optimal Sampler for $\mathcal{H}_\infty$ Control," *Proceedings of the 32rd IEEE Conference on Decision and Control*, San Antonio, TX, pp. 777–782 (1993).
43. Er-Wei Bai, "Uncertainty Bound of Sampled-Data Systems," *Systems & Control Letters* **19**, 151–156 (1992).
44. G. Dullerud and K. Glover, "Necessary and Sufficient Conditions for Robust Stability of SISO Sampled-Data Systems to LTI Perturbations," *Proceedings of the American Control Conference*, Chicago, IL, pp. 2644–2648 (1992).
45. G. Dullerud and K. Glover, "Robust Stabilization of Sampled-Data Systems to Structured LTI Pertubations," *IEEE Trans. Automat. Control* **AC-38**, 1497–1508 (1993).
46. K. Poolla and T. Ting, "Nonlinear Time-Varying Controllers for Robust Stabilization," *IEEE Trans. Automat. Control* **32**, 195–200 (1987).
47. J. C. Willems, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
48. H. Hu and C. V. Hollot, "Boundedness of the $\mathcal{L}_2$-induced Norm Implies Quadratic Stability for Uncertain Sampled-Data Control Systems," *Proceedings of the American Control Conference*, San Fransisco, CA, pp. 445–449 (1993).
49. I. Yaesh and U. Shaked, "A Transfer Function Approach to the Problems of Discrete-Time Systems: H_∞-Optimal Linear Control and Filtering," *IEEE Trans. Automat. Control* **36**, 1264–1271 (1991).
50. H. T. Toivonen, "Discretization of Analog Filters via H_∞ Model Matching Theory," *Int. J. Adaptive Control and Signal Processing* **6**, 499–514 (1992).

TECHNIQUES IN ON-LINE PERFORMANCE EVALUATION OF MULTILOOP DIGITAL CONTROL SYSTEMS AND THEIR APPLICATION

Carol D. Wieseman, Vivek Mukhopadhyay, and Sherwood Tiffany Hoadley
Langley Research Center
National Aeronautics and Space Administration
Hampton, Virginia 23681-0001

Anthony S. Pototzky
Lockheed Engineering and Sciences Company

I. INTRODUCTION

Active controls are becoming an increasingly important means to enhance the performance of aircraft. Because the process of designing multi-input/multi-output (MIMO) digital control laws uses relatively untested theoretical methods, it is crucial to evaluate the performance of designed control laws through experimentation. In this chapter performance of the control is measured in terms of stability and robustness. A stable system is one in which all the poles of the system are on the left-hand side of the complex plane and robustness means the tolerance of system stability to plant uncertainty, which can be measured in terms of minimum singular values or gain and phase margins. The results of the performance evaluations can then be used to help evaluate the design methods. For classical single-input/single-output (SISO) control systems, analysis tools such as Nyquist diagrams are often used to determine the stability and robustness of the closed-loop system. For MIMO systems, Nyquist techniques are inadequate. Consequently, analytical methods based on the use of singular values of return-difference matrices at various points in the control loop were developed (references [1] - [3]) to examine the stability and robustness of a control system (SISO or MIMO).

For examining the stability and robustness of digital control systems during testing, the plant is excited by a known input. Experimental time-history data consisting of the excitation and system responses (both plant and controller outputs) are acquired. These time history data are then transformed to the frequency domain using Fast Fourier Transform (FFT) methods so that transfer matrices and the return-difference matrices can be computed. Singular values are then determined to obtain measures of system stability and robustness. The steps from acquiring the data through interpreting the singular values comprise a methodology referred to herein as Controller Performance Evaluation (CPE). The methodology is generic in nature and can be used in many types of multi-loop digital controller applications including flight control systems, digitally controlled spacecraft structures, and actively controlled wind-tunnel models. These CPE methods were employed during recent actively controlled wind-tunnel testing to check the stability of the closed-loop system to reduce the risk of damage to the wind-tunnel model and the tunnel.

The present chapter describes the implementation of the CPE capability, structure of the data flow, signal processing methods used to process the data, and the software developed to generate transfer functions. A brief development of the equations used to obtain the open-loop plant, controller transfer matrices, and return-difference matrices are given. Results of applying the CPE methodology to provide on-line evaluation of digital flutter suppression systems tested on the Rockwell Active Flexible Wing (AFW) wind-tunnel model (references [4]-[6]), using the AFW digital controller described in reference [7], are presented to demonstrate the CPE capability.

II. NOTATION AND DEFINITIONS

det(•)	determinant
I	identity matrix

$\mathbf{X}_u$ controller output transfer matrix

$\mathbf{Y}_u$ plant output transfer matrix

$\bar{\sigma}$ maximum singular value

$\underline{\sigma}$ minimum singular value

ω frequency

Subscripts

A additive uncertainty

M_I multiplicative uncertainty at plant input point

M_O multiplicative uncertainty at plant output point

u excitation

x controller output

y plant output

Sub-subscripts

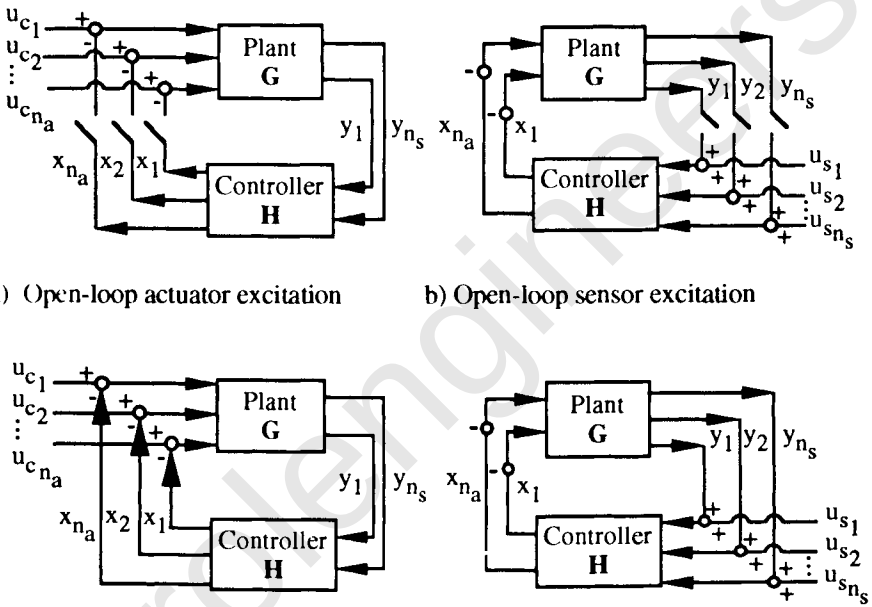
c signal added at the command location

s signal added at the sensor location

III. CONTROLLER PERFORMANCE EVALUATION

Block diagrams of the basic open- and closed-loop control problems with negative feedback are presented in Figure 1. The plant to be controlled is represented mathematically by a frequency domain transfer matrix, **G**, with n_s outputs and n_a inputs, where n_s is the number of control-law-input sensor measurements and n_a is the number of control-law-output actuator commands . The controller is represented mathematically with a transfer matrix, **H**, with n_s inputs, and n_a outputs.

The excitation is used to derive transfer functions between outputs and inputs in either the open- or closed-loop system. The open-loop system is one in which either the control law outputs (commands required for controlling plant response) are not fed back into the system as in figure 1(a), or the sensors are not fed back into the controller as in figure 1(b).



a) Open-loop actuator excitation

b) Open-loop sensor excitation

c) Closed-loop actuator excitation

d) Closed-loop sensor excitation

Figure 1. Controller- plant diagrams depicting the control problem with negative feedback.

Figures 1(a) and (c) depict the case when the external excitation, u, used to excite the system is added at the plant input point. Specifically, in 1(a), the i'th plant input, u_{c_i}, is equal to u and the others are zero. In the closed-loop system, 1(c), the i'th input, u_{c_i}-x_i, is equal to u-x_i, and the others are -x_i. When there are more sensor inputs than control outputs, then the excitation is added at the controller input point as depicted by figures 1(b) and (d). Specifically, the i'th

input to the controller, u_{si}, is equal to u and the others are zero in the open-loop case. In the closed-loop case, 1(d), the i'th controller input, $y_i + u_{si}$, is equal to $y_i + u$ and the others are y_j. In both cases 1(c) and (d), the negative of the controller outputs are input to the plant

When the number of sensors is equal to the number of control outputs, excitations can be applied at either location.

Controller performance evaluation is a two-mode, four-step process. The two modes are open- and closed-loop, and each mode consists of two steps. The process is outlined conceptually for the flutter suppression system application as follows:

A. OPEN-LOOP

Step 1: Verify the controller, **H**, by comparing with the designed control law transfer matrix.

Step 2: Predict closed-loop performance based on the open-loop performance to determine whether the control law will stabilize or destabilize the system when the loop is closed.

B. CLOSED-LOOP

Step 1: Determine the stability margins of the closed-loop system during the closed-loop testing by evaluating the singular values of return-difference matrices, **(I+GH)**, **(I+HG)**, and $\mathbf{H(I+GH)}^{-1}$.

Step 2: Determine open-loop plant stability during the closed-loop testing to determine the open-loop flutter boundary.

IV. CPE COMPUTATIONS AND PROCEDURES

The CPE computations involve generating frequency domain transfer functions of plant outputs, y, and control law feedback commands, x, due to an excitation, u. Fast Fourier Transform techniques are used to convert time-domain data to the frequency domain and transfer functions are calculated from the corresponding frequency-domain functions. The controller, **H**, and the return-difference matrices and their singular values are then calculated using matrix operations. The computations are described in the following paragraphs. Figure 2 is a flowchart which outlines the CPE procedures.

A. TRANSFER FUNCTION CALCULATIONS

The method used to compute transfer functions is described in reference [8]. The method therein was extended in the present study to include additional data-windowing capabilities and overlap averaging. Although most control designers used Hanning windows to help smooth data when evaluating their control laws during actual testing, windowing capabilities also include ramp-in/ramp-out, cosine taper, and cosine bell. The overlap-averaging capability allows long time histories to be partitioned into shorter time spans, taking advantage of long periods of time history data to average out noise thereby increasing the statistical quality of the data sample. A zero-fill capability is available to zero fill time history data to an exact block-size needed for FFT computations. The overlap-averaging capability with zero-fill provided optimum use of the experimental time history data which were obtained. The controller-output transfer matrix, $\mathbf{X_u}$, is the matrix of ratios of the cross-spectra, S_{ux}, of the controller outputs, x, due to the excitations, u, to the auto-spectra, S_{uu} of the excitations, u; all spectra are obtained from the FFT's of the time histories. Each i,j element of the transfer matrix is given by equation 1.

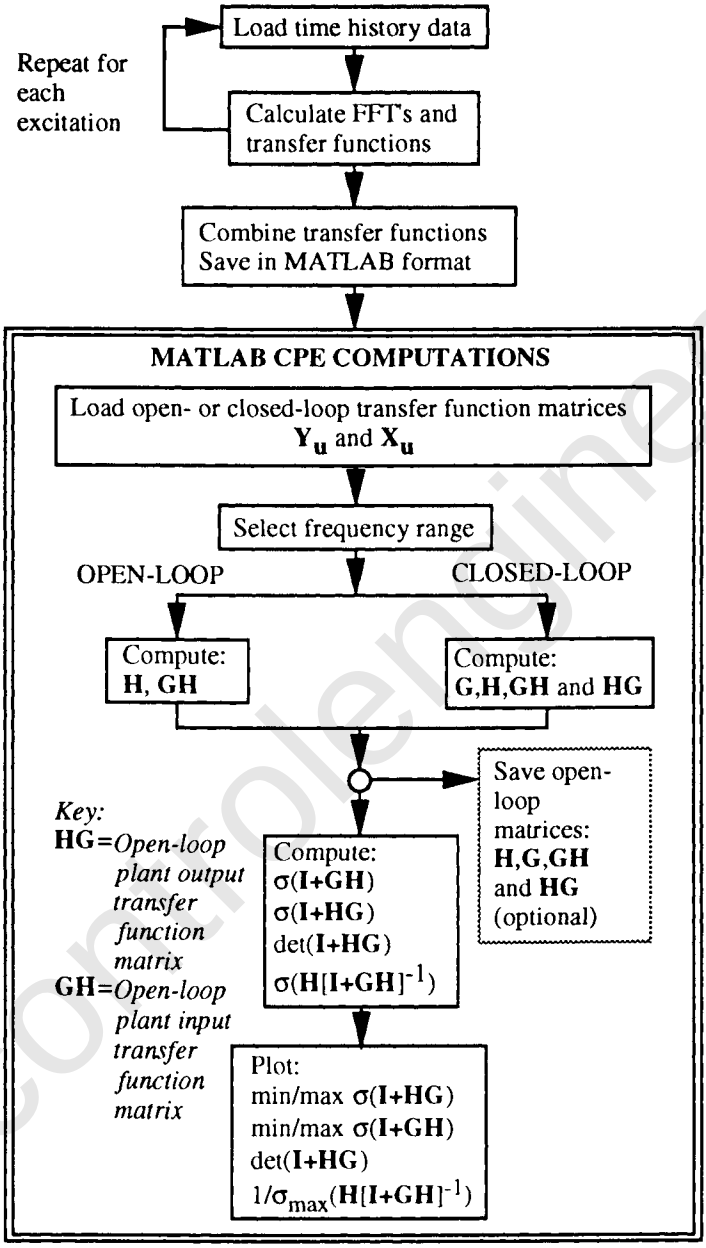


Figure 2. Flowchart of CPE procedures

$$[\mathbf{X_u}(\omega)]_{ij} = \frac{\sum_{m=1}^{N}\left(S_{u_i x_j}(\omega)\right)_m}{\sum_{m=1}^{N}\left(S_{u_i u_i}(\omega)\right)_m} \tag{1}$$

where N is the number of time history segments. Similarly, each element of the plant-output transfer matrix, $\mathbf{Y_u}$, is given by Eq. (2).

$$[\mathbf{Y_u}(\omega)]_{ij} = \frac{\sum_{m=1}^{N}\left(S_{u_i y_j}(\omega)\right)_m}{\sum_{m=1}^{N}\left(S_{u_i u_i}(\omega)\right)_m} \tag{2}$$

The two matrices, $\mathbf{X_u}$ and $\mathbf{Y_u}$, are the basis of all subsequent CPE computations. Note that if excitations are added at the command location, figures 1(a) and (c), the dimensions of $\mathbf{X_{uc}}$ and $\mathbf{Y_{uc}}$ are $n_a \times n_a$ and $n_s \times n_a$, respectively; whereas, if the excitations are added at the sensor location, figures 1(b) and (d), the dimensions of $\mathbf{X_{us}}$ and $\mathbf{Y_{us}}$ are $n_a \times n_s$ and $n_s \times n_s$.

B. CPE PROCEDURES

For both open- and closed-loop analysis, the return difference matrices are required. This involves computing **HG** and **GH** as well as the plant, **G**, and the controller, **H**. Table 1 summarizes the order in which these matrices are computed and the basic equations used in calculating them for both the open- and closed-loop cases in which excitation is added to either the command input to the plant or the sensor input to the controller as depicted in figures 1(a)-(d). To avoid rank-deficient matrices in Eqs. (5), excitations should be added at the command location, figures 1(a) and (c), corresponding to equations 5(a) and (c), if

$n_a > n_s$; and they should be added at the sensor location, figures 1(b) and (d), corresponding to equations 5(b) and (d), if $n_a < n_s$. Reference [9] provides a more detailed development of the equations for the case in which the excitations are added at the command location.

Table 1. Basic CPE Matrix Equations*

Open-Loop	
Command Excitation	Sensor Excitation
$\mathbf{G} = \mathbf{Y}_{uc}$ (3a)	$\mathbf{H} = \mathbf{X}_{us}$ (3b)
$\mathbf{HG} = \mathbf{X}_{uc}$ (4a)	$\mathbf{GH} = -\mathbf{Y}_{us}$ (4b)
$\mathbf{H}=\left([\mathbf{Y}_{uc}\,\mathbf{Y}_{uc}^T]^{-1}\,[\mathbf{Y}_{uc}\,\mathbf{X}_{uc}^T]\right)^T$ (5a)	$\mathbf{G}=-\left([\mathbf{X}_{us}\,\mathbf{X}_{us}^T]^{-1}\,[\mathbf{X}_{us}\,\mathbf{Y}_{us}^T]\right)^T$ (5b)
$\mathbf{GH} = \mathbf{G} \cdot \mathbf{H}$ (6a)	$\mathbf{HG} = \mathbf{H} \cdot \mathbf{G}$ (6b)
Closed-Loop	
Command Excitation	Sensor Excitation
$\mathbf{G} = \left([\mathbf{I} - \mathbf{X}_{uc}^T]^{-1}\mathbf{Y}_{uc}^T\right)^T$ (3c)	$\mathbf{H} = \left([\mathbf{I} - \mathbf{Y}_{us}^T]^{-1}\mathbf{X}_{us}^T\right)^T$ (3d)
$\mathbf{HG} = \left([\mathbf{I} - \mathbf{X}_{uc}^T]^{-1}\mathbf{X}_{uc}^T\right)^T$ (4c)	$\mathbf{GH} = -\left([\mathbf{I} - \mathbf{Y}_{us}^T]^{-1}\mathbf{Y}_{us}^T\right)^T$ (4d)
$\mathbf{H} = \left([\mathbf{Y}_{uc}\,\mathbf{Y}_{uc}^T]^{-1}\,[\mathbf{Y}_{uc}\,\mathbf{X}_{uc}^T]\right)^T$ (5c)	$\mathbf{G}= -\left([\mathbf{X}_{us}\,\mathbf{X}_{us}^T]^{-1}\,[\mathbf{X}_{us}\,\mathbf{Y}_{us}^T]\right)^T$ (5d)
$\mathbf{GH} = \mathbf{G} \cdot \mathbf{H}$ (6c)	$\mathbf{HG} = \mathbf{H} \cdot \mathbf{G}$ (6d)

* All matrices are functions of ω.

1. OPEN-LOOP

To perform the first step of the open-loop CPE, the controller transfer matrix, **H**, computed using either Eq. (5a) or Eq. (3b) is compared with the designed control law transfer matrix to verify the implementation of the

controller. Specifically, the transfer functions are compared for each output/input pair.

To perform the second step in the open-loop CPE (predicting closed-loop performance based on the open-loop performance to decide whether the control law will stabilize or destabilize the system when the loop is closed), it is convenient with a MIMO system to evaluate robustness with respect to plant uncertainties. Multiplicative uncertainties at the plant input and plant output points by examining the minimum singular values† of the return-difference matrices:

$$\sigma_{M_I}(\omega) = \sigma_{min}\,((\mathbf{I}+\mathbf{HG})(\omega)) \qquad \text{and}$$
$$\sigma_{M_O}(\omega) = \sigma_{min}\,((\mathbf{I}+\mathbf{GH})(\omega)) \tag{7}$$

and robustness with respect to an additive uncertainty by examining:

$$\sigma_A(\omega) = \left(\frac{1}{\sigma_{max}\left((\mathbf{H}[\mathbf{I}+\mathbf{GH}]^{-1})(\omega)\right)} \right). \tag{8}$$

The matrix product **HG** is obtained from either Eq. (4a) or (6b). The matrix product **GH** can be obtained using (6a) or (4b). The singular values of the return-difference matrices can then be determined and observations of the minimum and maximum values over the entire frequency range can be made.

A system crosses the stability boundary at frequencies when **I+HG** or **I+GH** is singular and the minimum singular value becomes zero. Therefore the proximity to zero indicates where the system is prone to go unstable and provides a quantitative measure of robustness. Reference [3] contains a derivation which relates guaranteed gain and phase margins to minimum singular

† The singular values, σ, of any matrix F are equal to $\sqrt{\lambda(F^*F)}$ where λ are the eigenvalues. The singular values, σ, are always non-negative real and F^* is the complex conjugate transpose of F.

values. This relationship is shown in Figure 3 of the present paper, which is a reproduction of figure 2 from reference [3], and will be referred to later when discussing results.

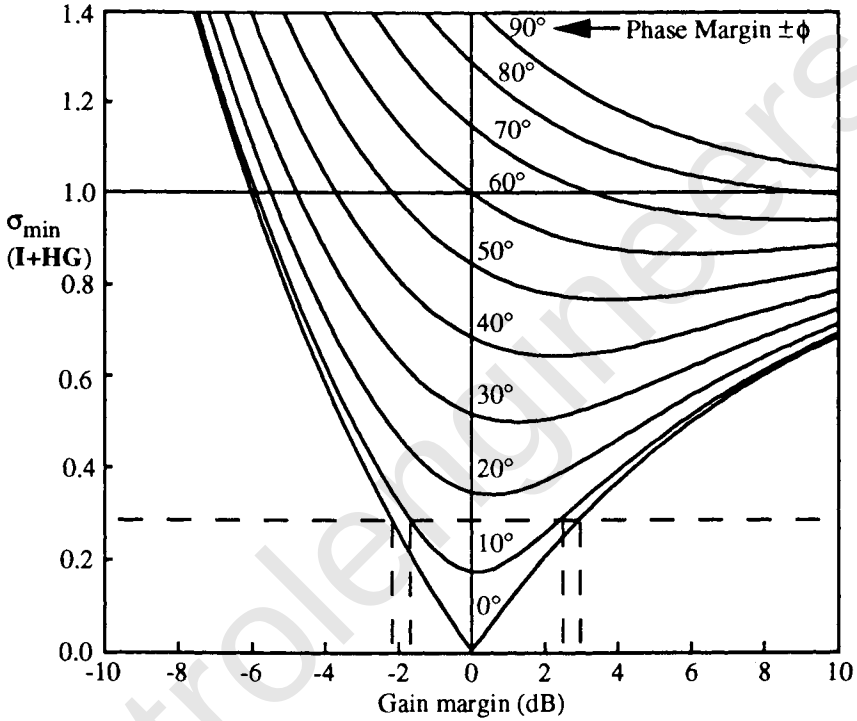


Figure 3. Universal gain and Phase Margin Diagram

The ratio of the maximum to minimum singular values of a return-difference matrix is the condition number. If a minimum singular value approaches zero (has low stability margins), the size of the condition number, especially when it is much larger than one, becomes an important indicator of the uncertainty in the measure of system stability; i.e. large condition numbers indicate that the predicted stability margin is very uncertain.

The determinant of the return difference matrix (**I HG**) can be used to determine open-loop system stability when the loop is closed. The locus of the determinant as a function of frequency has the property that if the open-loop system is stable, a clockwise encirclement of the critical point (0,0) indicates that the controller is destabilizing. Furthermore, the proximity of the determinant locus to the critical point is a direct indication of how near the control system is to an instability. For SISO controllers, this is analogous to the Nyquist plot since in that case

$$\det(\mathbf{I}+\mathbf{HG}) = \det(\mathbf{I}+\mathbf{GH}) = \det(\mathbf{I}) + \det(\mathbf{GH}) = 1 + \det(\mathbf{GH}). \quad (9)$$

Hence, for a SISO system only, the Nyquist plot is simply a translation of the plot of the determinant of the return-difference matrix about a different critical point, namely (-1,0).

2. CLOSED-LOOP

The main difference between the closed-loop and open-loop computations is that the measurements of x and y are obtained experimentally from the closed-loop system rather than the open-loop system. The transfer matrices, G and H, therefore be extracted from the closed-loop system. The matrix product **HG** is obtained using either Eq. (4c) or (6d), and the matrix product **GH** can be obtained using (6c) or (4d). Singular values of the return difference matrices can then be determined as for the open-loop case. The singular-value robustness plots are interpreted the same way for closed-loop testing as they were for open-loop testing. The det(I+HG) can be used to predict open-loop plant instability from a stable closed-loop system. Care must be taken however, in interpreting the determinant plots. If the closed-loop plant is stable and the open-loop plant is stable, there should be no net encirclement of the origin. On the contrary, if the closed-loop plant is stable and the open-loop plant is unstable with a pair of

complex unstable poles, then the determinant plot will show one net counter-clockwise encirclement of the origin. As with the open-loop case, the proximity of the determinant locus to the critical point and the proximity of the minimum singular values to zero are used as measures of closed-loop stability margins.

To obtain results for step 2 of the closed-loop mode, the minimum of the inverse maximum singular value (IMSV)

$$\min_{\omega} \left(\frac{1}{\sigma_{max}\,(\mathbf{G}(\omega))} \right) \tag{10}$$

of the open-loop plant transfer matrix, **G** defined by either Eq. (3c) or (5d), is an excellent indicator of poles in the proximity of the imaginary axis. The frequency of instability is determined by where the minimum of the IMSV of the plant approaches zero with increasing dynamic pressure. Tracing values of closest approach was a useful way of determining the open-loop plant flutter boundary with respect to some changing test condition, such as dynamic pressure.

V. PLANT ESTIMATION

To determine the plant in the case when there is no control law operating, the plant transfer matrix can be derived directly from the calculated transfer functions. In the case when there was a control law operating, the plant has to be extracted from the closed-loop system. In either case, the purpose of plant determination is two-fold. The first is to provide transfer function data to engineers for their use in redesigning control laws and the second purpose is to use the open-loop plant to evaluate open-loop plant stability. Some elements of the plant transfer matrix can be extracted during CPE calculations; however, an

additional capability is required to calculate the remaining elements of the plant transfer matrix.

Figure 4 shows a block diagram of the plant and controller. The "c" subscript refers to the control law elements. The "e" subscript refers to elements external to the control law tested. Table 3 outlines the equations needed in order to calculate all the elements of the plant transfer matrix:

$$G = \begin{bmatrix} G_{cc} & G_{ec} \\ G_{ce} & G_{ee} \end{bmatrix}$$

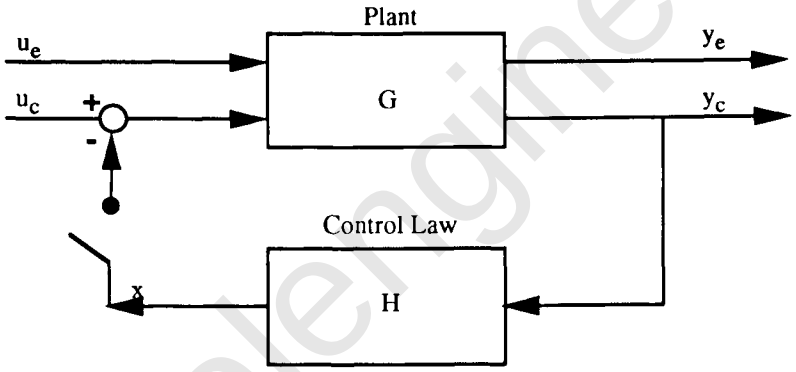


Figure 4. Controller-Plant diagrams depicting the control problem with negative feedback

Table 3. Basic Plant Equations*

Open-Loop	Closed-Loop
$G_{cc} = Y_{cc}$	$G_{cc} = \left([I - X_c^T]^{-1} Y_{cc}^T\right)^T$
$G_{ec} = Y_{ec}$	$G_{ec} = \left([I - X_c^T]^{-1} Y_{ec}^T\right)^T$
$G_{ce} = Y_{ce}$	$G_{ce} = Y_{ce} + G_{cc} X_e$
$G_{ee} = Y_{ee}$	$G_{ee} = Y_{ee} + G_{ec} X_e$

* All matrices are functions of ω.

In the table, $\mathbf{X}_c$ and $\mathbf{X}_e$ are the transfer functions of the control law outputs, x, with respect to u_c (excitations of control surfaces used by the control law) and u_e (those not used by the control law). $\mathbf{Y}_{cc}$ and $\mathbf{Y}_{ce}$ are the transfer functions of the plant outputs, y_c, used by the controller with respect to u_c and u_e, respectively. $\mathbf{Y}_{ec}$ and $\mathbf{Y}_{ee}$ are the transfer functions of the plant outputs, y_e, not used by the controller with respect to u_c and u_e, respectively.

VI. SUMMARY OF FLUTTER-SUPPRESSION TESTING

During flutter suppression testing, the control systems were operated in both open-loop and closed-loop modes. For the purpose of maintaining both model and tunnel safety, each candidate control law was initially tested open loop to insure that the control law itself would not destabilize the wind-tunnel model during closed-loop tests. The feedback was digitally switched open at the control law output point, figure 1(a) and the responses and excitation were collected at the control law input and the control law output locations. The appropriate transfer functions were generated from these responses and then the CPE capability was exercised to predict the closed-loop system stability while the loop was still open (figure 2). If the control law was predicted to be stable, the switch was closed and the closed-loop flutter suppression (FS) testing for that candidate control law commenced. During the closed-loop testing, the same excitations were inserted and responses were saved as during initial open-loop testing. At each test point, stability margins and open-loop plant stability were determined before proceeding to the next test point.

VII. DESCRIPTION OF THE CPE IMPLEMENTATION

The digital excitation, actuator commands, and sensor measurements used by the control law were transferred during testing to a SUN 3/160 computer with a SKY Warrior II array processor board where the FFT computations were performed using a FORTRAN 77 program, optimized to take advantage of the vector processing capabilities of the array processor. Each 2k FFT calculation took approximately .003 seconds. The matrix computations to obtain the singular values of return-difference matrices were also performed on the SUN 3/160 using MATLAB software operations (reference [10]). Figure 2 outlines the separate codes.

VIII. RESULTS AND DISCUSSION

Both SISO and MIMO flutter suppression control laws were designed for the AFW wind-tunnel model. During the wind-tunnel test, four FS control laws were tested using the AFW Digital Controller (reference [7]). Experimental data were used to evaluate their performance using the CPE capability presented in this paper. The process of obtaining experimental data and evaluating performance is described in the following.

The data for performing CPE was obtained by exciting, one at a time, all pairs of control surfaces used by the control law. Results from two tests performed in 1989 and in 1991 are presented in this paper. The excitation usually used in the 1989 tests was a 150 sec. logarithmic sine sweep (LogSS) over a frequency range varying from 4 to 35 Hz. However, low amplitude excitations, having low signal to noise ratios, were required at high dynamic pressures in order to keep the excitation itself from inducing flutter. The resulting CPE was poor and sometimes inconclusive. In the 1991 tests of the same model, a periodic pseudo noise (PPN) over a frequency range varying from 3 to 20 Hz was

generally used. The PPN (described in Appendix A) was designed to maximize the amplitude of the excitation (allowing maximum signal to noise ratio) within the actuator rate limits for a prescribed frequency range and specified frequency resolution. Typical transfer functions at one test condition resulting from both a logarithmic sine sweep and a PPN excitation are shown in figure 5. These results demonstrate that the higher amplitude PPN excitation, over the frequency range for which it was designed, provides smoother and more reliable results than the LogSS excitation.

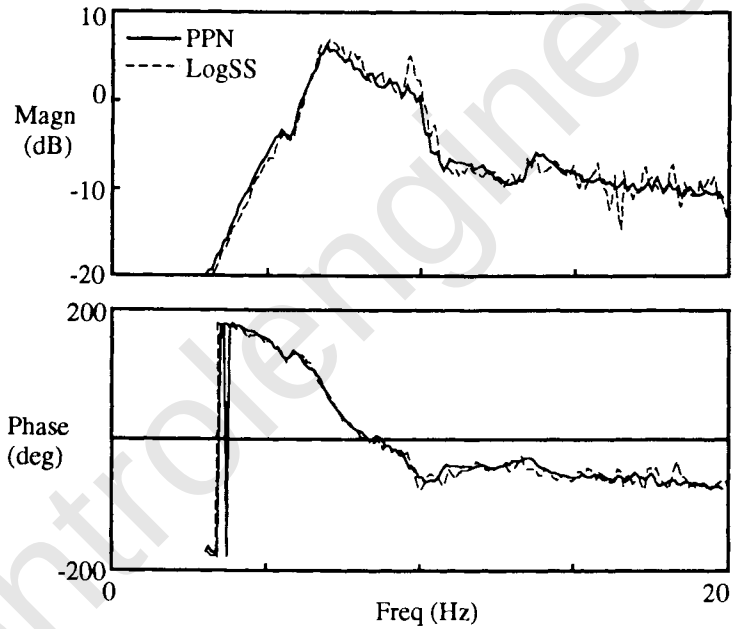


Figure 5. Comparison of transfer function $\ddot{z}_{tip} / \delta_{teo}$ for periodic random noise and logarithmic sine sweep excitation.

When FS control laws were required to control both symmetric and antisymmetric flutter, the CPE excitation was added to the control surfaces either symmetrically or antisymmetrically depending upon which symmetry was being evaluated. The responses, y, were then summed or differenced, depending upon

symmetry, before saving. While stability computations were being performed for symmetric control, the antisymmetric excitations could be performed and the transfer matrices generated. For this test, final results and plots were available within two minutes after the last excitation was performed, allowing near real-time controller performance evaluation.

Figure 6 shows an h-line (dynamic pressure vs. Mach number) plot with three test points identified which correspond to the points at which test results are presented herein. Points A and B correspond to cases for which a SISO control law is operating closed-loop and the plant is stable (A), and unstable (B). Point (C) identifies a test point at which another control law (in this case a MIMO) would have destabilized a stable plant if the loop were closed. CPE results are presented in the following discussions.

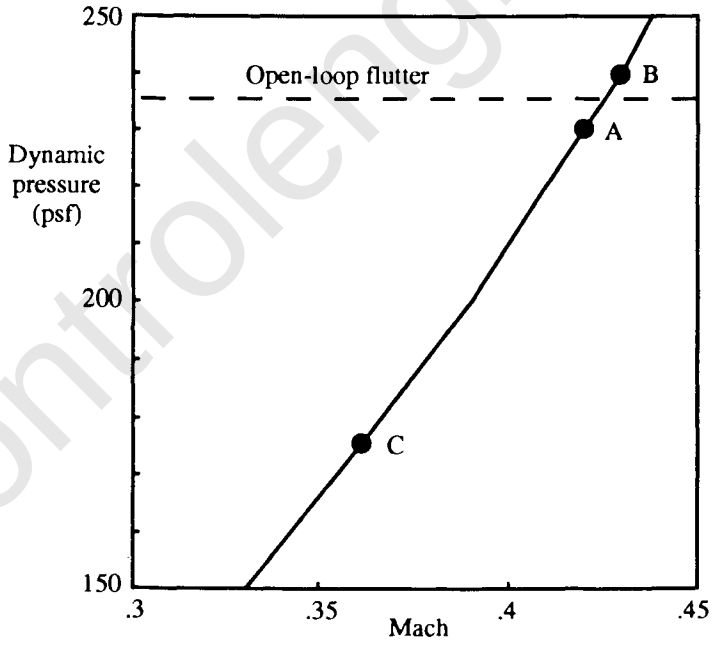
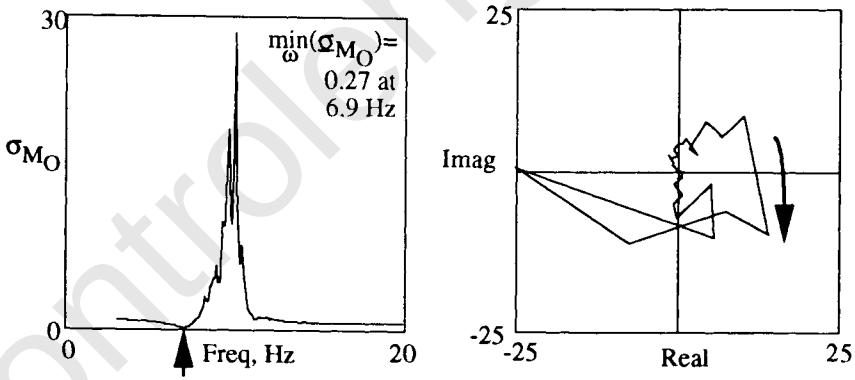


Figure 6. Atmospheric h-line showing flutter boundary and points for which results will be presented.

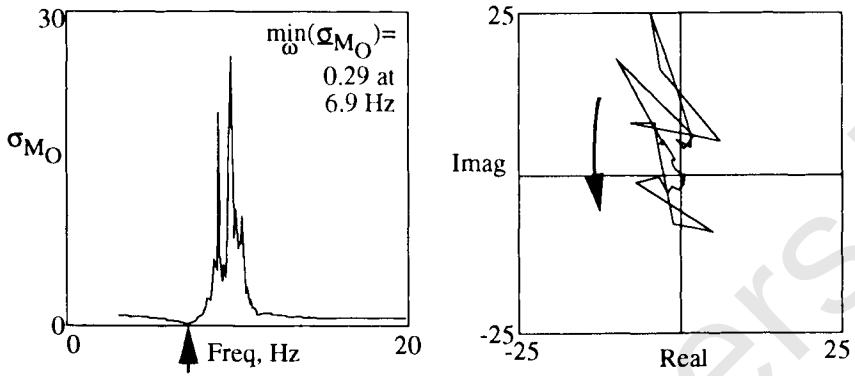
A. SISO CONTROL LAW

Typical CPE results for a symmetric SISO control law obtained during the closed-loop wind-tunnel tests are shown in figures 7 and 8. The determinant plot in figure 7 shows no encirclement about the origin (the critical point) at a dynamic pressure of 200 psf where the open-loop plant is known to be stable (point A of figure 6). Figure 8 shows the CPE results of a closed-loop system where the plant is unstable (point B on figure 6). Since there is one net counterclockwise encirclement of the critical point, these plots indicate that the controller is stabilizing the plant. Using the minimum singular value from figure 8, guaranteed stability margins can be obtained from the universal gain and phase diagram of figure 3. Since the minimum of σ_{M_O} is 0.29, the gain margin for zero phase margins are approximately -2.2 dB and +3.0 dB. For a 10 degree phase margin, however, the gain margins are -1.8 dB and +2.5 dB.



a) Singular value plot. b) Determinant plot.

Figure 7. Closed-loop CPE results for a symmetric SISO control law (open-loop plant is stable), M=.42, q=230 psf.



a) Singular value plot. b) Determinant plot.

Figure 8. Closed-loop CPE results for a symmetric SISO control law (open-loop plant is unstable), M=.43, q=240 psf..

B. MIMO CONTROL LAW

Results for a MIMO FS control law design (reference [6] with increased gain) are presented next. The initial open-loop testing performed at 150 psf indicated that the controller would not destabilize the model; hence, the loop was closed and testing continued. At approximately 175 psf, the closed-loop system appeared to become unstable where the open-loop plant was known to be stable. Consequently, open-loop testing was performed at 175 psf. The plots of $\underline{\sigma}_{M_I}$ and $\underline{\sigma}_{M_O}$ for the MIMO system are shown in figure 9 along with the maximum singular values, $\bar{\sigma}_{M_I}$ and $\bar{\sigma}_{M_O}$, in order to provide a visual indication of the condition number. As discussed previously, large condition numbers indicate uncertainty in the computation of the minimum singular values. Referring to the upper plot of figure 9, the ratio of the maximum, $\bar{\sigma}_{M_O}$, to the minimum singular value, $\underline{\sigma}_{M_O}$, (i.e. condition number) is large only in the vicinity of 7 Hz. Therefore, the low stability margins indicated by the minimum singular values in the vicinity of 20 Hz are fairly certain. Figure 9 also contains the

singular values for the evaluation of the additive perturbation, σ_A. The locus of the determinants of **I+HG** (lower right) shows an encirclement of the origin where the open-loop plant is known to be stable, thus indicating that the MIMO control law would be destabilizing as indicated during previous closed-loop testing. Upon further investigation, one of the elements of this destabilizing controller transfer matrix was examined and plots of the transfer function showed a peak magnitude close to 20 Hz.

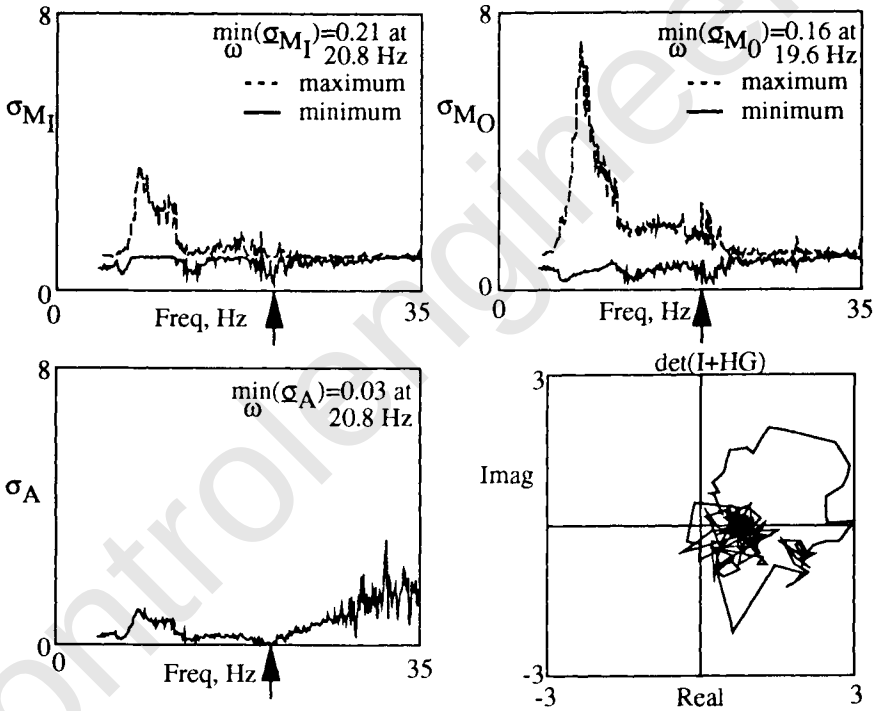


Figure 9. Open-loop results for a symmetric MIMO control law (open-loop plant is stable), M=0.36, q=175 psf.

C. FLUTTER BOUNDARY PREDICTION

One of the objectives of the wind-tunnel testing was to determine the open-loop flutter dynamic pressure from closed-loop experimental data. A method of determining whether the open-loop plant is stable or unstable is to count encirclements of the critical point in the determinant plot, as described in section IV. Figure 7 shows no net counterclockwise encirclement of the critical point. Since the closed-loop system is stable, this indicates the open-loop plant is stable. Figure 8 shows one counterclockwise encirclement of the critical point indicating the open-loop plant is unstable with one unstable pole (for positive frequencies). Figures 7 and 8 established that the flutter boundary was between 230 and 240 psf for the single-input single-output control system. Even though this prediction does not give a definite quantitative measure of the flutter boundary, it does set limits on where open-loop flutter occurs by setting upper and lower boundaries of where the open-loop system reaches neutral stability.

A more quantitative definition of the open-loop flutter boundary is obtained by tracking the minimums of the inverse maximum singular values (IMSV) (Eq. 10) of the plant extracted from closed-loop experimental data as a function of dynamic pressure. To do this, the IMSV are plotted as a function of frequency at each dynamic pressure. An example for a dynamic pressure of 150 psf is given in figure 10. The global minimum of the curve, identified by the arrow, indicates the mode which is going unstable and its frequency. The magnitude of this point approaching zero, indicates the proximity of the flutter mode of the open-loop system to neutral stability. Curves with two minimums approaching zero would indicate flutter is probably a result of two modes coalescing. In this case, the frequencies of the two modes coalesce at flutter.

A plot of the minimums from figures such as figure 10 are then plotted as a function of dynamic pressure. In this example, shown in figure 11, only one "global" minimum is being traced. Since the number of test points was limited,

the dashed part of the curve indicates the "best estimate" of the trajectory toward an instability. The point at which the inverse maximum singular values is zero is the point at which open-loop flutter occurs. This point, approximately 232 psf, is the predicted flutter dynamic pressure. Later open-loop testing to determine the actual open-loop flutter boundary indicated that the actual symmetric boundary point was 235 psf which corresponds well with the flutter prediction using plant transfer matrices extracted from closed-loop experimental data.

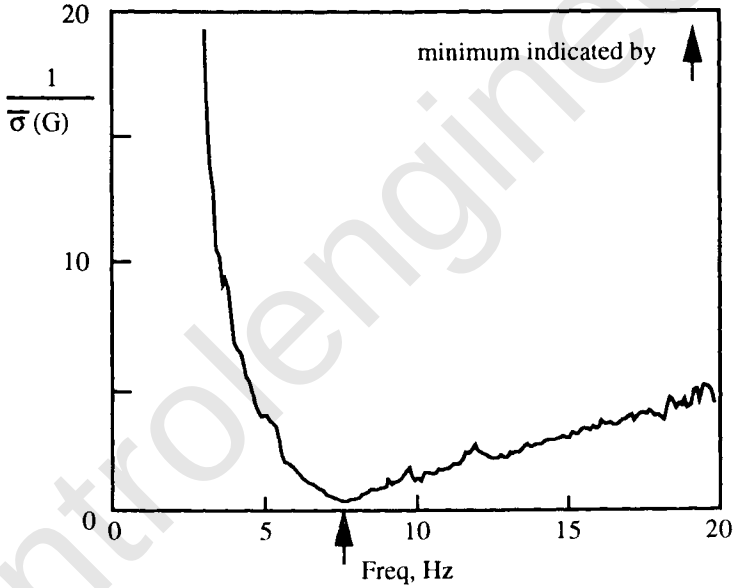


Figure 10. Plot of inverse maximum singular values of the open-loop plant transfer matrix, q=150 psf.

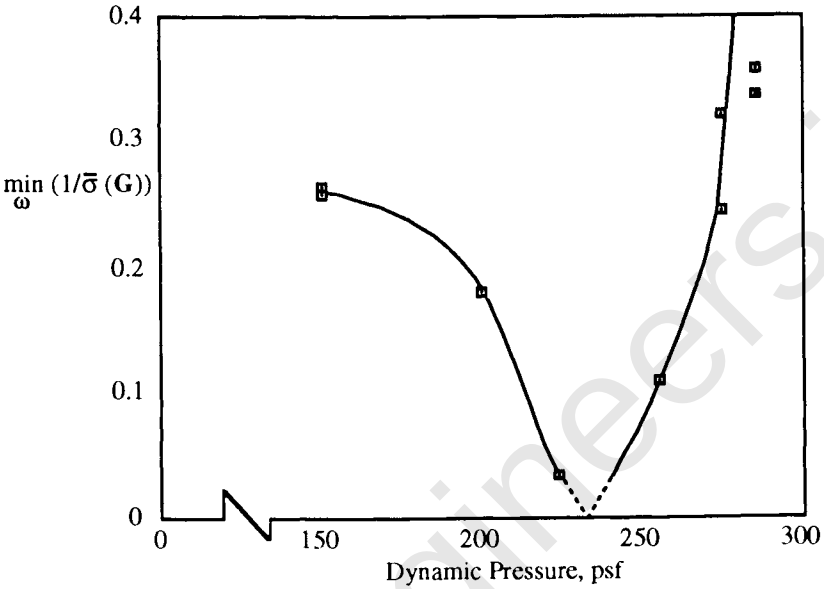


Figure 11. Flutter prediction using closed-loop CPE results.

X. CONCLUSIONS

A Controller Performance Evaluation (CPE) methodology was developed to evaluate the performance of multivariable, digital control systems. The method was used and subsequently validated during the wind-tunnel testing of an aeroelastic model equipped with a digital flutter suppression controller. Through the CPE effort, a wide range of sophisticated real-time analysis tools were developed. These tools proved extremely useful and worked very well during wind-tunnel testing. Moreover, results from open-loop CPE were the sole criteria for beginning closed-loop testing. In this way, CPE identified potentially destabilizing controllers before actually closing the loop on the control system, thereby helping to avoid catastrophic damage to the wind-tunnel model or the tunnel. CPE results also proved useful in determining open-loop plant stability during closed-loop test conditions.

X. APPENDIX A

The periodic pseudo noise (PPN) excitation developed for use in the AFW wind-tunnel test was developed to have a specific frequency content and to allow for maximum excitation amplitude subject to constraints on the rate. It is similar to periodic random noise or pseudo random noise described in reference [11], but it is not truly random and has a specified frequency content. It is generated by picking a block size which determines the frequency resolution. Time histories of sine sweeps with these frequencies over a finite time range defined by the block size are added or subtracted. Whether they are added or subtracted depends on which causes the least amount of increase in the maximum rate. After all the time histories have been combined the excitation is divided by the maximum amplitude to obtain an excitation with a unity maximum amplitude. The time histories are combined together starting with the sine sweep of the highest frequency.

XI. ACKNOWLEDGEMENTS

The authors wish to thank William M. Adams, Jr. for his continued insistence on the need for an on-line real-time controller performance evaluation capability and to acknowledge his original outline for the CPE procedures and the development of the Periodic Pseudo Noise excitation described herein. Any questions with regard to this PPN should be addressed to Mr. Adams. We would also like to thank Boyd Perry III for his coordination and guidance in carrying through the CPE effort.

XII. REFERENCES

1. Mukhopadhyay, V. and Newsom, J. R., "A Multiloop System Stability Margin Study Using Matrix Singular Values," *Journal of Guidance, Control, and Dynamics*, Vol. 7, September-October 1984, pp. 582-587.
2. Mukhopadhyay, V. and Newsom, J. R., "Application of Matrix Singular Value Properties for Evaluating Gain and Phase Margins of Multiloop Systems," AIAA Paper 82-1574, August 1982.
3. Mukhopadhyay, V.; Pototzky, A. S. and Fox, Matthew E.: "A Scheme for Theoretical and Experimental Evaluation of Multivariable System Stability Robustness," Proceedings of the 1990 American Control Conference, Paper No. FP 14-5, May, 1990, San Diego, California, pp. 3046-3047.
4. Miller, G. D., "Active Flexible Wing (AFW) Technology" AFWAL TR-87-3096, Feb. 1988.
5. Noll, T.; et al., "Aeroservoelastic Wind Tunnel Investigations Using Active Flexible Wing Model - Status and Recent Accomplishments," AIAA Paper 89-1168 (also NASA TM-101570), April 1989.
6. Perry III, B.; Mukhopadhyay, V.; Tiffany Hoadley, S.; Cole, S. R.; Buttrill, C. S.. and Houck, J. A., "Digital-Flutter-Suppression-System Investigations for the Active Flexible Wing Wind-Tunnel Model," AIAA Paper 90-1074 (also NASA TM-102618), April 1990.
7. Hoadley, S. H.; Buttrill, C.S.; McGraw, S. M.; and Houck, J. A., "Development, Simulation Validation, and Wind-Tunnel Testing of a Digital Controller System for Flutter Suppression", Paper presented at the Fourth Workshop on Computational Control of Flexible Aerospace Systems, July 11-13, 1990, Williamsburg, Virginia. NASA CP 10065, March 1991, pp. 583-613.

8. Adams, Jr., William M.; Tiffany, Sherwood H.; and Bardusch, Richard E., "Active Suppression of an 'Apparent Shock Induced Instability'", AIAA Paper 87-0881-CP, April 1987.

9. Pototzky, Anthony S.; Wieseman, Carol D.; Hoadley, Sherwood T.; and Mukhopadhyay, Vivek, "Development and Testing of Methodology for Evaluating the Performance of Multi-Input/Multi-Output Digital Control Systems," AIAA Paper 90-3501 (also NASA TM-102704), August 1990.

10. PRO-MATLAB User's Guide, The MathWorks Inc.; 24 Prime Park Way; Natick, MA 01760.

11. Olsen, Norm, "Excitation Functions for Structural Frequency Response Measurments", 2nd International Modal Analysis Conference, 1984.

This Page Intentionally Left Blank

Impulse Control of Piecewise Deterministic Systems

Oswaldo L. V. Costa

University of São Paulo
Dept. of Electronics Engineering
05508 900 São Paulo SP Brazil

I. INTRODUCTION

A variety of examples in the literature show the importance of impulse control problems in the theory of stochastic processes; inventory models, resource allocation problems and maintenance-replacement systems are some particular examples. In all these cases, control is taken by intervention, that is, the decision to act is taken at discrete times, in response to the random evolution of the system, and the process moved to a new point in the state space. Problems of this kind were first studied by Bensoussan and Lions [1], in the context of diffusion processes, as an application of variational and quasi-variational inequalities, and further developed in [2], [3]. Kushner [4], [5] considered similar problems, approximating the diffusion by controlled Markov chains. In [6] Lepeltier and Marchal considered the impulse control problem of right processes under very general assumptions by formulating it as a sequence of optimal stopping problems, an approach due to Robin [7]. Impulse control problems for Feller processes with the long run average cost were studied in [8], [9], [10], [11], [12]. Problems involving interventions in Markov decision drift processes [13] can be viewed as impulse control and have been analyzed in [14], [15] by discrete-time approximations and in [16], [17] by a direct deduction of optimality conditions.

CONTROL AND DYNAMIC SYSTEMS, VOL. 71

Piecewise deterministic Markov processes (PDP's) were introduced by Davis [18] as a general family of non-diffusion stochastic models suitable for formulating many optimization problems in several areas of operations research [19]. The motion of a PDP depends on three local characteristics, namely the flow ϕ, the jump rate λ and the transition measure Q, which specifies the post-jump location. Starting from x the motion of the process follows the flow $\phi(t,x)$ until the first jump time T_1 which occurs either spontaneously in a Poisson-like fashion with rate $\lambda(\phi(t,x))$ or when the flow hits the boundary of the state space. In either case the location of the process at the jump time T_1 is selected by the transition measure $Q(\ .\ ;\phi(T_1,x))$ and the motion restarts from this new point as before.

There are several different ways of approach to impulse control of PDP's. Dempster and Ye [20] transformed the problem of continuous control plus impulse control of PDP's into an equivalent continuous control problem. Gatarek [21], [22], [23] treated the impulse control problem by means of variational and quasi-variational inequalities with integral and first order differential terms. In this work we develop a more direct approach by formulating the impulse control of PDP's as a sequence of optimal stopping problems and using the special structure of PDP's to obtain recursive methods to characterize the value function of the problem. By taking this direct approach more specific results can be obtained and in particular the convergence of a discretization technique which leads to computational methods is derived.

This work is organized in the following way. Optimal stopping and impulse control are closely related since the latter can be regarded as an implicit optimal stopping problem where the gain function depends on the value function of the impulse control problem. Therefore optimal stopping constitutes an important step towards solving the more general problem of impulse control of PDP's and it is analyzed in section II. Characterization results for the value function, optimality conditions and discretization techniques leading to computational methods are presented. Our computational technique is attractive in that it reduces to a sequence of one-dimensional minimizations. In section III we apply the results of section II to obtain some recursive methods which characterize the value function of the impulse control problem of PDP's as well as provide a discretization technique. An illustrative example of this approximation method to the optimal maintenance of complex systems is presented. Finally in section IV some characterization results and optimality conditions for the value function of the long run average impulse control problem of PDP's are presented.

II. OPTIMAL STOPPING OF PDP's

A. PRELIMINARIES

In this section we deal with characterization results and approximation techniques for the optimal stopping of a PDP. The results obtained here are essential for the impulse control problem which we will see in sections III and IV. PDP's are right processes (but not Feller processes) and therefore the general theory in [6] can be applied. However due to the special features of PDP's more specific results can be obtained by a direct approach.

This section is organized in the following way. In subsection B we give the main definitions and general assumptions. Gugerli [24] showed that the value function of the optimal stopping problem of a PDP is "continuous along trajectories" of the flow $\phi(t,x)$ provided that the gain function also satisfies this property. However under general assumptions it is not to be expected that the "implicit" gain function of an impulse control problem of a PDP (seen as an "implicit" optimal stopping problem) will be "continuous along trajectories". But this result will hold if we replace "continuity along trajectory" by "lower semi-analicity". Subsection C presents an extension of the results in [24] to the case when the gain function is lower semi-analytic so that this generality can be applied in section III for the impulse control problem. In subsection D some optimality equations similar to the ones in [17] are derived. In subsection E we present a computational technique for solving the optimal stopping problem of a PDP. We construct a discretized PDP which retains the main characteristics of the original process and we show convergence of the payoff functions. This technique reduces the optimal stopping problem to a sequence of one dimensional minimizations. The results of this section follow those derived in [25].

B. NOTATIONS AND DEFINITIONS

For any Borel space $\mathfrak{S}$ we denote by $\sigma(\mathfrak{S})$ the Borel σ-field generated by $\mathfrak{S}$ and by $B^*(\mathfrak{S})$ ($B(\mathfrak{S})$, $C(\mathfrak{S})$ respectively) the space of real valued bounded lower semi-analytic (Borel measurable, continuous) functions on $\mathfrak{S}$. The set of non-negative integers $\{0,1,...\}$ and non-negative reals is denoted by $\mathbf{N}$ and R_+ respectively. Let E be an open subset of R^d, ∂E the boundary of E, $\mathcal{E}$ the borel σ-field of E and $\mathfrak{U}(E)$ the universally measurable σ-field of E. Let Ω denote the space of E-valued functions on

$[0,\infty)$ which are right continuous and have left hand limits at each $t < \infty$, $x_t(\omega)=\omega_t$ for every $\omega \in \Omega$, $\mathcal{F}_t^0=\sigma\{x_s,\ s \le t\}$ and $\mathcal{F}^0=\mathcal{F}_\infty$. We will consider a PDP taking values in E and determined by the following parameters :

a) the flow $\phi(t,x)$ of a Lipschitz continuous vector field $\mathcal{L}$ (cf. [18]).

b) the jump rate $\lambda(.) : E \to R_+$.

c) the transition measure $Q(.;.)$: $Ex(E \cup \partial^*E) \to [0,1]$ where

$$\partial^*E := \{\ z \in \partial E;\ \phi(-t,z) \in E \ \text{ for all } t \in (0,\epsilon) \text{ and some } \epsilon > 0\ \}.$$

Note that ∂^*E represents those boundary points at which the flow exits from E. We define

$$t^*(x) := \inf\{\ t > 0\ ;\ \phi(t,x) \in \partial^*E\ \}$$

for every $x \in E$. The general assumptions of the PDP, which we shall denote by (X_t), are as in Davis [18]. In addition to those assumptions we will also suppose that λ is bounded. The motion of the process (X_t) starting from x is constructed in the following way. Take a random variable T_1 such that :

$$P(T_1 > t) = \begin{cases} \exp\left(-\int_0^t \lambda(\phi(s,x))ds\right) & , t < t^*(x) \\ 0 & , t \ge t^*(x) \end{cases}$$

Now select independently an E-valued random variable having distribution $Q(.;\phi(T_1,x))$. The trajectory of X_t for $t \le T_1$ is given by

$$X_t = \begin{cases} \phi(t,x) & , t < T_1 \\ Z_1 & , t = T_1 \end{cases}$$

Starting from $X_{T_1} = Z_1$ we now select the next inter-jump time T_2-T_1 and post-jump location $X_{T_2} = Z_2$ in a similar way. This gives a piecewise deterministic trajectory for the process (X_t) with jump times T_1, T_2, ... and post-jump locations Z_1, Z_2, ... It is convenient to write that $T_0 = 0$ and $Z_0 = x$ (the initial point). We will denote P_x the probability law on $(\Omega,\mathcal{F}^0)$ of the PDP starting at x. We assume that

$$\lim_{n\to\infty} T_n = \infty \quad P_x\text{-a.s.} .$$

For $x \in E$ we define :

i) $Qv(x) := \int_E v(y)Q(dy;x)$ where $v \in B^*(E)$. Note that $Q(.;x) : E \to [0,1]$ can be uniquely extended to $\mathcal{U}(E)$ and thus the above integral is well defined on the measurable space $(E,\mathcal{U}(E))$.

ii) $\Lambda(t,x) := \int_0^t \lambda(\phi(s,x))ds$ where $0 \leq t \leq t^*(x)$.

Finally we say that $g \in B^c(E)$ if $g \in B(E)$, $g(\phi(.,x)): [0,t^*(x)) \to R$ is continuous and $\lim_{t\to t^*(x)} g(\phi(t,x))$ exists whenever $\exp(-\Lambda(t^*(x),x)) \neq 0$.

C. CHARACTERIZATION RESULTS

Let $\mathcal{F}$ be the universal completion of $\mathcal{F}^0$ and $\mathcal{F}_t$ the right continuous universal completion of $\mathcal{F}^0_t$. We denote by $\mathcal{M}$ the set of all $\mathcal{F}_t$-stopping times which are P_x-a.s. finite for all $x \in E$. We will consider the following optimal stopping problem for $x \in E$

$$\rho'(x) := \inf_{\tau \in \mathcal{M}} E_x(g (X_\tau)). \qquad (1)$$

In [24] Gugerli studied (1) when $g \in B^c(E)$. In this subsection we will extend the results in [24] to the case when $g \in B^*(E)$. As mentioned before this extension is needed to study the impulse control problem of PDP's.

For $x \in E$, $0 \leq t < t^*(x)$ and v_1, v_2 in $B^*(E)$ define

a) $J(v_1,v_2)(t,x) := E_x(v_1(\phi(t,x))1_{\{T_1>t\}} + v_2(Z_1)1_{\{T_1 \leq t\}}) =$

$$v_1(\phi(t,x))e^{-\Lambda(t,x)} + \int_0^t Qv_2(\phi(s,x))\lambda(\phi(s,x))e^{-\Lambda(s,x)}ds$$

$$\text{b) } Kv_2(x) := E_x(v_2(Z_1)) = \int_0^{t^*(x)} Qv_2(\phi(s,x))\lambda(\phi(s,x))e^{-\Lambda(s,x)}ds \; +$$

$$Qv_2(\phi(t^*(x),x))e^{-\Lambda(t^*(x),x)}$$

$$\text{c) } L(v_1,v_2)(x) := \{ \inf_{0 \le t < t^*(x)} J(v_1,v_2)(t,x)\} \wedge Kv_2(x).$$

It follows from the semi-group property of the drift ϕ that

$$L(v_1,v_2)(\phi(t,x)) = e^{\Lambda(t,x)}\Big(\{ \inf_{t \le s < t^*(x)} J(v_1,v_2)(s,x)\} \wedge Kv_2(x) -$$

$$E_x(v_2(Z_1)1_{\{T_1 \le t\}})\Big). \qquad (2)$$

We can easily show that $L(v_1,v_2)(\,.\,)\colon E \to R$ is in $B^*(E)$. Define the sequence of functions ρ_m, $m \in \mathbb{N}$ by

$$\rho_0 := g \,, \; \rho_{m+1} := L(g,\rho_m).$$

Then $\rho_m \in B^*(E)$ for all $m \in \mathbb{N}$ and since $\rho_1 \le g$, $\rho_{m+1} \le \rho_m$ for all $m \in \mathbb{N}$. Therefore

$$\rho := \inf_m \rho_m = \lim_{m \to \infty} \rho_m$$

exists and from Lemma 7.30(2) of [26], $\rho \in B^*(E)$. Corollary 1 of Gugerli [24] is readily modified to show the following results.

Proposition 1 : Let ρ and ρ_m be defined as above . Then

a) $\rho_{m+1} = L(\rho_m,\rho_m)$.

b) ρ is the biggest solution of

$$\begin{cases} v = L(g,v) \\ v \in B^*(E). \end{cases}$$

c) ρ is the biggest solution of

$$\begin{cases} v = L(v,v) \\ v \le g \;, \; v \in B^*(E). \end{cases}$$

Function ρ can also be characterized in the following way;

Proposition 2 : ρ is the biggest solution of

$$\begin{cases} v(x) \leq E_x(v(X_{t \wedge T_1})), \forall\, t \in [0,t^*(x)] \ , \ \forall\, x \in E \\ v \leq g \ , \ v \in B^*(E) \end{cases} \tag{3}$$

Proof : Noting that

$$L(\rho,\rho)(x) = \{ \inf_{0 \leq t < t^*(x)} E_x(\rho(X_{t \wedge T_1}))\} \wedge E_x(\rho(Z_1))$$

it follows from Prop. 1c) that for any $x \in E$

$$\rho(x) \leq E_x(\rho(X_{t \wedge T_1})) \text{ for all } t \in [0,t^*(x)) \text{ and } \rho(x) \leq E_x(\rho(Z_1)) .$$

Hence ρ is a solution of Eq. (3). Suppose v is another solution of Eq. (3). We use induction in m to show that $\rho_m \geq v$ for all $m \in \mathbb{N}$. For $m = 0$, $\rho_0 = g \geq v$. Suppose now that $\rho_m \geq v$. It follows from Prop. 1a) that $\rho_{m+1} = L(\rho_m,\rho_m) \geq L(v,v) \geq v$. Therefore

$$\rho = \lim_{m \to \infty} \rho_m \geq v,$$

proving the Proposition. □

The importance of the function ρ is that it equals the value function of the optimal stopping problem ρ' defined in Eq. (1) for all points in E, that is, the following result, which was proved in [25], holds:

Proposition 3 : $\rho(x) = \rho'(x)$ for all $x \in E$.

D. OPTIMALITY EQUATIONS

This subsection is devoted to show a connection between the solutions of $v = L(g,v)$ and the optimality equations of the kind presented in [17] for the optimal stopping problem of a PDP. Let $B^{ac}(E)$ denote the space of functions g in $B^c(E)$ such that for each $x \in E$, $g(\phi(.,x))$: $[0,t^*(x)) \to R$ is absolutely continuous. First we are going to show an auxiliary result . Define

$$L(g,v)(t,x) := \{ \inf_{t \leq s < t^*(x)} J(g,v)(s,x) \} \wedge Kv(x) \ , \ t \in [0,t^*(x)).$$

Proposition 4 : Suppose that $g \in B^{ac}(E)$ and $v \in B(E)$. Then for every $x \in E$, $L(g,v)(\ .\ ,x) : [0,t^*(x)) \to R$ is absolutely continuous.

Proof : Since $g(\phi(\ .\ ,x))$ is absolutely continuous it is clear that

$$J(g,v)(t,x) = e^{-\alpha t - \Lambda(t,x)} g(\phi(t,x)) + \int_0^t e^{-\alpha s - \Lambda(s,x)} \lambda(\phi(s,x)) Q v(\phi(s,x)) ds$$

is also absolutely continuous in t. Therefore for every compact set $\mathcal{C}$ of $[0,t^*(x))$ given $\epsilon > 0$ there exists $\delta > 0$ such that for each finite collection $[t_i,r_i]$, $i = 1,...,n$ of non-overlapping intervals of $\mathcal{C}$

$$\sum_{i=1}^{n} | J(g,v)(r_i,x) - J(g,v)(t_i,x) | < \epsilon \quad \text{if} \quad \sum_{i=1}^{n} (r_i - t_i) < \delta .$$

For each $i = 1,...,n$ we have one of the possibilities below :

i) $L(g,v)(t_i,x) = J(g,v)(t_i',x)$ for some $t_i \le t_i' < r_i$ and in this case it is clear that

$$L(g,v)(r_i,x) - L(g,v)(t_i,x) = L(g,v)(r_i,x) - L(g,v)(t_i',x) =$$

$$L(g,v)(r_i,x) - J(g,v)(t_i',x) \le J(g,v)(r_i,x) - J(g,v)(t_i',x).$$

Note also that $r_i - t_i' \le r_i - t_i$.

ii) $L(g,v)(t_i,x) = L(g,v)(r_i,x)$ and in this case we define $t_i' = r_i$.

Then from the fact that $L(g,v)(t,x)$ is increasing in t we have

$$\sum_{i=1}^{n} | L(g,v)(r_i,x) - L(g,v)(t_i,x) | = \sum_{i=1}^{n} (L(g,v)(r_i,x) - L(g,v)(t_i,x)) =$$

$$\sum_{i=1}^{n} (L(g,v)(r_i,x) - L(g,v)(t_i',x)) \le \sum_{i=1}^{n} (J(g,v)(r_i,x) - J(g,v)(t_i',x)) \le$$

$$\sum_{i=1}^{n} | J(g,v)(r_i,x) - J(g,v)(t_i',x) | < \epsilon$$

since $\sum_{i=1}^{n} (r_i - t_i') \le \sum_{i=1}^{n} (r_i - t_i) < \delta .$ □

Consider $v \in B(E)$ such that $v(x) = L(g,v)(x)$, $g \in B^{ac}(E)$ for all $x \in E$. Since from Eq. (2),

$$v(\phi(t,x)) = e^{\alpha t+\Lambda(t,x)}\Big(L(g,v)(t,x) - \int_0^t e^{-\alpha s-\Lambda(s,x)}\lambda(\phi(s,x))Qv(\phi(s,x))ds\Big)$$

it is clear from Proposition 4 that $v \in B^{ac}(E)$. Let $\mathcal{L}$ and $\mathbb{A}$ be respectively the vector field and the extended generator of the PDP (X_t). The domain of $\mathbb{A}$, $\mathfrak{D}(\mathbb{A})$, has been characterized in Theorem 5.5 of Davis [18]. We have that if $v \in B^{ac}(E)$ and $v(x) := Qv(x)$ for $x \in \partial^* E$ we get $v \in \mathfrak{D}(\mathbb{A})$ and

$$\mathbb{A}v(x) = \mathcal{L}v(x) + \lambda(x)\int_E (v(y) - v(x))Q(dy,x), \quad \forall\ x \in \mathbb{E}$$

where

$$\mathbb{E} := \{\ x \in E;\ \mathcal{L}v(x) = D^+v(x) := \frac{dv(\phi(t,x))}{dt}\Big|_{t=0} \quad \text{exists at } x\ \}$$

(recall that since $v(\phi(\ .\ ,x))$ is absolutely continuous the derivative exists almost everywhere on $[0,t^*(x))$).

Proposition 5 : For $g \in B^{ac}(E)$ let v be a solution of

$$\begin{cases} v(x) = L(g,v)(x), & \forall\ x \in E \\ v \in B^{ac}(E) \end{cases}$$

Then

$$\begin{cases} \mathbb{A}v(x) - \alpha v(x) \geq 0 & , \quad \forall\ x \in \mathbb{E} \\ v(x) \leq g(x) & , \quad \forall\ x \in E \\ (\mathbb{A}v(x) - \alpha v(x))(v(x) - g(x)) = 0 & , \quad \forall\ x \in \mathbb{E} \end{cases}$$

Proof : Since $L(g,v)(\ .\ ,x) : [0,t^*(x)) \to R$ is absolutely continuous and increasing we have that $\dfrac{dL(g,v)(t,x)}{dt}$ exists almost everywhere and is positive. From

$$v(\phi(t,x)) = e^{\alpha t+\Lambda(t,x)}\Big(L(g,v)(t,x) - \int_0^t e^{-\alpha s-\Lambda(s,x)}\lambda(\phi(s,x))Qv(\phi(s,x))ds\Big)$$

we get that for $x \in \mathbb{E}$

$$\mathcal{L}v(x) = D^+v(x) = (\alpha + \lambda(x))v(x) + D^+L(g,v)(0,x) - \lambda(x)Qv(x)$$

and thus

$$\mathcal{L}v(x) + \lambda(x)\int_E (v(y) - v(x))Q(dy;x) - \alpha v(x) =$$

$$\mathcal{A}v(x) - \alpha v(x) = D^+L(g,v)(0,x) \geq 0 .$$

From $v(x) = L(g,v)(x)$ it is immediate that $v(x) \leq g(x)$, $\forall\, x \in E$. Now if for $x \in \mathbb{E}$, $L(g,v)(0,x) = v(x) < g(x) = J(g,v)(0,x)$ then from continuity of $L(g,v)(.,x)$ and $J(g,v)(.,x)$ we have that for some $\epsilon > 0$, $L(g,v)(t,x) < J(g,v)(t,x)$ for all $t \in [0,\epsilon)$ and therefore $L(g,v)(t,x)$ is constant on $[0,\epsilon)$. It means that if $v(x) < g(x)$ and $x \in \mathbb{E}$ then $D^+L(g,v)(0,x) = 0 = \mathcal{A}v(x) - \alpha v(x)$, proving the Proposition. □

E. DISCRETIZATION RESULTS

1. AUXILIARY RESULT

In this subsection we will consider the discounted optimal stopping problem

$$\rho'(x) = \inf_{\tau \in \mathcal{M}_\infty} E_x(e^{-\alpha\tau}g(X_\tau))$$

where $\alpha > 0$ and $\mathcal{M}_\infty$ represents the set of all $\mathcal{F}_t$-stopping times (including ∞). The results of subsection C can be readily modified for this case and, as we saw in Prop. 3, ρ is equal to ρ' for all points in E. Suppose $g \in B^*(E)$ is defined and bounded on ∂^*E. For $x \in \partial^*E$ set:

$$\rho_{m+1}(x) := g(x) \wedge E_x(\rho_m(Z_1)), \quad \rho_0(x) := g(x)$$

$$\rho(x) := g(x) \wedge E_x(\rho(Z_1)).$$

Remark : Note that for $x \in \partial^*E$, $\rho'(x) = E_x(\rho(Z_1))$ and thus may be different from $\rho(x)$.

Define $\tilde{E} := E \cup \partial^*E$. The following Proposition can be easily proved.

Proposition 6 : For $x \in \tilde{E}$ and $m \in \mathbb{N}$

$$0 \leq \rho_m(x) - \rho(x) \leq E_x(e^{-\alpha T_m}(g(Z_m) - \rho(Z_m))) .$$

2. ASSUMPTIONS

We add the following assumptions to the previous ones:

1) $t^*(.) : \tilde{E} \to R_+ \cup \{ \infty \}$ is continuous ($t^*(x) := 0$ for $x \in \partial^* E$).

2) $\lambda(\phi(.,y)) \to \lambda(\phi(.,x))$ as $y \to x$ Lebesgue a. s. on $[0,t^*(x))$ for all $x \in E$.

3) For every $v \in C(E)$

3.i) $Qv(.) : E \to R$ is continuous.

3.ii) $Qv(.) : \partial^* E \to R$ is continuous.

4) g is continuous and bounded by a_1 on $\tilde{E}$.

5) $\Lambda(t^*(x),x) = \infty$ whenever $t^*(x) = \infty$.

From the assumptions above we can easily show the following Proposition.

Proposition 7 : $\Lambda(.,.): \{(t,x);\ x \in \tilde{E},\ 0 \le t \le t^*(x)\ \} \to R_+ \cup \{\infty\}$ is continuous.

For $x \in \tilde{E}$ let us denote by μ_x the joint probability measure of (T_1,Z_1) when $Z_0 = x$ ((T_1^x,Z_1^x) for short) on $(R^{d+1},\sigma(R^{d+1}))$. We have that for any $t \in R_+ \cup \{\infty\}$ and $A \in E$

$$P_x(\ T_1 \le t\ ,\ Z_1 \in A\ ) = \int_0^{t \wedge t^*(x)} Q(\ A\ ;\ \phi(s,x)\ )\lambda(\phi(s,x))\ e^{-\Lambda(s,x)} ds$$

$$+\ Q(A;\phi(t^*(x),x)\ )e^{-\Lambda(t^*(x),x)} 1_{\{\ [t^*(x),\infty]\ \}}(t)\ .$$

Denoting by $\mathcal{P}(R^{d+1})$ the space of all probability measures over $\sigma(R^{d+1})$, we have the following Proposition:

Proposition 8 : $\mu_{(.)}: \tilde{E} \to \mathcal{P}(R^{d+1})$ is continuous with respect to the weak topology on $\mathcal{P}(R^{d+1})$.

Outline of the proof : It is enough to show that for every real valued continuous and bounded function f on $R_+ \times E$ and arbitrary $x \in E \cup \partial^* E$,

$$\lim_{y \to x} E_y(f(T_1,Z_1)) = E_x(f(T_1,Z_1)).$$

□

3. DISCRETIZED PROCESS

Let $D := \{z_1, z_2, \ldots\}$ be a countable dense set in E and $D^N := \{z_1, \ldots, z_N\}$. Define the sets $\{A_i^N\}_{i=1}^N$ in the following form:

$$B_i^N := \{z \in E;\ |z - z_i| \leq |z - z_j| \text{ for every } j = 1,\ldots,N\},\ i = 1,\ldots,N$$

$$A_1^N := B_1^N, \quad A_i^N := B_i^N - \bigcup_{j=1}^{i-1} A_j^N \qquad i = 2,\ldots,N\ .$$

The discretized PDP, denoted by (X_t^N), is determined by:

a) the flow $\phi(t,x)$.

b) the jump rate λ.

c) the transition probability Q^N defined as follows :

$$Q^N(\{z_i\}; x) := Q(A_i^N; x) \quad \text{for } i = 1,\ldots, N \text{ and all } x \in \tilde{E}.$$

Therefore jumping from x the discretized process can go to a finite number of points $\{z_1,\ldots,z_N\}$ only. From now on whenever necessary we will use the superscript N to distinguish the discretized process from the original one.

We will be concerned now with the construction of (T_1, Z_1) and (T_1^N, Z_1^N). For arbitrary $x \in \tilde{E}$ we have seen in Prop. 8 that $\mu_y \to \mu_x$ weakly as $y \to x$. Let P denote the Lebesgue-[0,1] measure. Then using Skorohod's Theorem (Billingsley [27] page 337) we can construct (T_1^y, Z_1^y) and (T_1^x, Z_1^x) measurable functions from the probability space $([0,1], \sigma([0,1]), P)$ into $(R^{d+1}, \sigma(R^{d+1}))$ such that (T_1^y, Z_1^y) has distribution μ_y, (T_1^x, Z_1^x) has distribution μ_x and

$$(T_1^y, Z_1^y) \to (T_1^x, Z_1^x) \quad \text{as } y \to x \qquad P\text{-a.s.}\ .$$

For the discretized case the construction is the following: for every $\nu \in E$ and $u \in [0,1]$

$$T_1^{N\nu}(u) = T_1^{\nu}(u)$$

$$Z_1^{N\nu}(u) = z_i \quad \text{if} \quad Z_1^{\nu}(u) \in A_i^N\ .$$

It is easy to show that $(T_1^{N\nu}, Z_1^{N\nu})$ as defined above agrees with the definition of the discretized process (X_t^N). For the next result we consider the construction seen above for some arbitrary fixed $x \in \tilde{E}$.

Proposition 9 : $Z_1^{Ny} \to Z_1^x$ as $y \to x$ and $N \to \infty$ P-a.s. .

Proof : Define the set $U := \{ u \in [0,1] ; Z_1^y(u) \to Z_1^x(u) \text{ as } y \to x \}$. As we saw above, $P(U) = 1$. Fix $u \in U$. For any $\epsilon > 0$ there exists $z_i \in D$ such that

$$| Z_1^x(u) - z_i | < \epsilon/2$$

because D is dense in E. Take $\delta > 0$ such that

$$| Z_1^y(u) - Z_1^x(u) | < \epsilon/4 \text{ whenever } | x-y | < \delta$$

and $N_0 \in \mathbb{N}$ such that $z_i \in D^N$ whenever $N \geq N_0$. Thus for $N \geq N_0$ and $| x - y | < \delta$,

$$\begin{aligned} | Z_1^x(u) - Z_1^{Ny}(u) | &\leq | Z_1^x(u) - Z_1^y(u) | + | Z_1^y(u) - Z_1^{Ny}(u) | \\ &\leq | Z_1^x(u) - Z_1^y(u) | + | Z_1^y(u) - z_i | \end{aligned}$$

where the last inequality follows from the definition of the sets $\{A_i^N\}$. Hence

$$\begin{aligned} | Z_1^x(u) - Z_1^{Ny}(u) | &\leq | Z_1^x(u) - Z_1^y(u) | + | Z_1^y(u) - z_i | \\ &\leq | Z_1^x(u) - Z_1^y(u) | + | Z_1^y(u) - Z_1^x(u) | + | Z_1^x(u) - z_i | \\ &= 2 | Z_1^x(u) - Z_1^y(u) | + | Z_1^x(u) - z_i | < \epsilon \end{aligned}$$

which proves the Proposition. □

The following result follows from the above Proposition:

Proposition 10 : For every $m \in \mathbb{N}$

a) $E_{(.)}(\exp\{-\alpha T_m\}) : \tilde{E} \to R_+$ is continuous.

b) $E_y^N (\exp \{ -\alpha T_m^N \}) \to E_x (\exp \{ -\alpha T_m \})$ as $y \to x$ and $N \to \infty$ for every $x \in \tilde{E}$.

Proof : We show b) only since the proof of a) is similar. For arbitrary x

in $\tilde{E}$ we use the construction seen above and prove the Proposition by induction on m. For m = 0 the result is clear. Suppose b) holds for m. Define for $z \in \tilde{E}$, $\chi(z) := E_z(\exp\{-\alpha T_m\})$ and $\chi^N(z) := E_z^N(\exp\{-\alpha T_m^N\})$. Then from time homogeneity and the strong Markov property

$$E_y^N(\exp\{-\alpha T_{m+1}^N\}) = E_y^N(\exp\{-\alpha T_1^N\} E^N(\exp\{-\alpha(T_{m+1}^N - T_1^N)\}/\mathcal{F}_{T_1^N})) =$$

$$E_y^N(\exp\{-\alpha T_1^N\}\chi^N(Z_1^N)) = E(\exp\{-\alpha T_1^{Ny}\}\chi^N(Z_1^{Ny})) \qquad (4)$$

where the last expectation is over [0,1] with measure P. Then from the construction of T_1^{Ny} ($= T_1^y$), Prop. 9 and the induction hypothesis for m

$$\exp\{-\alpha T_1^{Ny}\}\chi^N(Z_1^{Ny}) \to \exp\{-\alpha T_1^x\}\chi(Z_1^x) \qquad (5)$$

as $y \to x$ and $N \to \infty$ P-a.s. From (4), (5) and the bounded convergence theorem we obtain the desired result. □

4. CONVERGENCE RESULTS

Define the following discretized optimal stopping problem

$$\rho^N(x) = \inf_{\tau \in \mathcal{M}_\infty^N} E_x^N(e^{-\alpha\tau} g^N(X_\tau^N))$$

where $g^N \in B^*(E)$, g^N is defined and bounded on $\partial^* E$, and $g^N \to g$ uniformly on compact sets of $\tilde{E}$ as $N \to \infty$. We allow this extra generality on g^N to use the results we get here in the impulse control problem. From the results of subsection C we have $\rho_m^N \downarrow \rho^N$ as $m \to \infty$. The following equivalence will be used in the sequel:

Proposition 11 : Suppose $v: \tilde{E} \to R$ is continuous. The following statements are equivalent :

a) $v^N \overset{N \to \infty}{\to} v$ uniformly on compact sets of $\tilde{E}$.

b) $v^N(y) \to v(x)$ as $y \to x$ and $N \to \infty$ for all $x \in \tilde{E}$.

Continuity of ρ and convergence of ρ^N to ρ will follow from the next Theorem.

Theorem 1 : For every $m \in \mathbb{N}$

a) $\rho_m(\,.\,) : \tilde{E} \to R$ is continuous.

b) $\rho_m^N(y) \to \rho_m(x)$ as $y \to x$ and $N \to \infty$ for every $x \in \tilde{E}$.

Proof : Again we show b) only by applying induction on m. For $m = 0$ the result is immediate from the hypothesis and using Prop. 11 ($\rho_0^N(y) = g^N(y) \to g(x) = \rho_0(x)$ as $y \to x$ and $N \to \infty$). Suppose b) holds for m. We will show the result for $x \in E$ (the proof for the case $x \in \partial^* E$ is similar). The proof follows from the following steps :

Step 1 : For any $t \in [0, t^*(x))$

i) $E_y^N(\exp\{-\alpha T_1^N\}\rho_m^N(Z_1^N)1_{\{T_1^N \le t\}}) \to E_x(\exp\{-\alpha T_1\}\rho_m(Z_1)1_{\{T_1 \le t\}})$
as $y \to x$ and $N \to \infty$.

ii) $E_y^N(\exp\{-\alpha T_1^N\}\rho_m^N(Z_1^N)) \to E_x(\exp\{-\alpha T_1\}\rho_m(Z_1))$ as $y \to x$ and $N \to \infty$.

Step 2 : For any sequence $y_k \to x$ as $k \to \infty$

$$\lim_{k \to \infty} P\left\{ T_1^x = t^*(x) \,,\, T_1^{y_k} < t^*(y_k) \right\} = 0.$$

Step 3 : For any sequence $y_k \to x$, $N_k \to \infty$, $t_k \to t \le t^*(x)$ as $k \to \infty$ where $t_k < t^*(y_k)$ for all k, we have that

$$E_{y_k}^{N_k}(\exp\{-\alpha T_1^{N_k}\}\rho_m^{N_k}(Z_1^{N_k})1_{\{T_1^{N_k} \le t_k\}}) \overset{k \to \infty}{\to}$$

$$\begin{cases} E_x(\exp\{-\alpha T_1\}\rho_m(Z_1)1_{\{T_1 \le t\}}) & \text{if } t < t^*(x) \\ E_x(\exp\{-\alpha T_1\}\rho_m(Z_1)1_{\{T_1 < t\}}) & \text{if } t = t^*(x) \end{cases}$$

Step 4 : $\inf_{0 \le t < t^*(y)} J^N(g^N, \rho_m^N)(t,y) \to \inf_{0 \le t < t^*(x)} J(g, \rho_m)(t,x)$ as $y \to x$ and $N \to \infty$.

From the construction of T_1^{Ny}, Proposition 11, the induction hypothesis for m, the bounded convergence theorem and the fact that

$$P(\, T_1^x = t\,) = P_x(\, T_1 = t\,) = 0$$

we get step 1. It is clear that

$$P\left\{T_1^{y_k} < t^*(y_k),\ T_1^x = t^*(x)\right\} = P\left\{\, T_1^x = t^*(x)\,\right\} - P\left\{\, T_1^{y_k} = t^*(y_k)\,\right\} +$$

$$P\left\{\, T_1^{y_k} = t^*(y_k)\ ,\ T_1^x < t^*(x)\,\right\} = e^{-\Lambda(t^*(x),x)} - e^{-\Lambda(t^*(y_k),y_k)} +$$

$$P\left\{T_1^{y_k} = t^*(y_k)\ ,\ T_1^x < t^*(x)\,\right\}.$$

From continuity of $\Lambda(t^*(\ .\ ),\ .\ )$ on E we get

$$\limsup_{k\to\infty} P\left\{T_1^{y_k} < t^*(y_k),\ T_1^x = t^*(x)\right\} = \limsup_{k\to\infty} P\left\{\, T_1^{y_k} = t^*(y_k),\ T_1^x < t^*(x)\,\right\}$$

$$\leq P\left\{\, \limsup_{k\to\infty} \{T_1^{y_k} = t^*(y_k)\}\ ,\ T_1^x < t^*(x)\,\right\}. \tag{6}$$

But since $T_1^{y_k} \to T_1^x$ as $k \to \infty$ P-a.s. and $t^*(y_k) \to t^*(x)$ as $k \to \infty$ it follows that $\limsup_{k\to\infty} \left\{T_1^{y_k} = t^*(y_k)\right\} \subset \left\{T_1^x = t^*(x)\,\right\}$ P-a.s.; thus (6) equals zero proving step 2. If $t < t^*(x)$ then $P(T_1^x = t) = 0$ and it is clear that

$$1_{\{T_1^{y_k} \leq t\}} \to 1_{\{T_1^x \leq t\}} \text{ as } k \to \infty \text{ P-a.s. .}$$

This, in conjunction with step 1, shows the first part of step 3. Suppose now that $t = t^*(x)$. Then we get

$$E\ |\ e^{-\alpha T_1^{y_k}} \rho_m^{N_k}(Z_1^{N_k y_k}) 1_{\{T_1^{y_k} \leq t_k\}} - e^{-\alpha T_1^x} \rho_m(Z_1^x) 1_{\{T_1^x < t^*(x)\}}\ |\ =$$

$$E\ \left|\left(e^{-\alpha T_1^{y_k}} \rho_m^{N_k}(Z_1^{N_k y_k}) 1_{\{T_1^{y_k} \leq t_k\}} - e^{-\alpha T_1^x} \rho_m(Z_1^x)\ \right) 1_{\{T_1^x < t^*(x)\}}\ \right| +$$

$$E\left|\left(e^{-\alpha T_1^{y_k}} \rho_m^{N_k}(Z_1^{N_k y_k}) 1_{\{T_1^{y_k} \leq t_k\}}\right) 1_{\{T_1^x = t^*(x)\}}\right|. \tag{7}$$

Since $T_1^{y_k} \to T_1^x$ as $k \to \infty$ P-a.s. and $t_k \to t^*(x)$ as $k \to \infty$ it is clear that $1_{\{T_1^{y_k} \leq t_k\}} \to 1$ as $k \to \infty$ P-a.s. on $\{T_1^x < t^*(x)\}$. Therefore as in

the proof of step 1 the first term of (7) goes to zero as $k\to\infty$. The second term is majorized by $a_1 P\left\{T_1^{y_k} \le t^*(y_k), T_1^x = t^*(x)\right\} \to 0$ as $k\to\infty$ from step 2, proving step 3. Let us show now step 4. Consider any sequence $y_k \to x$ and $N_k \to \infty$ as $k \to \infty$. We will show that

$$\text{i) } \limsup_{k\to\infty}\left(\inf_{0 \le t<t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k)\right) \le \inf_{0 \le t<t^*(x)} J(g,\rho_m)(t,x)$$

$$\text{ii) } \liminf_{k\to\infty}\left(\inf_{0 \le t<t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k)\right) \ge \inf_{0 \le t<t^*(x)} J(g,\rho_m)(t,x).$$

Let us show i) first. Given $\epsilon > 0$ take $t_\epsilon < t^*(x)$ such that t_ϵ is ϵ-optimal for the right hand side of i). By virtue of continuity of $t^*(\,.\,)$ we can get $k_0 \in \mathbb{N}$ such that $t_\epsilon < t^*(y_k)$ for $k \ge k_0$. It is clear then that for $k \ge k_0$

$$\left\{ \inf_{0 \le t<t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k) \right\} - \left\{ \inf_{0 \le t<t^*(x)} J(g,\rho_m)(t,x) \right\} \le$$

$$J^{N_k}(g^{N_k},\rho_m^{N_k})(t_\epsilon,y_k) - J(g,\rho_m)(t_\epsilon,x) + \epsilon \le$$

$$\left(g^{N_k}(\phi(t_\epsilon,y_k))e^{-\Lambda(t_\epsilon,y_k)} - g(\phi(t_\epsilon,x))e^{-\Lambda(t_\epsilon,x)} \right) +$$

$$\left(E_{y_k}^{N_k}\left(e^{-\alpha T_1^{N_k}} \rho_m^{N_k}(Z_1^{N_k}) 1_{\{T_1^{N_k} \le t_\epsilon\}}\right) - E_x(e^{-\alpha T_1}\rho_m(Z_1)1_{\{T_1 \le t_\epsilon\}})\right) + \epsilon.$$

By virtue of continuity of $\Lambda(\,.\,,\,.\,)$ and the assumptions on g^N and g the first term above goes to zero as $k \to \infty$. From step 1 we see that the second term also goes to zero as $k \to \infty$. Thus

$$\limsup_{k\to\infty} \left\{ \inf_{0 \le t<t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k)\right\} - \inf_{0 \le t<t^*(x)} J(g,\rho_m)(t,x) \le \epsilon$$

and since it holds for every $\epsilon > 0$, part i) is proved . Let us show ii) now. Given $\epsilon > 0$ we can find for each $k \in \mathbb{N}$, $s_k < t^*(y_k)$ which is ϵ-optimal for

$$\inf_{0 \le t<t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k).$$

Consider now any convergent subsequence

$$J^{N_{k_i}}(g^{N_{k_i}},\rho_m^{N_{k_i}})(s_{k_i},y_{k_i}) \text{ of } J^{N_k}(g^{N_k},\rho_m^{N_k})(s_k,y_k).$$

There is no loss of generality in assuming that s_{k_i} converges to some s in $R_+ \cup \{\infty\}$. Since $s_{k_i} < t^*(y_{k_i}) \to t^*(x)$ as $i \to \infty$ it is clear that $s \leq t^*(x)$. If $s < t^*(x)$ then

$$J(g,\rho_m)(s,x) - J^{N_{k_i}}(g^{N_{k_i}},\rho_m^{N_{k_i}})(s_{k_i},y_{k_i}) =$$

$$\left(g(\phi(s,x))e^{-\Lambda(s,x)} - g^{N_{k_i}}(\phi(s_{k_i},y_{k_i}))e^{-\Lambda(s_{k_i},y_{k_i})}\right) +$$

$$\left(E_x(e^{-\alpha T_1}\rho_m(Z_1)1_{\{T_1 \leq s\}}) - E_{y_{k_i}}^{N_{k_i}}(e^{-\alpha T_1^{N_{k_i}}}\rho_m^{N_{k_i}}(Z_1^{N_{k_i}})1_{\{T_1^{N_{k_i}} \leq s_{k_i}\}})\right)$$

which goes to zero as $i \to \infty$ from the assumptions on g^N and g, continuity of Λ and step 3. If $s = t^*(x)$ then from continuity of g on $\tilde{E}$ we have

$$\left(\lim_{t \to t^*(x)} J(g,\rho_m)(t,x)\right) - J^{N_{k_i}}(g^{N_{k_i}},\rho_m^{N_{k_i}})(s_{k_i},y_{k_i}) =$$

$$\left(g(\phi(t^*(x),x))e^{-\Lambda(t^*(x),x)} - g^{N_{k_i}}(\phi(s_{k_i},y_{k_i}))e^{-\Lambda(s_{k_i},y_{k_i})}\right) +$$

$$\left(E_x(e^{-\alpha T_1}\rho_m(Z_1)1_{\{T_1 < t^*(x)\}}) - E_{y_{k_i}}^{N_{k_i}}(e^{-\alpha T_1^{N_{k_i}}}\rho_m^{N_{k_i}}(Z_1^{N_{k_i}})1_{\{T_1^{N_{k_i}} \leq s_{k_i}\}})\right)$$

which goes to zero as $i \to \infty$ from the assumptions on g^N and g, continuity of Λ and step 3 . Then we can conclude that

$$\inf_{0 \leq t < t^*(x)} J(g,\rho_m)(t,x) \leq \lim_{i \to \infty} J^{N_{k_i}}(g^{N_{k_i}},\rho_m^{N_{k_i}})(s_{k_i},y_{k_i}).$$

Since it is true for any convergent subsequence of $J^{N_k}(g^{N_k},\rho_m^{N_k})(s_k,y_k)$, we get that

$$\inf_{0 \leq t < t^*(x)} J(g,\rho_m)(t,x) \leq \liminf_{k \to \infty} \left\{J^{N_k}(g^{N_k},\rho_m^{N_k})(s_k,y_k)\right\}$$

$$\leq \liminf_{k \to \infty} \left\{\inf_{0 \leq t < t^*(y_k)} J^{N_k}(g^{N_k},\rho_m^{N_k})(t,y_k)\right\} + \epsilon$$

proving step 4. From step 1.ii) and step 4 we get

$$\lim_{\substack{y\to x\\ N\to\infty}} \rho^N_{m+1}(y) = \lim_{\substack{y\to x\\ N\to\infty}} \left(\left\{ \inf_{0\le t<t^*(y)} J^N(g,\rho^N_m)(t,y) \right\} \wedge K^N\rho^N_m(y) \right) =$$

$$\left(\lim_{\substack{y\to x\\ N\to\infty}} \left\{ \inf_{0\le t<t^*(y)} J^N(g,\rho^N_m)(t,y) \right\} \right) \wedge \left(\lim_{\substack{y\to x\\ N\to\infty}} K^N\rho^N_m(y) \right) =$$

$$\left\{ \inf_{0\le t<t^*(x)} J(g,\rho_m)(t,x) \right\} \wedge K\rho_m(x) = \rho_{m+1}(x)$$

proving the Theorem. □

The next results show continuity of ρ and convergence of ρ^N to ρ.

Corollary 1 : ρ is continuous on $\tilde{E}$.

Proof : From Prop. 6 we have for any $x \in \tilde{E}$

$$0 \le \rho_m(x) - \rho(x) \le E_x(e^{-\alpha T_m}(g(Z_m) - \rho(Z_m))) \le 2a_1 E_x(e^{-\alpha T_m}).$$

The sequence $E_x(e^{-\alpha T_m})$ is obviously decreasing and converges to zero as $m \to \infty$. From Prop. 10, $E_{(.)}(e^{-\alpha T_m})$ is continuous on $\tilde{E}$. By virtue of Dini's Theorem (see [28] page 135) $E_x(e^{-\alpha T_m}) \to 0$ as $m \to \infty$ uniformly on compact sets of $\tilde{E}$. So it is clear that $\rho_m \to \rho$ as $m \to \infty$ uniformly on compact sets of $\tilde{E}$. From Theorem 1, ρ_m is continuous on $\tilde{E}$ which implies that ρ is continuous on $\tilde{E}$. □

Corollary 2 : $\rho^N \overset{N\to\infty}{\to} \rho$ uniformly on compact sets of $\tilde{E}$.

Proof : For any $x \in \tilde{E}$ we get from Prop. 6 , Prop. 10 and Theorem 1

$$\begin{aligned} |\rho(x) - \rho^N(y)| &\le |\rho(x) - \rho_m(x)| + |\rho_m(x) - \rho^N_m(y)| + \\ &\quad |\rho^N_m(y) - \rho^N(y)| \\ &\le 2a_1 E_x(e^{-\alpha T_m}) + |\rho_m(x) - \rho^N_m(y)| + 2a_1 E^N_y(e^{-\alpha T_m}) \\ &\to 4a_1 E_x(e^{-\alpha T_m}) \quad \text{as } y \to x \text{ and } N \to \infty . \end{aligned}$$

Taking the limit as $m \to \infty$ we obtain that $\rho^N(y) \to \rho(x)$ as $y \to x$ and $N \to \infty$ and from Proposition 11 the Corollary is proved. □

III. IMPULSE CONTROL PROBLEM

A. PRELIMINARIES

In the previous section we presented some characterization results and a numerical technique for the optimal stopping problem of a PDP. In this section we will apply these results to the impulse control problem. We characterize the value function by some recursive methods rather than by variational inequalities. From the general theory of impulse control of right processes (see [6]) such problems can be written as a sequence of optimal stopping problems and therefore by using the results of the previous section we can develop a numerical technique for computing optimal impulse controls for PDP's.

This section is organized in the following way. We keep the same notations, definitions and general assumptions as in subsection II, B. In subsection B we give the formulation of the problem. In subsection C we study the impulse control of PDP's under general conditions. We show that by iteration of the single-jump-or-intervention operator we obtain a sequence of functions converging to the value function of the problem. In subsection D a connection between the representation results of subsection C and the Bellman inequalities as defined by Yushkevich [17] is presented. In subsection E we present a numerical technique for the impulse control of PDP's. This technique consists of solving a sequence of one dimensional minimizations. We conclude that subsection by presenting a numerical example. The results of this section follow those obtained in [29].

B. PROBLEM FORMULATION

We shall now describe the construction of the canonical space which we shall use for the formulation of the impulsive control problem. Let $\{\Delta\}$ be a cemetery state and let $\hat{\Omega}$ be the space of functions $\hat{\omega} : R_+ \to E \cup \{\Delta\}$ which satisfies one of the following properties :

- $\hat{\omega} \in \Omega$; in this case we define $\eta(\hat{\omega}) := \infty$.
- for some $a \in R_+$ $\hat{\omega}_t$ is right continuous with left limit in $[0,a]$ and $\hat{\omega}_t = \Delta$ for $t \in (a,\infty)$; we define $\eta(\hat{\omega}) := a$.
- $\hat{\omega}_t = \Delta$ for every $t \in R_+$; in this case we denote $\hat{\omega}$ by Δ and define $\eta(\Delta) := 0$.

For $\hat{\omega} \in \hat{\Omega}$ define $\hat{x}_t(\hat{\omega}) := \hat{\omega}_t$, $\hat{\mathcal{F}}_t^0 := \sigma\{ \hat{x}_s ; s \leq t \}$, $\hat{\mathcal{F}}^0 := \hat{\mathcal{F}}_\infty$ and $\hat{\mathcal{F}}$ the universal completion of $\hat{\mathcal{F}}^0$. Let $(\hat{\Omega}_i, \hat{\mathcal{F}}_i)$ be a copy of the measurable space $(\hat{\Omega}, \hat{\mathcal{F}})$. Now define $\mathcal{W} := \prod_{i=1}^{\infty} \hat{\Omega}_i$, $\mathcal{G}^0 := \sigma\{ \prod_{i=1}^{\infty} \hat{\mathcal{F}}_i \}$ and $\mathcal{G}$ the universal completion of $\mathcal{G}^0$. We will denote the elements of $\mathcal{W}$ by $w = (\hat{\omega}_1, \hat{\omega}_2, \ldots)$. Let $\mathcal{W}_k := \prod_{i=1}^{k} \hat{\Omega}_i$, $\mathcal{G}_k^0 := \sigma\{ \prod_{i=1}^{k} \hat{\mathcal{F}}_i \}$, $\mathcal{G}_k$ the universal completion of $\mathcal{G}_k^0$ and $w_k : \mathcal{W} \to \mathcal{W}_k$ the projection map $w_k(w) = (\hat{\omega}_1, \ldots, \hat{\omega}_k)$. We now set

$$\tau_k(w) = \tau_k(w_k) := \sum_{i=1}^{k} \eta(\hat{\omega}_i) \qquad (\infty + \infty := \infty)$$

$$\tau_\infty(w) := \lim_{k \to \infty} \tau_k(w)$$

and define the path $(Y_t(w))_{t \geq 0}$ by

$$Y_0(w) := \hat{x}_0(\hat{\omega}_1)$$

$$Y_t(w) := \begin{cases} \hat{x}_{t-\tau_i(w)}(\hat{\omega}_{i+1}) & \text{if } \tau_i(w) < t \leq \tau_{i+1}(w) \\ \Delta & \text{if } t > \tau_\infty(w) \end{cases}$$

Let Γ be an analytic set of $E \times E$ and let $\Gamma(x)$ denote the x-section of Γ, i.e., $\Gamma(x) = \{y \in E: (x,y) \in \Gamma\}$. We impose the following conditions :

i) $\{ \Gamma(x) \}_{x \in E}$ is a family of non-empty sets of E.

ii) for every $x \in E$ and $y \in \Gamma(x)$, $\Gamma(y) \subset \Gamma(x)$.

We define the class of admissible strategies **S** as the set of sequences $\mathcal{S} := (S_n , R_n)_{n=1}^{\infty}$ which verify:

a) for every $n \geq 1$, $R_n(.)$: $\mathcal{W}_n \to E \cup \{\Delta\}$ is a universally measurable random variable such that

$$R_n(w_n) \begin{cases} = \Delta & \text{if } \tau_n(w_n) = \infty \\ \in \Gamma(\hat{x}_{\eta(\hat{\omega}_n)}(\hat{\omega}_n)) & \text{otherwise} \end{cases}$$

b) for every $n \geq 2$, $S_n(.,.)$: $\mathcal{W}_{n-1} \times \Omega \to R_+ \cup \{\infty\}$ is such that for $\omega \in \Omega$ fixed, $S_n(.,\omega)$ is a universally measurable function and for $w_{n-1} \in \mathcal{W}_{n-1}$ fixed, $S_n(w_{n-1}, .)$ is an $\mathcal{F}_t$-stopping time of a PDP (X_t) starting from

$R_{n-1}(w_{n-1})$. For n=1 , S_1 is an $\mathcal{F}_t$-stopping time of a PDP (X_t) starting from x.

c) for every $n \geq 2$, $S_n(w_{n-1}, \ .\ )$ is zero if there is $k \leq n-1$ such that $\eta(\hat{\omega}_k) = \infty$.

d) for every $n \geq 2$, S_n is strictly positive except for case c).

Remark : Note that given $w_{n-1} \in \mathcal{W}_{n-1}$, $(R_{n-1}(w_{n-1}), S_n(w_{n-1},.))$ defines a mapping $\mathcal{Y}$ from Ω into $\hat{\Omega}$ in the following way. If $R_{n-1}(w_{n-1}) = \Delta$ then $\mathcal{Y}(\omega) = \Delta \in \hat{\Omega}$ for all $\omega \in \Omega$. Consider now $R_{n-1}(w_{n-1}) \neq \Delta$. If $S_n(w_{n-1},\omega) = \infty$ then $\mathcal{Y}(\omega) = \omega \in \hat{\Omega}$; otherwise

$$\hat{x}_t(\mathcal{Y}(\omega)) = \begin{cases} x_t(\omega) & \text{if } t \leq S_n(w_{n-1},\omega) \\ \Delta & \text{if } t > S_n(w_{n-1},\omega) \end{cases}$$

and thus $\mathcal{Y}(\omega) \in \hat{\Omega}$. From the definition of $\mathcal{Y}$ we have $\eta(\mathcal{Y}(\omega)) = S_n(w_{n-1},\omega)$, $\forall\ \omega \in \Omega$.

Suppose $\mathcal{S} = (\ S_n\ ,\ R_n\ )_{n=1}^{\infty}$ satisfy the conditions a), b), c) and d) above. Then S_1 and P_x will induce a probability measure μ_x^1 on $(\hat{\Omega}_1, \hat{\mathcal{F}}_1)$. Given w_{n-1} we define $\mu_x^n(\{\Delta\};w_{n-1}) = 1$ if $\tau_{n-1}(w_{n-1}) = \infty$; otherwise $S_n(w_{n-1},.)$ and $P_{R_{n-1}(w_{n-1})}$ will induce a probability measure $\mu_x^n(.,w_{n-1})$ on $(\hat{\Omega}_n, \hat{\mathcal{F}}_n)$. The family (μ_x^n) defines a probability measure $P_x^{\mathcal{S}}$ on $(\mathcal{W},\mathcal{G})$. The last condition for $\mathcal{S}$ be admissible is the following :

e) $\tau_\infty = \infty$ $P_x^{\mathcal{S}}$ - a.s. .

Let $f : E \to R_+$ and $c : \Gamma \to R_+$ be bounded Borel functions satisfying the following conditions:

C1) $c(x,y) \geq c_0 > 0$ for some $c_0 \in R_+$ and every $(x,y) \in \Gamma$; $c(x,\Delta) = 0$, for any $x \in E$.

C2) for every $x \in E$, $y \in \Gamma(x)$ and $z \in \Gamma(y) \subset \Gamma(x)$, $c(x,y) + c(y,z) \geq c(x,z)$.

For each admissible strategy $\mathcal{S} \in \mathbf{S}$ we associate the following cost :

$$V^{\vartheta}(x) := E_x^{\vartheta}\Big(\int_0^{\tau_\infty} e^{-\alpha s} f(Y_s)ds \; + \; \sum_{i=1}^{\infty} e^{-\alpha \tau_i} c(Y_{\tau_i}, Y_{\tau_i}+) \Big)$$

where $\alpha > 0$ and the expectation is over $\mathcal{W}$ with probability measure P_x^{ϑ}. The value function of the impulsive control problem is :

$$\hat{\rho}'(x) := \inf_{\vartheta \in \mathbf{S}} V^{\vartheta}(x).$$

C. CHARACTERIZATION RESULTS

For $x \in E$, $0 \leq t < t^*(x)$ and v_1, v_2 in $B^*(E)$ we re-define the monotone operator J, K and define M, $\mathcal{A}$ in the following way:

a) $$J(v_1,v_2)(t,x) := E_x\Big(\int_0^{T_1 \wedge t} e^{-\alpha s} f(\phi(s,x))ds + e^{-\alpha t} v_1(\phi(t,x)) 1_{\{T_1 > t\}} + e^{-\alpha T_1} v_2(Z_1) 1_{\{T_1 \leq t\}} \Big)$$

b) $$Kv_2(x) := E_x\Big(\int_0^{T_1} e^{-\alpha s} f(\phi(s,x))ds + e^{-\alpha T_1} v_2(Z_1) \Big) \quad (= J(v_1,v_2)(\infty,x))$$

c) $$Mv_2(x) := \inf_{y \in \Gamma(x)} \{ c(x,y) + v_2(y) \}$$

d) $$\mathcal{A}v_2(x) := \inf_{\tau \in \mathcal{M}_{b\infty}} E_x\Big(\int_0^{\tau} e^{-\alpha s} f(X_s)ds + e^{-\alpha \tau} Mv_2(X_\tau) \Big) .$$

Recall that the definition of L is :

e) $$L(v_1,v_2)(x) := \{ \inf_{0 \leq t < t^*(x)} J(v_1,v_2)(t,x) \} \wedge Kv_2(x).$$

Since M maps $B^*(E)$ into $B^*(E)$ ([26], Prop. 7.47) we have from the results for the optimal stopping of PDP's (Propositions 1 and 3) that $\mathcal{A}$ also maps $B^*(E)$ into $B^*(E)$. Define

$$h(x) := E_x\Big(\int_0^{\infty} e^{-\alpha s} f(X_s)ds \Big)$$

which corresponds to the cost of "no intervention" strategy.

The following Proposition can be easily proved.

Proposition 12 : The function h is the smallest solution of the system

$$\begin{cases} v = Kv \\ v \geq 0 \quad , \quad v \in B(E) \end{cases}$$

Moreover if we define $h_{n+1} = Kh_n$, $h_0 = 0$ then $h_n \uparrow h$ as $n \to \infty$.

Since $\mathcal{A}$ is monotone, it is clear that $\mathcal{A}^n h$, $n = 0,1, \ldots$ is a decreasing sequence of functions in $B^*(E)$ (note that for $n = 1$, $\mathcal{A}h \leq h$ by definition of the operator $\mathcal{A}$). Therefore we have that the limit of $\mathcal{A}^n h$ as n goes to infinity exists and from Lemma 7.30(2) of [26] it is in $B^*(E)$. From the general results of Lepeltier and Marchal [6] on impulse control of right processes and the fact that PDP's are right processes we have that the following result holds:

Proposition 13 : $\hat{\rho}'$ is the biggest solution of the system

$$\begin{cases} v = \mathcal{A}v \\ v \in B^*(E) \end{cases}$$

Moreover $\hat{\rho}' = \lim_{n \to \infty} \mathcal{A}^n h$.

Outline of the proof : The second part of the statement follows from Propositions 22 and 23 of [6] (see also [29]). From the bounded convergence Theorem it is clear that

$$\hat{\rho}'(x) = \inf_n \mathcal{A}^{n+1} h(x) = \inf_n \inf_{\tau \in \mathcal{M}_\infty} E_x\left(\int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} M \mathcal{A}^n h(X_\tau) \right)$$

$$= \inf_{\tau \in \mathcal{M}_\infty} E_x\left(\int_0^\tau e^{-\alpha s} f(X_s) ds + e^{-\alpha \tau} M(\inf_n \mathcal{A}^n h)(X_\tau) \right) = \mathcal{A}\hat{\rho}'(x) .$$

Moreover if $v = \mathcal{A}v$, $v \in B^*(E)$ then clearly $v = \mathcal{A}v \leq h$ and thus using monotonicity of $\mathcal{A}$ again we have

$$v = \mathcal{A}^n v \leq \mathcal{A}^n h \xrightarrow{n \to \infty} \hat{\rho}'.$$

□

From the above representation we have that $\hat{\rho}'$ is the limit of a sequence of optimal stopping problems. Note also that the function $\mathcal{A}^n h$ can be regarded as the value function of the impulsive control problem which allows only n interventions. For $v \in B^*(E)$ define the operator

$\hat{L}(v)(x) := L(Mv,v)(x)$. This operator when applied to v can be seen as the value function of the single-jump-or-intervention problem with gain function v. The next proposition links the operator $\hat{L}$ with $\hat{\rho}'$.

Proposition 14 : The cost function $\hat{\rho}'$ is the biggest solution of the system

$$\begin{cases} v = \hat{L}(v) \\ v \leq h \quad , \quad v \in B^*(E) \end{cases} \tag{8}$$

Moreover if we define $\hat{\rho}'_{n+1} = \hat{L}(\hat{\rho}'_n)$, $\hat{\rho}'_0 = h$ then $\hat{\rho}'_n \downarrow \hat{\rho}'$ as $n \to \infty$.

Proof : By induction arguments and monotonicity of the operator $\hat{L}$, it is clear that the sequence of functions $\hat{\rho}'_n$ is decreasing in $B^*(E)$ (note that $\hat{\rho}'_1(x) = \hat{L}(h)(x) \leq Kh(x) = h(x)$ from Prop. 12). Therefore

$$\ell(x) := \lim_{n\to\infty} \hat{\rho}'_n(x)$$

exists and is in $B^*(E)$. From the same arguments as for the proof of the first part of Proposition 13 we get that ℓ is the biggest solution of system (8). It remains to show that $\ell = \hat{\rho}'$. From the optimal stopping results and the fact that $\hat{\rho}' = \mathcal{A}\hat{\rho}'$ we obtain that

$$\hat{\rho}' = L(M\hat{\rho}', \hat{\rho}') = \hat{L}(\hat{\rho}') \quad \text{(from optimal stopping)},$$

$$\hat{\rho}' \leq h \text{ (from } \hat{\rho}' = \mathcal{A}\hat{\rho}' \text{ and definition of the operator } \mathcal{A}).$$

So $\hat{\rho}'$ is a solution of (8) which implies that $\hat{\rho}' \leq \ell$ since ℓ is the biggest solution of (8). If we show that $\ell = \mathcal{A}\ell$ then from Proposition 13, $\ell \leq \hat{\rho}'$ and the result will be proved. So all we need to show is that $\ell = \mathcal{A}\ell$. From the theoretical results for optimal stopping we know that

$$\mathcal{A}\ell = \inf_n \ell_n \quad \text{where } \ell_{n+1} := L(M\ell, \ell_n) \ , \ \ell_0 := M\ell.$$

We will use induction on n to prove that $\ell \leq \ell_n$ for all $n \in \mathbb{N}$. For $n = 0$ we have, by virtue of $\ell = \hat{L}(\ell) = L(M\ell,\ell)$, that $\ell \leq M\ell = \ell_0$. Suppose now that $\ell \leq \ell_n$. Then

$$\ell = \hat{L}(\ell) = L(M\ell,\ell) \leq L(M\ell,\ell_n) = \ell_{n+1}$$

proving that $\ell \leq \ell_n$ for all $n \in \mathbb{N}$. Taking the infimum over n we obtain $\ell \leq \mathcal{A}\ell$. From the bounded convergence theorem we have

$$\mathcal{A}\ell(x) = \inf_{\tau \in \mathcal{M}_\infty} E_x\Big(\int_0^\tau e^{-\alpha s} f(X_s)ds + e^{-\alpha\tau} M(\inf_n \hat{\rho}'_n)(X_\tau) \Big) =$$

$$\inf_n \Big(\inf_{\tau \in \mathcal{M}_\infty} E_x\Big(\int_0^\tau e^{-\alpha s} f(X_s)ds + e^{-\alpha\tau} M\hat{\rho}'_n(X_\tau) \Big) \Big) = \inf_n \mathcal{A}\hat{\rho}'_n(x) .$$

We will show by induction that $\mathcal{A}\hat{\rho}'_n \leq \hat{\rho}'_{n+1}$ for all $n \in \mathbf{N}$. For $n = 0$ we get $\mathcal{A}\hat{\rho}'_0 = \mathcal{A}h \leq h$ (by definition of $\mathcal{A}$) and by virtue of the results for optimal stopping, $\mathcal{A}h = L(Mh,\mathcal{A}h)$. Thus $\mathcal{A}\hat{\rho}'_0 = \mathcal{A}h = L(Mh,\mathcal{A}h) \leq L(Mh,h) = \hat{L}(h) = \hat{\rho}_1'$. Suppose $\mathcal{A}\hat{\rho}'_{n-1} \leq \hat{\rho}'_n$. Then $\mathcal{A}\hat{\rho}'_n \leq \mathcal{A}\hat{\rho}'_{n-1}$ (since $\hat{\rho}'_m$, $m \in \mathbf{N}$ is decreasing) $\leq \hat{\rho}'_n$ and $\mathcal{A}\hat{\rho}'_n = L(M\hat{\rho}'_n,\mathcal{A}\hat{\rho}'_n) \leq L(M\hat{\rho}'_n,\hat{\rho}'_n) = \hat{L}(\hat{\rho}'_n) = \hat{\rho}'_{n+1}$ proving that $\mathcal{A}\hat{\rho}'_n \leq \hat{\rho}'_{n+1}$, $\forall\, n \in \mathbf{N}$. Hence $\mathcal{A}\ell = \inf_n \mathcal{A}\hat{\rho}'_n \leq \inf_n \hat{\rho}'_{n+1} = \ell$. □

It is interesting to note that the function $\hat{\rho}'_n$ can be understood as the value function of the impulse control problem of a PDP where only n jumps plus interventions are allowed and after that there are no further interventions.

Suppose now that $v_0 \in B^*(E)$ and $v_0 \geq h$. Define $v_{n+1} := \hat{L}(v_n)$. The following result, established in [29], shows that we only need to find an upper bound for h in order to get a sequence of functions converging to $\hat{\rho}'$.

Proposition 15 : $\lim_{n\to\infty} v_n(x) = \hat{\rho}'(x)$ for all $x \in E$.

In [30] chapter 9, Zabczyk studies the properties of the operator $\mathcal{A}$ when the process is Feller and there is continuity of the parameters c, f and of the operator M. Theorem 9.2 of [30] can be readily modified to show the following Proposition. Let f be bounded by a_3 and recall that $c(x,y) \geq c_0$ for all $(x,y) \in \Gamma$. Define $\upsilon := 1/(1 + (\alpha c_0/a_3))$.

Proposition 16 : For all $x \in E$ and $m \in \mathbf{N}$

$$0 \leq \mathcal{A}^{m-1}h(x) - \mathcal{A}^m h(x) \leq \upsilon^m \mathcal{A}^{m-1}h(x) .$$

Remark : From Prop. 16 it follows that

$$0 \leq \mathcal{A}^m h(x) - \hat{\rho}'(x) \leq \frac{\upsilon^{m+1}}{1-\upsilon} h(x) \leq \frac{\upsilon^{m+1}}{\alpha(1-\upsilon)} a_3 .$$

D. OPTIMALITY EQUATIONS

The purpose of this subsection is to show a connection between the solutions of $v = \hat{L}(v)$ and the optimality equations obtained in [17] when there is only impulse control. Define for $v \in B^*(E)$,

$$\mathcal{U}v(x) := f(x) + \lambda(x) \int_E (v(y) - v(x))\, Q(dy;x).$$

We have the following result:

Proposition 17 : Suppose that for every $x \in E$, $v \in B^*(E)$ satisfies

$$v(x) = \hat{L}(v)(x),\ v(x) \geq 0.$$

Then for all $x \in E$,

i) $v(x) \leq Mv(x)$ (9)

ii) $v(\phi(\,.\,,x)) : [0,t^*(x)) \to R$ is a Borel function, its derivative exists almost everywhere and

$$\int_0^t (\mathcal{U}v(\phi(s,x)) - \alpha v(\phi(s,x)))\, ds \geq v(x) - v(\phi(t,x)), \tag{10}$$

for all $t \in [0,t^*(x))$. If moreover for some $x \in E$, $Mv(\phi(t,x))$ is continuous on $[0,t^*(x))$ then for each open interval $(s_1,s_2) \in \mathfrak{R}_x^c$ (the complement of $\mathfrak{R}_x$), where $\mathfrak{R}_x := \{0 \leq t < t^*(x);\ v(\phi(t,x)) = Mv(\phi(t,x))\}$, the function $v(\phi(t,x))$ is absolutely continuous on $[s_1,s_2]$ ($[s_1,\infty)$ if $s_2 = \infty$) and

$$\mathcal{U}v(\phi(t,x)) - \alpha v(\phi(t,x)) + \frac{dv(\phi(t,x))}{dt} = 0 \tag{11}$$

almost everywhere on $[s_1,s_2]$.

Proof : Equation (9) is immediate from $v(x) = \hat{L}(v)(x)$. Equation (2) (modified to include α and f) yields for any $t \in [0,t^*(x))$

$$v(\phi(t,x)) = e^{\alpha t + \Lambda(t,x)} \Big\{ L(Mv,v)(t,x) - \int_0^t \Big(f(\phi(s,x)) + Qv(\phi(s,x))\lambda(\phi(s,x)) \Big) e^{-\alpha s - \Lambda(s,x)} ds \Big\} \tag{12}$$

where

$$L(Mv,v)(t,x) = \Big(\inf_{t \le s<t^*(x)} J(Mv,v)(s,x) \Big) \wedge Kv(x) .$$

Since $L(Mv,v)(t,x)$ is increasing on $[0,t^*(x))$ it is clear that it is a Borel function and by virtue of Theorem 2 on page 96 of [33], $\frac{dL(Mv,v)(t,x)}{dt}$ exists almost everywhere on $[0,t^*(x))$. From (12) it follows that $v(\phi(t,x))$ is Borel measurable, has derivatives almost everywhere and

$$\frac{dv(\phi(t,x))}{dt} + \mathcal{U}v(\phi(t,x)) - \alpha v(\phi(t,x)) =$$

$$e^{\alpha t+\Lambda(t,x)} \frac{dL(Mv,v)(t,x)}{dt} \ge 0 \qquad (13)$$

a.e. on $[0.t^*(x))$ where the last inequality is due to the fact that $L(Mv,v)(.,x)$ is increasing. It is clear from (13) that for $t \in [0,t^*(x))$

$$\int_0^t \Big(\mathcal{U}v(\phi(s,x)) - \alpha v(\phi(s,x)) \Big) ds \ge - \int_0^t \frac{dv(\phi(s,x))}{ds} ds$$

and to show (10) it remains to prove

$$\int_0^t \frac{dv(\phi(s,x))}{ds} ds \le v(\phi(t,x)) - v(x).$$

By applying again Theorem 2, page 96, of [33] on the increasing function

$$\mathcal{V}(t) := e^{\alpha t+\Lambda(t,x)} L(Mv,v)(t,x)$$

we get that it is differentiable a.e. on $[0.t^*(x))$, the derivative is Borel measurable and

$$\int_0^t \frac{d\mathcal{V}(s)}{ds} ds \le \mathcal{V}(t) - \mathcal{V}(0). \qquad (14)$$

Since the function

$$\mathcal{R}(t) := - e^{\alpha t+\Lambda(t,x)} \int_0^t \Big(f(\phi(s,x)) + Qv(\phi(s,x))\lambda(\phi(s,x)) \Big) e^{-\alpha s-\Lambda(s,x)} ds$$

is absolutely continuous on $[0,t^*(x))$ we have for any $t \in [0,t^*(x))$

$$\int_0^t \frac{d\mathcal{R}(s)}{ds}\, ds = \mathcal{R}(t) - \mathcal{R}(0). \tag{15}$$

Equations (12), (14) and (15) yields

$$v(\phi(t,x)) = \mathcal{V}(t) + \mathcal{R}(t) \geq \mathcal{V}(0) + \int_0^t \frac{d\mathcal{V}(s)}{ds}\, ds + \mathcal{R}(0) + \int_0^t \frac{d\mathcal{R}(s)}{ds}\, ds$$

$$= \mathcal{V}(0) + \mathcal{R}(0) + \int_0^t \frac{d\,(\,\mathcal{V}(s) + \mathcal{R}(s)\,)}{ds}\, ds$$

$$= \mathcal{V}(0) + \mathcal{R}(0) + \int_0^t \frac{dv(\phi(s,x))}{ds}\, ds$$

and since $\mathcal{V}(0) = L(Mv,v)(0,x) = v(x)$, $\mathcal{R}(0) = 0$, it is clear that

$$\int_0^t \frac{dv(\phi(s,x))}{ds}\, ds \leq v(\phi(t,x)) - v(x)$$

proving (10). Finally suppose $Mv(\phi(t,x))$ is continuous on $[0,t^*(x))$ for some $x \in E$. Then clearly $J(Mv,v)(t,x)$, $L(Mv,v)(t,x)$ and $v(\phi(t,x))$ are continuous on $[0,t^*(x))$. Also, $v(\phi(t,x)) < Mv(\phi(t,x))$ iff $L(Mv,v)(t,x) < J(Mv,v)(t,x)$. For any open interval $(s_1,s_2) \in \mathfrak{R}_X^c$ and every $t \in (s_1,s_2)$, $v(\phi(t,x)) < Mv(\phi(t,x))$ and thus $L(Mv,v)(t,x) < J(Mv,v)(t,x)$. Therefore $L(Mv,v)(t,x)$ must be constant on (s_1,s_2) and from continuity of $L(Mv,v)(t,x)$, it is constant on $[s_1,s_2]$ ($[s_1,\infty)$ if $s_2 = \infty$). By virtue of (12) and (13) we get that $v(\phi(t,x))$ is absolutely continuous on $[s_1,s_2]$ and (11) holds since

$$\frac{dL(Mv,v)(t,x)}{dt} = 0. \qquad \square$$

E. DISCRETIZATION RESULTS

1. ASSUMPTIONS

In this subsection we extend the definition of the sets $\Gamma(y)$ to the points y in ∂^*E and assume that $\Gamma(y)$ is non-empty in $\tilde{E}$ for $y \in \partial^*E$. Conditions C1 and C2 of subsection III, B are modified replacing E by $\tilde{E}$. We extend the domain of the operator $\mathcal{A}$ to $B^*(\tilde{E})$ in the following way: for $v \in B(\tilde{E})$ and $x \in E$, $\mathcal{A}v(x)$ is defined as before and for $x \in \partial^*E$, $\mathcal{A}v(x) = Mv(x) \wedge E_x(\mathcal{A}v(Z_1))$ where the process jumps instantaneously when it starts from any point in ∂^*E. Note that this definition is consistent since $Z_1 \in E$ and we first evaluate $\mathcal{A}v(x)$ for $x \in E$. We can easily show that Propositions 12 and 16 still hold replacing E by $\tilde{E}$. Define for $x \in \tilde{E}$, $\hat{\rho}(x) := \lim_{n\to\infty} \mathcal{A}^n h(x)$. Note that $\hat{\rho}'(x) = \hat{\rho}(x)$ for $x \in E$ but for $x \in \partial^*E$, $\hat{\rho}(x)$ may be less than $\hat{\rho}'(x)$ $(= E_x(\hat{\rho}'(Z_1)))$.

Let $2^{\tilde{E}}$ denote the collection of all non-empty compact sets of $\tilde{E}$. The Hausdorff metric $d(.,.)$ in $2^{\tilde{E}}$ is defined in the following way (see Bertsekas-Shreve [26] appendix C): for A, B in $2^{\tilde{E}}$ and x in $\tilde{E}$,

$$d(x,A) := \min_{a \in A} |x-a|, \quad d(A,B) := \max \{ \max_{a \in A} d(a,B), \max_{b \in B} d(b,A) \}.$$

For $x \in \tilde{E}$ let $\bar{\Gamma}(x)$ be the closure of $\Gamma(x)$ in $\tilde{E}$. In addition to Assumptions 1, 2, 3 and 5 of subsection II, E, 2 we impose that:

6) $\bar{\Gamma}(.)$ is a continuous set-valued mapping from $\tilde{E}$ into $2^{\tilde{E}}$ with respect to the Hausdorff metric in $2^{\tilde{E}}$.

7) $c(.,.) : \tilde{E} \times \tilde{E} \to R_+$ is bounded (by a_4) and continuous.

8) $f(.) : E \to R_+$ is bounded (by a_3) and continuous.

2. AUXILIARY RESULTS

For $v \in C(\tilde{E})$ and $x \in \tilde{E}$ define

$$\bar{M}v(x) := \min_{z \in \bar{\Gamma}(x)} \{ c(x,z) + v(z) \}.$$

From our assumptions it is clear that $Mv(x) = \bar{M}v(x)$. From Proposition C.3 (appendix C) of Bertsekas-Shreve [26] and our assumptions the following result can be proved (cf. [29]):

Proposition 18 : M maps $C(\tilde{E})$ into $C(\tilde{E})$.

Let $D := \{z_1, z_2, \ldots\}$ be the countable dense set in E seen in subsection II, E, 3 and $D^N := \{z_1, \ldots, z_N\}$. Define the sets $\{ \tilde{A}_i^N \}_{i=1}^N$ as :

$$\tilde{B}_i^N := \{z \in \tilde{E};\ |z - z_i| \leq |z - z_j| \text{ for every } j = 1,\ldots, N\}, \quad i = 1,\ldots,N$$

$$\tilde{A}_1^N := \tilde{B}_1^N \quad , \quad \tilde{A}_i^N := \tilde{B}_i^N - \bigcup_{j=1}^{i-1} \tilde{A}_j^N \qquad i = 2,\ldots,N.$$

So clearly $A_i^N = \tilde{A}_i^N - \partial^* E$. For each $N \in \mathbb{N}$ and $x \in \tilde{E}$ define

$$\Gamma^N(x) := \{ z_i \in D^N ;\ \tilde{A}_i^N \cap \tilde{\Gamma}(x) \neq \emptyset \} \tag{16}$$

and for $v^N \in B^*(\tilde{E})$

$$M^N v^N(x) := \min_{y \in \Gamma^N(x)} \{ c(x,y) + v^N(y) \}.$$

With these definitions we have the following Proposition :

Proposition 19 : Suppose $v \in C(\tilde{E})$, $v^N \in B^*(\tilde{E})$ and $v^N \overset{N \to \infty}{\to} v$ uniformly on compact sets of $\tilde{E}$. Then $M^N v^N \overset{N \to \infty}{\to} Mv$ uniformly on compact sets of $\tilde{E}$.

Proof : For arbitrary $x \in \tilde{E}$ consider any sequence $y_k \to x$ and $N_k \to \infty$ as $k \to \infty$. For some $\varphi_x \in \tilde{\Gamma}(x)$, $\tilde{M}v(x) = c(x,\varphi_x) + v(\varphi_x)$ and, from continuity of $\tilde{\Gamma}$, we can find $\varphi_k \in \tilde{\Gamma}(y_k)$ such that $\varphi_k \to \varphi_x$ as $k \to \infty$. Define for each $k \in \mathbb{N}$,

$$\varphi_k^{N_k} := z_j \text{ where } j \in \{1,\ldots,N_k\} \text{ is such that } \varphi_k \in \tilde{A}_j^{N_k}.$$

From Proposition 9 it follows that $\varphi_k^{N_k} \to \varphi_x$ as $k \to \infty$. From the above construction it is clear that $\varphi_k^{N_k} \in \Gamma^{N_k}(y_k)$. Thus

$$M^{N_k} v^{N_k}(y_k) \leq c(y_k, \varphi_k^{N_k}) + v^{N_k}(\varphi_k^{N_k}) \to c(x,\varphi_x) + v(\varphi_x) = \tilde{M}v(x)$$

as $k \to \infty$ where we have used above continuity of c on $\tilde{E} \times \tilde{E}$, continuity of v on $\tilde{E}$ and uniform convergence of v^N to v on compact sets of $\tilde{E}$ (see Proposition 11). Therefore $\limsup_{k \to \infty} M^{N_k} v^{N_k}(y_k) \leq \tilde{M}v(x)$. Consider now $\varphi_k^{N_k} \in \Gamma^{N_k}(y_k)$ such that

$$M^{N_k}v^{N_k}(y_k) = c(y_k,\varphi_k^{N_k}) + v^{N_k}(\varphi_k^{N_k}) .$$

For each $\varphi_k^{N_k}$ we have from the definition of the set $\Gamma^{N_k}(y_k)$ (see Eq. (16)) that

$$\varphi_k^{N_k} = z_j \text{ where } j \in \{1,...,N_k\} \text{ and } \bar{\Gamma}(y_k) \cap \tilde{A}_j^{N_k} \neq \emptyset.$$

Thus associated to each $\varphi_k^{N_k}$ there is at least one $\varphi_k \in \bar{\Gamma}(y_k)$ such that $\varphi_k \in \tilde{A}_j^{N_k}$. From continuity of $\bar{\Gamma}$ we can find a subsequence y_{k_i} of y_k and φ_{k_i} of φ_k such that $\varphi_{k_i} \overset{i \to \infty}{\to} \varphi_x \in \bar{\Gamma}(x)$ and

$$\begin{aligned}\liminf_{k \to \infty} M^{N_k}v^{N_k}(y_k) &= \liminf_{k \to \infty} \left(c(y_k,\varphi_k^{N_k}) + v^{N_k}(\varphi_k^{N_k}) \right) \\ &= \lim_{i \to \infty} \left(c(y_{k_i},\varphi_{k_i}^{N_{k_i}}) + v^{N_{k_i}}(\varphi_{k_i}^{N_{k_i}}) \right) .\end{aligned}$$

Then again by virtue of Proposition 9 we get $\varphi_{k_i}^{N_{k_i}} \to \varphi_x$ as $i \to \infty$ and as before due to continuity of c on $\tilde{E} \times \tilde{E}$, continuity of v on $\tilde{E}$ and uniform convergence of v^N to v on compact sets of $\tilde{E}$ we obtain that

$$\begin{aligned}\liminf_{k \to \infty} M^{N_k}v^{N_k}(y_k) &= \lim_{i \to \infty} \left(c(y_{k_i},\varphi_{k_i}^{N_{k_i}}) + v^{N_{k_i}}(\varphi_{k_i}^{N_{k_i}}) \right) \\ &= c(x,\varphi_x) + v(\varphi_x) \geq \bar{M}v(x) .\end{aligned}$$

So $\limsup_{k \to \infty} M^{N_k}v^{N_k}(y_k) \leq \bar{M}v(x) \leq \liminf_{k \to \infty} M^{N_k}v^{N_k}(y_k)$ and hence

$$M^{N_k}v^{N_k}(y_k) \overset{k \to \infty}{\to} \bar{M}v(x) = Mv(x).$$

Since it holds for every sequence $y_k \to x$, $N_k \to \infty$ as $k \to \infty$ and x is arbitrary in $\tilde{E}$ we get that $M^N v^N(y) \to Mv(x)$ as $y \to x$ and $N \to \infty$ for all $x \in \tilde{E}$. The result follows from Proposition 11 and continuity of Mv on $\tilde{E}$ (Proposition 18). □

3. CONVERGENCE RESULTS

We will consider the impulse control problem for the discretized PDP (X_t^N) as defined in subsection II, E, 3 with the sets $\{\Gamma^N(y)\}_{y \in \tilde{E}}$ as defined in (16) above. We will use the superscript N to distinguish the discretized problem from the original one. The following Proposition can be easily proved:

Proposition 20 : $h \in C(\tilde{E})$ and $h^N \to h$ as $N \to \infty$ uniformly on compact sets of $\tilde{E}$.

The next Proposition is important to show continuity of $\hat{\rho}$ on $\tilde{E}$.

Proposition 21 : For every $n \in \mathbb{N}$, $\mathcal{A}^n h \in C(\tilde{E})$.

Proof : We use induction on n to show the result. For $n = 0$ it is immediate from the previous Proposition. Suppose it holds for n, that is, $\mathcal{A}^n h \in C(\tilde{E})$. Then from Proposition 18, $M\mathcal{A}^n h \in C(\tilde{E})$. Since $\mathcal{A}^{n+1} h$ is the value function of an optimal stopping problem with gain function $M\mathcal{A}^n h$ we obtain by virtue of Corollary 1 that $\mathcal{A}^{n+1} h \in C(\tilde{E})$ proving the Proposition . □

From uniform convergence of $\mathcal{A}^n h$ to $\hat{\rho}$ (Proposition 16) and continuity of $\mathcal{A}^n h$ on $\tilde{E}$ the next Corollary is immediate.

Corollary 3 : $\hat{\rho} \in C(\tilde{E})$.

The next Proposition is needed to show convergence of $\hat{\rho}^N$ to $\hat{\rho}$.

Proposition 22 : For every $n \in \mathbb{N}$, $(\mathcal{A}^N)^n h^N \to \mathcal{A}^n h$ as $N \to \infty$ uniformly on compact sets of $\tilde{E}$.

Proof : We again use induction on n . For $n = 0$ the result is clear from Prop. 20. Suppose it holds for n, that is, $(\mathcal{A}^N)^n h^N \overset{N}{\to} \mathcal{A}^n h$ uniformly on compact sets of $\tilde{E}$. Since $\mathcal{A}^n h \in C(\tilde{E})$ (Prop. 21) and $(\mathcal{A}^N)^n h^N \in B^*(\tilde{E})$ we have from Proposition 19 that $M^N(\mathcal{A}^N)^n h^N \overset{N}{\to} M\mathcal{A}^n h$ uniformly on compact sets of $\tilde{E}$. As mentioned before $\mathcal{A}^{n+1} h$ ($(\mathcal{A}^N)^{n+1} h^N$ respectively) is the value function of an optimal stopping problem of the original (discretized) process with gain function $M\mathcal{A}^n h$ ($M^N(\mathcal{A}^N)^n h^N$). By virtue of Corollary 2

$$(\mathcal{A}^N)^{n+1} h^N \overset{N}{\to} \mathcal{A}^{n+1} h$$

uniformly on compact sets of $\tilde{E}$ proving the Proposition. □

Noting that the constant v seen in Prop. 16 is the same for both the original problem and the discretized problem we obtain the following result:

Corollary 4 : $\hat{\rho}^N \overset{N\to\infty}{\to} \hat{\rho}$ uniformly on compact sets of $\tilde{E}$.

Proof : $|\hat{\rho} - \hat{\rho}^N| \leq (\mathcal{A}^n h - \hat{\rho}) + |\mathcal{A}^n h - (\mathcal{A}^N)^n h^N| + ((\mathcal{A}^N)^n h^N - \hat{\rho}^N)$

$$\leq 2\frac{v^{n+1}a_3}{(1-v)\alpha} + |\mathcal{A}^n h - (\mathcal{A}^N)^n h^N| \overset{N\to\infty}{\to} 2\frac{v^{n+1}a_3}{(1-v)\alpha}$$

uniformly on compact sets of $\tilde{E}$ by virtue of Proposition 22. Taking the limit as $n \to \infty$ we obtain the desired result. □

4. NUMERICAL EXAMPLE

Let us consider the following example of preventive maintenance of complex systems to illustrate our method. This example is similar to the one presented in [31] which was solved for the long run cost using the theory of generalized Markovian decision processes. It was also solved via linear programming [32] using a state space and time discretization. We consider a system with two units which, once operating, are subject to random failures. These units operate separately but they share one repair facility. It means that a queuing situation for the repair facility may arise. We assume that each unit has a fixed maintenance/repair time r_i, a failure rate $\lambda_i(s)$ and a return rate $C_i(s)$ where i=1,2 and s is the age of unit i. Upon unit failure there is an immediate cost $B_i(s)$ plus the repair cost $M_i(s)$. If a unit is taken out of operation preventively and the repair facility is available, only repair costs $M_i(s)$ have to be paid; but if the repair facility is engaged then a stopping cost $S_i(s)$ is incurred plus the repair cost $M_i(s)$ (where again i = 1,2 and s is the age of unit i). We also assume that when the age of unit i reaches t_i^* then unit i is taken out of operation with no extra costs incurred but $M_i(s)$. If an unit is out of order and the repair facility is available then it is immediately sent to be repaired.

The above model can be seen as a PDP with a 15-components state space. The three main state space components are :

$E_1 = [0,t_1^*)x[0,t_2^*)$ where units 1 and 2 are working

$E_2 = [0,t_1^*)x[0,r_2)$ where unit 1 is working and unit 2 is under repair

$E_3 = [0,r_1)x[0,t_2^*)$ where unit 1 is under repair and unit 2 is working

The other components are mainly cemetery states and are not important for the dynamic of the system. The decision maker can at any time send unit 1 or 2 to maintenance when the process is in E_1 and stop unit 1 (2 respectively) when the process is E_2 (E_3). We want to minimize the total

discounted cost. The parameters considered are shown in Table I:

parameters	unit 1	unit 2
Barrier t_i^*	20	10
Maintenance/repair time r_i	2	1
Failure rate $\lambda_i(s)$	0.05	$0.13s^{0.3}$
Failure cost $B_i(s)$	9 + 0.55s	2.20
Maintenance/repair costs $M_i(s)$	$2 + 0.01s^2$	0.5 + 0.35s
Stopping costs $S_i(s)$	0.5	0.1
Rate of return $C_i(s)$	8.5 - 0.55s	8.1 - 0.58s

Table I: Parameters of the numerical simulation

Three values for the discount factor α were considered, $\alpha = 0.5$, 0.25 and 0.1. Figure I shows the results obtained.

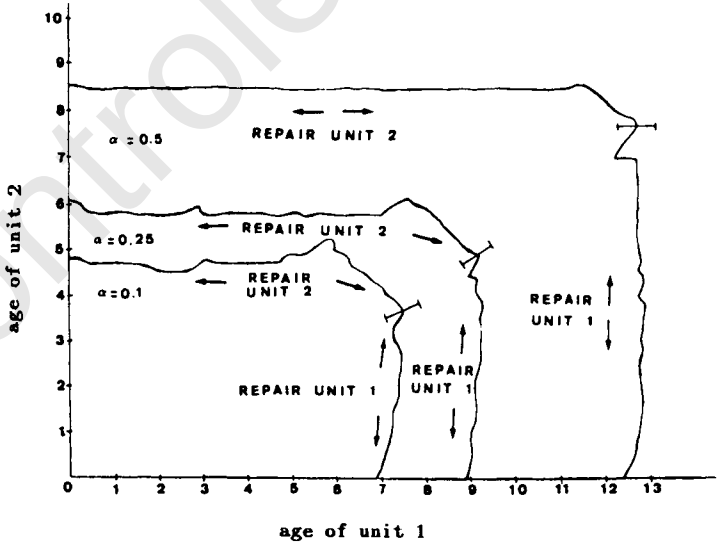


Fig. I. Intervention boundary for the numerical example

IV. LONG RUN AVERAGE IMPULSE CONTROL PROBLEM

A. PRELIMINARIES

In this section we present some characterization results and optimality equations for the long run average impulse control of a PDP. First we consider in subsection B the case when there are no restrictions to where the system can be sent to after an intervention and the cost for impulses from x to y is in a separated form c(x)+d(y). These assumptions will imply that we can concentrate on the class of strategies which always return to the same point after an intervention and use the same stopping time as intervention time (thus the process with intervention will be a regenerative process). These hypothesis simplify considerably the problem and an iterative technique which consists of solving a sequence of optimal stopping problems is derived. In subsection C we consider the more general case where constrains on the location of the process after an intervention are imposed and the cost per intervention may not be in a separated form. An optimality equation for this case is derived. The results of subsections B and C below follow those derived in [34] and [35] respectively. Convergence of discretization methods similar to those seen in the previous sections can also be obtained for these cases and the reader is referred to [34], [35], [36].

The construction of the admissible strategies and the assumptions for the impulse control problem are defined as in subsection III, B. Associated to an admissible strategy $\mathcal{S} \in \mathbf{S}$ we define the long run average cost as

$$V^{\mathcal{S}}(x) := \liminf_{t\to\infty} \frac{E_x^{\mathcal{S}}(G_t)}{t}$$

where

$$G_t := \int_0^t f(Y_s)ds + \sum_{i=0}^{\infty} 1_{\{\tau_i \le t\}} c(Y_{\tau_i}, Y_{\tau_i}+) .$$

We re-define $\hat{\rho}'$ as the payoff function of the long run average impulse control problem in the following way:

$$\hat{\rho}'(x) := \inf_{\mathcal{S}\in\mathbf{S}} V^{\mathcal{S}}(x).$$

B. THE CASE WITH NO RESTRICTIONS AFTER IMPULSE

1. ASSUMPTIONS AND AUXILIARY RESULTS

The assumptions on the PDP are the same as in subsection II, B. We assume that $\Gamma(x) = E$ for all $x \in E$, $c(x,y) = c(x) + d(y)$ for any $x,y \in E$, where c is a strictly positive function on $B^c(E)$ (thus there exists $c_0 > 0$ such that $c(x) \geq c_0$ for all $x \in E$) and d is a positive function on $B(E)$. Besides these assumptions, we also consider the following ones. Let $A_1 := \{ x \in E; t^*(x) < \infty \}$ and $A_2 := \{ x \in E; t^*(x) = \infty \}$.

Assumption A1 : There exist $\lambda_{max} \geq 0$, $\lambda_{min} > 0$, $t^*_{max} > 0$ such that
i) $\lambda(x) \leq \lambda_{max}$ for all $x \in E$
ii) $\lambda(x) > \lambda_{min}$ for all $x \in A_2$ and $t^*(x) < t^*_{max}$ for all $x \in A_1$

Assumption A2 (Doeblin condition for the embedded Markov chain) : There exist a finite valued non zero measure φ, an integer $m^* \geq 1$ and positive ϵ^* such that for all $A \in \mathcal{E}$ and $x \in E$,

$$\varphi(A) \leq \epsilon^* \Rightarrow P_x(Z_{m^*} \in A) \leq 1 - \epsilon^*$$

Note that Assumption A1.ii) implies that for some $a_0 > 0$, $E_x(T_m) \leq m a_0$ for all $x \in E$ and $m = 0,1,2,\ldots$ .We assume that f is a positive function on $B(E)$ and, to avoid trivialities, $\|f\| > 0$ (the sup norm). Define

$$\beta^* := \inf_{x \in E} \inf_{\tau \in \mathcal{M}_\infty} \frac{E_x\left(\int_0^\tau f(X_s)ds + c(X_\tau) \right) + d(x)}{E_x(\tau)} \qquad (17)$$

where if $E_x(\tau) = \infty$, then

$$\frac{E_x\left(\int_0^\tau f(X_s)ds + c(X_\tau) \right) + d(x)}{E_x(\tau)} :=$$

$$\liminf_{t \to \infty} \frac{E_x\left(\int_0^{\tau \wedge t} f(X_s)ds + c(X_{\tau \wedge t}) \right) + d(x)}{E_x(\tau \wedge t)} . \qquad (18)$$

By using the general results of optimal stopping of PDP's obtained by Gugerli [24], the first part of Theorem 1 (section 2) of [12] can be modified to show the following result:

Proposition 23 : $\hat{\rho}'(x) = \beta^*$ for all $x \in E$.

Therefore in order to solve the long run average impulse control for PDP's we have to find the solutions of (17). For $m = 1,2,\ldots$, $x \in E$ and $\beta \in R_+$ define

$$\psi_m(\beta,x) := \inf_{\tau \in \mathcal{M}_\infty} E_x\Big(\int_0^{\tau \wedge T_m} (f(X_s) - \beta) \, ds + c(X_{\tau \wedge T_m}) \Big)$$

$$l_m(\beta) := \inf_{x \in E} \Big(\psi_m(\beta,x) + d(x) \Big)$$

$$\hat{\beta} := \sup \{ \beta \in R_+; \; l_m(\beta) \geq 0 \text{ for all } m=1,2, \ldots \}.$$

We have the following results.

Proposition 24 : There exists $\beta \in R_+$ such that $l_m(\beta) \geq 0$ for all $m=1,2,\ldots$ and $\beta = \hat{\beta}$.

Proof : It is easy to see that $0 \in \{\beta \in R_+; \; l_m(\beta) \geq 0 \text{ for all } m=1,2, \ldots\} \neq \emptyset$ since f, c and d are positive. Fixing $x' \in E$ and noting that $0 < E_{x'}(T_1) \leq a_0$ we have that for all $\beta > \| f \| + \dfrac{\| c \| + \| d \|}{E_{x'}(T_1)}$, $l_1(\beta) < 0$ and thus

$$\{\beta \in R_+; \; l_m(\beta) \geq 0 \text{ for all } m=1,2,\ldots\} \subset [0, \| f \| + \frac{\| c \| + \| d \|}{E_{x'}(T_1)}].$$

Let $\hat{\beta}_k \in \{\beta \in R_+; \; l_m(\beta) \geq 0 \text{ for all } m=1,2, \ldots\}$ be an increasing sequence converging to $\hat{\beta} < \infty$ as $k \to \infty$. Then it is easy to see from Assumption A1 and the bounded convergence theorem that

$$l_m(\hat{\beta}) = l_m(\sup_k \hat{\beta}_k) = \inf_{x \in E} \inf_{\tau \in \mathcal{M}_\infty} \Big(\inf_k E_x\Big(\int_0^{\tau \wedge T_m} (f(X_s) - \hat{\beta}_k) ds + c(X_{\tau \wedge T_m}) \Big) + d(x) \Big) = \inf_k \inf_{x \in E} \inf_{\tau \in \mathcal{M}_\infty} \Big(E_x\Big(\int_0^{\tau \wedge T_m} (f(X_s) - \hat{\beta}_k) ds + c(X_{\tau \wedge T_m}) \Big) + d(x) \Big) = \inf_k l_m(\hat{\beta}_k) \geq 0$$

for all $m = 1,2, \ldots$ which proves the desired result. □

Proposition 25 : $\hat{\beta} = \beta^*$.

Proof : For any $x \in E$ and $\tau \in \mathcal{M}_\infty$ such that $0 < E_x(\tau) < \infty$,

$$\hat{\beta} \leq \frac{E_x\left(\int_0^{\tau \wedge T_m} f(X_s)ds + c(X_{\tau \wedge T_m}) \right) + d(x)}{E_x(\tau \wedge T_m)} \quad \text{for all } m = 1,2, \ldots .$$

Letting $m \to \infty$ we have by virtue of the bounded convergence theorem and $T_m \to \infty$ P_x a.s. that

$$\hat{\beta} \leq \frac{E_x\left(\int_0^{\tau} f(X_s)ds + c(X_\tau) \right) + d(x)}{E_x(\tau)} .$$

From the definition for the case $E_x(\tau) = \infty$ (see (18)) it is clear that $\hat{\beta} \leq \beta^*$. On the other hand for all $m = 1,2, \ldots$

$$l_m(\beta^*) = \inf_{x \in E} \left\{ \left\{ c(x) + d(x) \right\} \wedge \left\{ \inf_{\substack{\tau \in \mathcal{M}_\infty \\ \tau > 0}} \left(E_x(\tau \wedge T_m) \right. \right. \right.$$

$$\left. \left. \left. \left(\frac{E_x\left(\int_0^{\tau \wedge T_m} f(X_s)ds + c(X_{\tau \wedge T_m}) \right) + d(x)}{E_x(\tau \wedge T_m)} - \beta^* \right) \right) \right\} \right\} \geq 0$$

and therefore $\beta^* \leq \hat{\beta}$ which completes the proof. □

A non-empty set $A \in E$ is called invariant for the embedded Markov chain (PDP respectively) if $P_x(Z_m \in A) = 1$ for all $x \in A$ and $m=1,2,\ldots$ ($P_x(X_t \in A)$ for all $x \in A$ and $t \in R_+$) and it is minimal if it does not contain another invariant set B such that $\varphi(B) < \varphi(A)$. Assumption A2 implies the next two Lemmas (cf. [37], pp 191-215).

Lemma 1 : There exists a maximal sequence of disjoint minimal invariant sets $E_1, \ldots, E_r$ for the embedded Markov chain (Z_m). Moreover there is a decomposition $C_{i1},\ldots,C_{id_i}$, $C_{ik} \cap C_{ij} = \emptyset$ for $k \neq j$, of $E_i = \bigcup_{j=1}^{d_i} C_{ij}$

and a family of measures π_{ij} on E, constants $b_1 \in R_+$, $\nu_1 \in [0,1)$ such that for all $i = 1,...,r$, $j = 1,...,d_i$, $k = 0,...,d_i - 1$, $x \in C_{ij}$ and $A \in E$,

$$| P_x(Z_{nd_i+k} \in A) - \pi_{im}(A) | \leq b_1 \nu_1^n \tag{19}$$

where $m = j + k \pmod{d_i}$. Moreover the measures

$$\pi_i(.) := \frac{1}{d_i} \sum_{k=1}^{d_i} \pi_{ik}(.), \; i = 1,..., r$$

are invariant for the Markov chain and any invariant measure for the Markov chain belongs to the convex hull generated by π_i.

Lemma 2 : Define $F := E - \bigcup_{i=1}^{r} E_i$. Then there exists $b_2 > 0$ and $\nu_2 \in [0,1)$ such that for all $x \in E$ and $m = 0,1,2,...$,

$$P_x(Z_m \in F) \leq b_2 \nu_2^m. \tag{20}$$

From Assumption A1 and the equivalence results in [38] we have that there exists a one to one mapping between the invariant measures for the PDP and the invariant measures for the embedded Markov chain. Thus denoting by μ_i the invariant measure for the PDP associated to π_i, $i = 1,...,r$, we have from Lemma 1 that any invariant measure for the PDP belongs to the convex hull generated by μ_i. From [38] we know that for $h \in B(E)$,

$$\int_E h(y)\, \mu_i(dy) = \frac{\displaystyle\int_E \int_0^{t^*(x)} h(\phi(t,x)) e^{-\Lambda(t,x)} dt \; \pi_i(dx)}{\displaystyle\int_E \int_0^{t^*(x)} e^{-\Lambda(t,x)} dt \; \pi_i(dx)}. \tag{21}$$

Remark : Although Assumption A2 implies (19) and (20), we cannot say that $P_x(X_t \in A)$ will converge as $t \to \infty$. Indeed consider

$$E = [0,1),\ \lambda = 0,\ Q(\{0\};1) = 1 \text{ and } \phi(t,x) = x + t.$$

Then $\pi(\{0\}) = 1 = P_x(Z_m=0)$ for all $m = 1,2, ...$ and all $x \in E$. However it is easy to see from (21) that $\mu([0,\zeta)) = \zeta$ for $\zeta \in [0,1)$ is the invariant measure for the PDP but $P_x(X_t \in [0,\zeta)) = 0$ or 1 does not converge to ζ as $t \to \infty$. Thus we can conclude that the Doeblin condition for the embedded Markov chain does not imply the Doeblin condition (as in [37],

p. 256) for the PDP. The reverse in general is not true either but if for all $t \in R_+$,

$$\lim_{m\to\infty} P_x(T_m \le t) = 0 \text{ uniformly in } E$$

then we can show that the Doeblin condition for the PDP implies the Doeblin condition for the embedded Markov chain. In most practical applications this result will hold. Another reason to use the latter is because when we look at the discretization technique (see [34]) it is easier and more general to formulate the assumptions in terms of the embedded Markov chain rather than the PDP itself.

Define for $i = 1,\ldots,r$,

$$\tilde{E}_i := \{x \in E;\ x = \phi(t,z), \text{ some } z \in E_i, \text{ some } t \in [0,t^*(z))\} \cup \{x \in \partial E_i;\ \phi(t,x) \in E_i \text{ for all } t \in (0,\epsilon), \text{ some } \epsilon > 0\}$$

$$\tilde{F} := E - \bigcup_{i=1}^{r} \tilde{E}_i, \quad \tilde{f}_i := \int_E f(y)\, \mu_i(dy)$$

$$\tilde{f}(x) := \sum_{i=1}^{r} \tilde{f}_i\, 1_{\tilde{E}_i}(x) + \| f \|\, 1_{\tilde{F}}(x).$$

The next two Propositions are consequences of Lemmas 1 and 2, and the proofs can be found in [34].

Proposition 26 : The sets $\tilde{E}_i$, $i=1,\ldots,r$ are invariant for the PDP and $\tilde{E}_i \cap \tilde{E}_j = \emptyset$ for $i \neq j$.

Proposition 27 : The function

$$v(x) := \sum_{m=0}^{\infty} E_x\Big(\int_{T_m}^{T_{m+1}} (\tilde{f}(X_t) - f(X_t))\, dt \Big), \quad x \in E$$

is bounded on E. Moreover for any $\tau \in \mathcal{M}_\infty$ such that $E_x(\tau) < \infty$, we have that

$$v(x) = E_x\Big(\int_0^{\tau} (\tilde{f}(X_t) - f(X_t))\, dt + v(X_\tau) \Big). \qquad (22)$$

2. POLICY ITERATION TECHNIQUE

We shall describe now a policy iteration technique to obtain the value β^* defined in (17). This technique consist of solving a sequence of truncated optimal stopping problems, where the truncation is on the number of jumps allowed for the PDP. As m, the maximal number of jumps allowed goes to infinity, a sequence β_m of positive numbers converges to β^*. Define $\tilde{f}^* := \min \{\tilde{f}_i;\ i = 1,...,r\}$. From (18) and Proposition 27 it is clear that $\beta^* \leq \tilde{f}^*$.

Policy-iteration algorithm :

1) start with $\beta_0 = \tilde{f}^*$

2) for $m = 1,2,...$ iterate

3) if $l_m(\beta_{m-1}) \geq 0$ then $\beta_m = \beta_{m-1}$ and go to 2); otherwise perform steps 4), 5) and 6) below.

4) for every $\epsilon > 0$, obtain $x_m^\epsilon \in E$ and $\tau_m^\epsilon \in \mathcal{M}_\infty$ such that

$$E_{x_m^\epsilon}\left(\int_0^{\tau_m^\epsilon \wedge T_m} (f(X_s) - \beta_{m-1})\, ds + c(X_{\tau_m^\epsilon \wedge T_m}) \right) + d(x_m^\epsilon) \leq$$

$$l_m(\beta_{m-1}) + \epsilon \qquad (23)$$

5) calculate $p_m := \limsup_{\epsilon \to 0} E_{x_m^\epsilon}\left(\tau_m^\epsilon \wedge T_m \right)$

6) $\beta_m = \beta_{m-1} + \dfrac{l_m(\beta_{m-1})}{p_m}$; go to 2).

Auxiliary definition : We also define for every $\epsilon > 0$, $m = 1,2, \ldots$ and x_m^ϵ, τ_m^ϵ as in step 4 above, $q_m := \liminf_{\epsilon \to 0} E_{x_m^\epsilon}\left(\tau_m^\epsilon \wedge T_m \right)$.

Remark : In step 4) above we have to work with ϵ-optimal solutions since in general we cannot guarantee the existence of exact solutions.

Proposition 28 : For every $m = 1,2,...$

i) $0 \leq \beta_m \leq \beta_{m-1}$

ii) if $l_m(\beta_{m-1}) \leq 0$ then $q_m \geq \dfrac{c_0}{\| f \|}$ (recall that $\| f \| > 0$)

Proof : The proof follows by using induction on m. We shall show only the case m = 1, since the proof that (m ⇒ m+1) is identical. Suppose $l_1(\tilde{f}^*) \leq 0$. Then for all $\epsilon > 0$,

$$- \| f \| E_{x_1^\epsilon}(\tau_1^\epsilon \wedge T_1) + c_0 \leq$$

$$E_{x_1^\epsilon}\Big(\int_0^{\tau_1^\epsilon \wedge T_1} (f(X_s) - \tilde{f}^*) \, ds + c(X_{\tau_1^\epsilon \wedge T_1}) \Big) + d(x_1^\epsilon) \leq$$

$$l_1(\tilde{f}^*) + \epsilon \leq \epsilon \qquad (24)$$

and thus $E_{x_1^\epsilon}(\tau_1^\epsilon \wedge T_1) \geq \frac{c_0 - \epsilon}{\| f \|}$. It follows that $q_1 \geq \frac{c_0}{\| f \|}$. From (24) we have

$$- \tilde{f}^* E_{x_1^\epsilon}(\tau_1^\epsilon \wedge T_1) \leq l_1(\tilde{f}^*) + \epsilon$$

and thus

$$\beta_1 = \tilde{f}^* + \frac{l_1(\tilde{f}^*)}{p_1} \geq \tilde{f}^* - \tilde{f}^* \Big(E_{x_1^\epsilon}(\tau_1^\epsilon \wedge T_1) + \epsilon \Big) / p_1 .$$

Taking the limit superior as $\epsilon \to 0$ we obtain $\beta_1 \geq \tilde{f}^* (1 - q_1/p_1) \geq 0$. Finally, by definition, it is clear that $\beta_1 \leq \tilde{f}^*$. □

Proposition 29 : If for some positive integer m_0, $l_{m_0}(\beta_{m_0-1}) \leq 0$ then $l_m(\beta_{m-1}) \leq 0$ for all $m \geq m_0$.

Proof : Suppose $l_{m_0}(\beta_{m_0-1}) \leq 0$ for some positive integer m_0. Let us show the result by induction on m. For $m = m_0$ the result is true by assumption. Suppose $l_m(\beta_{m-1}) \leq 0$ for $m > m_0$. Then for every $\epsilon > 0$ and x_m^ϵ , τ_m^ϵ as in (23), we have

$$l_{m+1}(\beta_m) = \inf_{x \in E} \Big(\psi_{m+1}(\beta_m, x) + d(x) \Big) \leq \inf_{x \in E} \Big(\psi_m(\beta_m, x) + d(x) \Big) \leq$$

$$E_{x_m^\epsilon}\Big(\int_0^{\tau_m^\epsilon \wedge T_m} (f(X_s) - \beta_{m-1}) ds + c(X_{\tau_m^\epsilon \wedge T_m}) \Big) + d(x_m^\epsilon) -$$

$$l_m(\beta_{m-1}) E_{x_m^\epsilon}(\tau_m^\epsilon \wedge T_m)/p_m \leq$$

$l_m(\beta_{m-1}) + \epsilon - l_m(\beta_{m-1})E_{x_m^\epsilon}(\tau_m^\epsilon \wedge T_m)/p_m$.

Taking the limit superior as $\epsilon \to 0$ and recalling that $-l_m(\beta_{m-1}) \geq 0$ by the induction hypothesis, we obtain that $l_{m+1}(\beta_m) \leq 0$. □

The sequence β_m is decreasing and thus $\beta' := \lim_{m\to\infty}\beta_m = \inf_m \beta_m$ exists and $\beta' \geq 0$. From the last Proposition we can have two cases:

a) either $l_m(\tilde{f}^*) \geq 0$ for all m = 1,2, ... and thus

$$\beta' = \beta_m = \tilde{f}^* \text{ for all } m = 1,2,...,$$

b) or for some positive integer m_0, $l_{m_0}(\tilde{f}^*) < 0$ and from Proposition 29, $l_m(\beta_{m-1}) \leq 0$ for all $m \geq m_0$ with $\beta' \leq \beta_{m_0} < \tilde{f}^*$.

We have the following Theorem.

Theorem 2 : $\beta' = \beta^*$.

Proof : Suppose case a) above first. We have from Proposition 24 that $\beta^* \geq \beta' = \tilde{f}^*$ but since $\beta^* \leq \tilde{f}^*$ it is clear that $\beta^* = \tilde{f}^* = \beta'$. Suppose case b) now. We shall prove the following steps:

Step 1 : $\beta^* \leq \beta'$

Step 2 : p_m is a bounded sequence

Step 3 : $\beta^* \geq \beta'$

Proof of step 1 : For $m \geq m_0$, consider x_{m+1}^ϵ , τ_{m+1}^ϵ as defined in (23). Then

$$E_{x_{m+1}^\epsilon}\left(\int_0^{\tau_{m+1}^\epsilon \wedge T_{m+1}} (f(X_s) - \beta_m)ds + c(X_{\tau_{m+1}^\epsilon \wedge T_{m+1}})\right) + d(x_{m+1}^\epsilon)$$

$$\leq l_{m+1}(\beta_m) + \epsilon \leq \epsilon . \qquad (25)$$

From the proof of Proposition 28,

$$E_{x_{m+1}^\epsilon}(\tau_{m+1}^\epsilon \wedge T_{m+1}) \geq \frac{c_0 - \epsilon}{\|f\|},$$

and from (25) with $\epsilon < c_0$,

$$\beta^* \leq \frac{E_{x^\epsilon_{m+1}}\left(\int_0^{\tau^\epsilon_{m+1} \wedge T_{m+1}} f(X_s)ds + c(X_{\tau^\epsilon_{m+1} \wedge T_{m+1}}) \right) + d(x^\epsilon_{m+1})}{E_{x^\epsilon_{m+1}}(\tau^\epsilon_{m+1} \wedge T_{m+1})}$$

$$\leq \beta_m + \frac{\epsilon}{E_{x^\epsilon_{m+1}}(\tau^\epsilon_{m+1} \wedge T_{m+1})} \leq \beta_m + \epsilon \frac{\| f \|}{c_0 - \epsilon} . \qquad (26)$$

Taking the limit as $\epsilon \to 0$ we get that $\beta^* \leq \beta_m$ and since m is arbitrary, $\beta^* \leq \beta'$.

Proof of step 2 : By contradiction, suppose p_m is unbounded. Then we can find a subsequence $x^{\epsilon_i}_{m_i}$, $\tau^{\epsilon_i}_{m_i}$ of x^ϵ_m , τ^ϵ_m (as defined in (23)) such that $\epsilon_i \downarrow 0$, $m_i \uparrow \infty$ and $E_{x^{\epsilon_i}_{m_i}}(\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}) \to \infty$ as $i \uparrow \infty$. From (22),

$$E_{x^{\epsilon_i}_{m_i}}\left(\int_0^{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}} f(X_s)ds \right) \geq \tilde{f}^* E_{x^{\epsilon_i}_{m_i}}\left(\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}\right) +$$

$$E_{x^{\epsilon_i}_{m_i}}\left(v(X_{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}})\right) - v(x^{\epsilon_i}_{m_i})$$

and from (26) with $\epsilon_i < c_0$,

$$\tilde{f}^* + \frac{E_{x^{\epsilon_i}_{m_i}}\left(v(X_{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}}) + c(X_{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}}) \right) - v(x^{\epsilon_i}_{m_i}) + d(x^{\epsilon_i}_{m_i})}{E_{x^{\epsilon_i}_{m_i}}\left(\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}\right)} \leq$$

$$\frac{E_{x^{\epsilon_i}_{m_i}}\left(\int_0^{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}} f(X_s)ds + c(X_{\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}}) \right) + d(x^{\epsilon_i}_{m_i})}{E_{x^{\epsilon_i}_{m_i}}\left(\tau^{\epsilon_i}_{m_i} \wedge T_{m_i}\right)} \leq$$

$$\beta_{m_i} + \epsilon_i \frac{\| f \|}{c_0 - \epsilon_i} .$$

Letting $i \to \infty$ and recalling that v, d and c are bounded we get that $\tilde{f}^* \leq \beta'$ which is absurd since, from case b), $\beta' < \tilde{f}^*$. Then p_m must be bounded.

Proof of step 3 : From Step 2 there exists $p \in R_+$ such that $p_m \leq p$ for all $m = 1,2, \ldots$. Then for all $m \geq m_0$, $0 \leq -l_m(\beta_{m-1}) = p_m(\beta_{m-1} - \beta_m) \leq p(\beta_{m-1} - \beta_m) \to 0$ as $m \to \infty$, and it means that for any sequence $\epsilon_i \to 0$ as $i \to \infty$ we can find $m_i \to \infty$ such that $l_{m_i+1}(\beta_{m_i}) + \epsilon_i \geq 0$. For any $x \in E$, $\tau \in \mathcal{M}_\infty$, $0 < E_x(\tau) < \infty$, we have

$$0 \leq l_{m_i+1}(\beta_{m_i}) + \epsilon_i \leq E_x\Big(\int_0^{\tau \wedge T_{m_i+1}} (f(X_s) - \beta_{m_i}) \, ds + c(X_{\tau \wedge T_{m_i+1}}) \Big) + d(x) + \epsilon_i$$

and thus

$$\beta' \leq \beta_{m_i} \leq \frac{E_x\Big(\int_0^{\tau \wedge T_{m_i+1}} f(X_s)ds + c(X_{\tau \wedge T_{m_i+1}}) \Big) + d(x) + \epsilon_i}{E_x(\tau \wedge T_{m_i+1})} .$$

Taking the limit as $i \to \infty$ we have from the bounded convergence theorem and $T_{m_i} \to \infty$ P_x-a.s. that

$$\beta' \leq \frac{E_x\Big(\int_0^{\tau} f(X_s)ds + c(X_\tau) \Big) + d(x)}{E_x(\tau)}$$

and from the definition for the case $E_x(\tau) = \infty$ (see (18)) it follows that $\beta' \leq \beta^*$. □

C. THE GENERAL CASE

In this subsection we assume that $\Gamma(x)$ is a compact set of E for each $x \in \tilde{E}$, and make the same assumptions as in subsections II, E, 2 and III, E, 1. We modify the definitions of the operators J, K and L of subsection

II, C in the following way. For v_0, v_1, v_2, v_3 in $C(\tilde{E})$, $x \in \tilde{E}$ and $0 \leq t \leq t^*(x)$ define

$$\mathfrak{B}(v_0,v_2)(x) := v_0(x) + \lambda(x)Qv_2(x)$$

$$J(v_0,v_1,v_2)(t,x) := \int_0^t e^{-\Lambda(s,x)}\mathfrak{B}(v_0,v_2)(\phi(s,x))ds + e^{-\Lambda(t,x)}v_1(\phi(t,x))$$

$$K(v_0,v_2)(x) := \int_0^{t^*(x)} e^{-\Lambda(s,x)}\mathfrak{B}(v_0,v_2)(\phi(s,x))ds + e^{-\Lambda(t^*(x),x)}Qv_2(\phi(t^*(x),x))$$

$$L(v_0,v_1,v_2)(x) := \inf_{0 \leq t < t^*(x)} J(v_0,v_1,v_2)(t,x) \wedge K(v_0,v_2)(x) \ .$$

The following Theorem characterize the optimality equation for the general average impulse control of PDP's.

Theorem 3 : If there exists $\psi \in C(\tilde{E})$ and $\beta^* \geq 0$ such that for all $x \in \tilde{E}$

$$\psi(x) = L(f-\beta^*,M\psi,\psi)(x)$$

then $\hat{\rho}'(x) = \beta^*$ for all $x \in E$.

This Theorem will follow from the Lemmas below.

Lemma 3 : The process

$$\int_0^t (f(X_s)-\beta^*)\, ds + \psi(X_t)$$

is an $\mathcal{F}_t$-submartingale.

Proof : Let us show by induction on m that

$$\mathfrak{S}_m(t,x) := E_x\left(\int_0^{t \wedge T_m} (f(X_s) - \beta^*)\, ds + \psi(X_{t \wedge T_m}) \right) \geq \psi(x)$$

for all $t \in R_+$, $x \in \tilde{E}$ and $m = 0,1,\ldots$. For $m = 0$ the result is

immediate. Suppose $\psi(y) \leq \mathcal{E}_m(t,y)$ for all $y \in \tilde{E}$ and $t \in R_+$. Then

$$\mathcal{E}_{m+1}(t,x) = E_x\Big(\int_0^{t \wedge T_1} (f(\phi(s,x)) - \beta^*)ds + \psi(\phi(t,x))1_{\{T_1>t\}} +$$

$$\mathcal{E}_m(t-T_1,Z_1)1_{\{T_1 \leq t\}}\Big) \geq$$

$$E_x\Big(\int_0^{t \wedge T_1} (f(\phi(s,x))-\beta^*)ds + \psi(\phi(t,x))1_{\{T_1>t\}} + \psi(Z_1)1_{\{T_1 \leq t\}}\Big) \geq$$

$$L(f-\beta^*,\psi,\psi)(x) = \psi(x)$$

where the last equality follows from Corollary 1 of [24]. Thus we have shown that $\psi(x) \leq \mathcal{E}_m(t,x)$. Applying the bounded convergence theorem and recalling the fact that $T_m \to \infty$ as $m \to \infty$ P_x-a.s we obtain that

$$\psi(x) \leq E_x\Big(\int_0^t (f(X_s) - \beta^*)\, ds + \psi(X_t) \Big)$$

for all $t \in R_+$ and $x \in \tilde{E}$. The Lemma follows from time homogeneity and strong Markov property of $\{X_t\}$. □

Lemma 4 : For any $\mathcal{G} \in \mathcal{S}$ and $x \in E$, $\beta^* \leq V^{\mathcal{G}}(x)$.

Proof : From Lemma 3, $\psi \leq M\psi$, and using arguments similar to those in the proof of Lemma 7 below we get

$$\psi(x) \leq E_x^{\mathcal{G}}\Big(\int_0^t (f(Y_s) - \beta^*)ds + \sum_{i=0}^{\infty} c(Y_{\tau_i},Y_{\tau_i}+)1_{\{\tau_i \leq t\}} + \psi(Y_t) \Big).$$

Taking liminf as $t \to \infty$ and from boundedness of ψ we obtain the desired result. □

Optimal stopping problems of PDP's may not have an optimal stopping time (cf [24]). Therefore we will need ϵ-optimal solutions, given in the next two Lemmas.

Lemma 5 : Let κ be an integer such that $\kappa c_0 > 2 \| \psi \|$ (the sup norm). For $0 < \epsilon < (\kappa c_0 - 2 \| \psi \|)/\kappa$ we can find a Borel measurable selector $R_\epsilon(.) : E \to E$ such that $R_\epsilon(x) \in \Gamma(x)$, $c(x,R_\epsilon(x)) + \psi(R_\epsilon(x)) \leq M\psi(x) + (\kappa-1)\epsilon$ and $M\psi(R_\epsilon(x)) > \psi(R_\epsilon(x)) + \epsilon$ for all $x \in E$.

Proof : See [35].

Lemma 6 : Define $U_\epsilon = \inf \{ t \geq 0; M\psi(X_t) \leq \psi(X_t) + \epsilon \}$ for $\epsilon > 0$. Then the process

$$\int_0^{t \wedge U_\epsilon} (f(X_s) - \beta^*)ds + \psi(X_{t \wedge U_\epsilon})$$

is an $\mathcal{F}_t$-martingale.

Proof : Lemma 2 in [24] is readily modified to show that

$$E_x \Big(\int_0^{t \wedge U_\epsilon \wedge T_m} (f(X_s) - \beta^*) \, ds + \psi(X_{t \wedge U_\epsilon \wedge T_m}) \Big) = \psi(x)$$

for all $m = 0,1,\ldots$, $t \geq 0$ and $x \in E$. The remainder of the proof is as in Lemma 3 above. □

For ϵ as in Lemma 5 let $\epsilon_i = \epsilon/2^i$, $i = 1,2, \ldots$ and define the strategy $\mathcal{S}^*$ by:

$S_1^*(\omega_1) = U_{\epsilon_1}(\omega_1)$, $\omega_1 \in \Omega$, $\hat{\omega}_1 \in \hat{\Omega}$ as in the remark of subsection III, B,

$$R_1^*(w_1) = \begin{cases} R_{\epsilon_2}(\hat{x}_{\eta(\hat{\omega}_1)}(\hat{\omega}_1)) \text{ if } \eta(\hat{\omega}_1) < \infty, \\ \Delta \text{ otherwise,} \end{cases}$$

$w_1 = (\hat{\omega}_1) \in \mathcal{W}_1$, and for $i = 2,3,\ldots$,

$$S_i^*(w_{i-1},\omega_i) = \begin{cases} U_{\epsilon_i}(\omega_i) \text{ if } R_{i-1}^*(w_{i-1}) \neq \Delta, \\ 0 \text{ otherwise,} \end{cases}$$

$\omega_i \in \Omega$, $\hat{\omega}_i \in \hat{\Omega}$ as in the remark of subsection III, B, and

$$R_i^*(w_i) = \begin{cases} R_{\epsilon_{i+1}}(\hat{x}_{\eta(\hat{\omega}_i)}(\hat{\omega}_i)) \text{ if } \eta(\hat{\omega}_i) < \infty \text{ and } R_{i-1}^*(w_{i-1}) \neq \Delta, \\ \Delta \text{ otherwise,} \end{cases}$$

$w_i = (\hat{\omega}_1,\ldots,\hat{\omega}_i) \in \mathcal{W}_i$, where R_{ϵ_i}, U_{ϵ_i} are as in Lemma 5 and 6 respectively, replacing ϵ by ϵ_i. Note that from the above construction and Lemma 5 we get that $\tau_{i+1} > \tau_i$ (on $\tau_i < \infty$) since that $M\psi(R_{\epsilon_{i+1}}(x)) > \psi(R_{\epsilon_{i+1}}(x)) + \epsilon_{i+1}$. Lemmas 4 and 7 below completes the proof of Theorem 3.

Lemma 7 : For all $x \in E$, $\beta^* \geq V^{\mathcal{P}^*}(x)$.

Proof : For all $t \in R_+$, $x \in E$ and $i = 1,2, \ldots$ define

$$C_i(t,x) := E_x\Big(\int_0^{t \wedge U_{\epsilon_i}} (f(X_s) - \beta^*)\, ds + \Big(c(X_{U_{\epsilon_i}}, R_{\epsilon_{i+1}}(X_{U_{\epsilon_i}})) +$$

$$\psi(R_{\epsilon_{i+1}}(X_{U_{\epsilon_i}}))\Big) 1_{\{U_{\epsilon_i} \leq t\}} + \psi(X_t) 1_{\{U_{\epsilon_i} > t\}} \Big) \leq \psi(x) + \kappa\epsilon_i$$

where the inequality above follows from Lemmas 5 and 6. We get that

$$E_x^{\mathcal{P}^*}\Big(\int_0^{t \wedge \tau_i} (f(Y_s) - \beta^*)\, ds + \sum_{j=1}^{i} \Big(c(Y_{\tau_j}, Y_{\tau_j}+)\, 1_{\{\tau_j \leq t\}}$$

$$+ \psi(Y_{\tau_j+} \wedge t)\Big)\Big) = E_x^{\mathcal{P}^*}\Big(\sum_{j=1}^{i} \Big(\Big\{ \int_{t \wedge \tau_{j-1}}^{t \wedge \tau_j} (f(Y_s) - \beta^*)\, ds +$$

$$\Big(c(Y_{\tau_j}, Y_{\tau_j}+) + \psi(Y_{\tau_j}+)\Big) 1_{\{\tau_j \leq t\}} +$$

$$\psi(Y_t)1_{\{\tau_j>t\}}\}1_{\{\tau_{j-1} \le t\}} + \psi(Y_t)1_{\{\tau_{j-1}>t\}}\Big)\Big) =$$

$$E_x^{\mathcal{P}^*}\Big(\sum_{j=1}^{i}\Big(c_j(t-\tau_{j-1},Y_{\tau_{j-1}+})1_{\{\tau_{j-1}\le t\}} + \psi(Y_t)1_{\{\tau_{j-1}>t\}}\Big)\Big) \le$$

$$E_x^{\mathcal{P}^*}\Big(\sum_{j=1}^{i}\psi(Y_{\tau_{j-1}+\wedge t})\Big) + \kappa\epsilon\ .$$

Therefore for all $i = 1, 2, \ldots$ and $x \in E$,

$$E_x^{\mathcal{P}^*}\Big(\int_0^{t\wedge\tau_i}(f(Y_s) - \beta^*)ds + \sum_{j=1}^{i} c(Y_{\tau_j},Y_{\tau_j+})\,1_{\{\tau_j\le t\}} + \psi(Y_{\tau_i+\wedge t})\Big)$$

$$\le \psi(x) + \kappa\epsilon\ .$$

From the above expression it is easy to check that $P_x^{\mathcal{P}^*}(\tau_i \le t) \le (2\|\psi\| + \kappa\epsilon + \beta^* t)/ic_0$ for any $t \in R_+$ and this result yields that

$$P_x^{\mathcal{P}^*}(\lim_{i\to\infty}\tau_i < \infty) = 0.$$

From the bounded convergence theorem it follows that

$$E_x^{\mathcal{P}^*}\Big(\int_0^{t}(f(Y_s) - \beta^*)\,ds + \sum_{j=1}^{\infty} c(Y_{\tau_j},Y_{\tau_j+})\,1_{\{\tau_j\le t\}} + \psi(Y_t)\Big) \le$$

$$\psi(x) + \kappa\epsilon\ .$$

Dividing by t, taking the liminf as $t \to \infty$ and from boundedness of ψ we get the desired result. □

V. REFERENCES

1. A. Bensoussan and J.L. Lions, "Nouvelles Methodes en Contrôle Impulsionnel", *Appl. Math . Optim.* **1**, pp 289-312 (1975).

2. A. Bensoussan and J.L. Lions, *Applications des Inequations Variationnelles en Contrôle Stochastique*, Dunod, Paris, (1978).

3. A. Bensoussan and J.L. Lions, *Contrôle Impulsionnel et Inequations Quasi-Variationnelles*, Dunod, Paris, (1982).

4. H.J. Kushner, "Approximations and computational methods for optimal stopping and stochastic impulsive control problems", *Appl. Math. Optim.* **3**, pp 81-100 (1977).

5. H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York (1977).

6. J.P. Lepeltier and B. Marchal, "Théorie générale du contrôle impulsionnel Markovien", *SIAM J. Control Optim.* **22**, pp 645-665 (1984).

7. M. Robin, *Contrôle impulsionnel des processus de Markov*, Thesis, Universite Paris IX (1978).

8. M. Robin, "On some impulsive control problems with long run average cost, *SIAM J. Control Optim.* **19**, pp 333-358 (1981).

9. D. Gatarek and L. Stettner, "On the compactness method in general ergodic impulsive control of Markov processes", *Stochastics* **31**, pp 15-25 (1990).

10. L. Stettner, "On the Poisson equation and optimal stopping of ergodic Markov processes", *Stochastics* **18**, 25-48 (1986).

11. L. Stettner, "On ergodic impulsive control problems", *Stochastics* **18**, 49-72 (1986).

12. L. Stettner, "On impulsive control with long run average cost criterion", *Studia Math.* **76**, pp 279-298 (1983).

13. F.A. Van der Duyn Schouten, "Markov Decision Drift Processes,

CWI Tract, Centre for Mathematics and Computer Science, Amsterdam (1983).

14. A. Hordijk and F.A. Van der Duyn Schouten, "Average optimal policies in Markov decision drift processes with applications to a queuing and a replacement model", *Adv. in Appl. Probab.* **15**, pp 274-303 (1983).

15. A. Hordijk and F.A. Van der Duyn Schouten, "Discretization and weak convergence in Markov decision drift processes", *Math. Oper. Res.* **9**, pp. 112-141 (1984).

16. A.A. Yushkevich, "Continuous-time Markov decision processes with intervention", *Stochastics* **9**, pp 235-274 (1983).

17. A.A. Yushkevich, "Bellman inequalities in Markov decision deterministic drift processes, *Stochastics* **23**, pp 25-77 (1987).

18. M.H.A. Davis, "Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models", J.*Royal Statistical Soc. (B)* **46**, pp 353-388 (1984).

19. M.H.A. Davis, *Markov Models and Optimization*, Monographs on Statistics and Applied Probability **49**, Chapman and Hall, London (1993).

20. M.A.H. Dempster and J.J. Ye, "Impulse control of piecewise-deterministic Markov processes", *Ann. Appl. Prob*, to appear.

21. D. Gatarek, "On value functions for impulsive control of piecewise-deterministic processes", *Stochastics* **32**, pp 27-52 (1990).

22. D. Gatarek, "Optimality conditions for impulsive control of piecewise-deterministic processes", *Math. Control Signals Systems* **5**, pp 217-232 (1992).

23. D. Gatarek, "Impulsive control of piecewise-deterministic processes with long run average cost", *Stochastics*, to appear.

24. U.S. Gugerli, "Optimal stopping of a piecewise-deterministic Markov process", *Stochastics* **19**, pp 221-236 (1986).

25. O.L.V. Costa and M.H.A. Davis, "Approximations for optimal

stopping of a piecewise-deterministic process", *Math. Control Signals Systems* **1**, pp 123-146 (1988).

26. D. Bertsekas and S.E. Shreve, *Stochastic Optimal Control : The Discrete-Time Case*, Academic Press, New York, (1978).

27. P. Billingsley, *Probability and Measure*, Wiley, New York, (1979).

28. M.E. Munroe, *Introductory Real Analysis*, Addison Wesley, Reading, MA (1964).

29. O.L.V. Costa and M.H.A. Davis, "Impulse control of piecewise-deterministic processes", *Math. Control Signals Systems* **2**, pp 187-206 (1989).

30. J. Zabczyk, Lectures in Stochastic Control, Control Theory Centre Report No 125, University of Warwick (1984).

31. A. Schornagel, "Optimum preventive maintenance for complex systems with independent equipment units, preprint, Shell Research Laboratory, Amsterdam (1987).

32. O.L.V. Costa, "Impulse control of piecewise-deterministic processes via linear programming", *IEEE Transc. Automatic Control* **AC-36**, pp 371-375 (1991).

33. H. L. Royden, *Real Analysis*, Macmillan, New York, (1968).

34. O.L.V. Costa, "Average impulse control of piecewise-deterministic processes", *IMA J. Math. Control Inform.* **6**, pp 375-397 (1989).

35. O.L.V. Costa, "Asymptotic convergence for the general average impulse control of piecewise deterministic processes", *IMA J. Math. Control Inform.* **8**, pp 1-27 (1991).

36. O.L.V. Costa, "Discretizations for the average impulse control of piecewise-deterministic processes", *J. Appl. Probab.* **30**, pp 405-420 (1993).

37. J. L. Doob, *Stochastic Processes*, Wiley, New York, (1953).

38. O.L.V. Costa, "Stationary distributions for piecewise-deterministic processes", *J. Appl. Probab.* **27**, pp 60-73 (1990).

INDEX

I

J

K

L

M

O

P

R

S

This Page Intentionally Left Blank