



# LUND UNIVERSITY

## New Approaches of the Numerical Solution of Optimal Control Problems

Mårtensson, Krister

1972

*Document Version:*

Publisher's PDF, also known as Version of record

[Link to publication](#)

*Citation for published version (APA):*

Mårtensson, K. (1972). *New Approaches of the Numerical Solution of Optimal Control Problems*. [Doctoral Thesis (monograph), Department of Automatic Control]. Department of Automatic Control, Lund Institute of Technology (LTH).

*Total number of authors:*

1

### General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117  
221 00 Lund  
+46 46-222 00 00

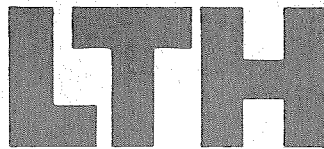
1002

REPORT 7206

MARCH 1972

# New Approaches to the Numerical Solution of Optimal Control Problems

KRISTER MÅRTENSSON

The logo for the Lund Institute of Technology (LTH) is displayed in a large, bold, black font. It consists of the letters 'LTH' in a stylized, blocky typeface. The 'L' and 'H' are connected at the top, and the 'T' is positioned between them. The logo is centered horizontally and is set against a dark, solid background that spans the width of the page.

Division of Automatic Control · Lund Institute of Technology  
Studentlitteratur

Krister Mårtensson

New Approaches to the Numerical Solution of  
Optimal Control Problems

Distribution: Studentlitteratur

Lund 1972

To Ulla, with thanks for her constant encouragement  
and support.

Printed in Sweden  
Studentlitteratur  
Lund 1972  
ISBN 91-44-08851-5



# TABLE OF CONTENTS

## INTRODUCTION 5

## PART 1 - ON THE MATRIX RICCATI EQUATION 15

1. INTRODUCTION 16
2. THE ALGEBRAIC EQUATION  $A^T X + XA -$   
 $- XBQ_2^{-1}B^T X + Q_1 = 0$  18
3. THE RICCATI EQUATION IN OPTIMAL CONTROL  
PROBLEMS 52
4. REFERENCES 68

## PART 2 - A NEW APPROACH TO CONSTRAINED FUNCTION OPTIMIZATION 71

1. INTRODUCTION 72
2. NECESSARY AND SUFFICIENT CONDITIONS FOR A  
CONSTRAINED LOCAL MINIMUM 74
3. LAGRANGE MULTIPLIER FUNCTIONS 76
4. EXAMPLES 84
5. ALGORITHMS AND COMPUTATIONAL ASPECTS 90
6. REFERENCES 100

## PART 3 - A CONSTRAINING HYPERPLANE TECHNIQUE FOR STATE VARIABLE CONSTRAINED OPTIMAL CONTROL PROBLEMS 103

1. INTRODUCTION 104
2. STATEMENT OF THE PROBLEM 106
3. SURVEY OF NECESSARY CONDITIONS FOR OPTIMA-  
LITY 107
4. THE CONSTRAINING HYPERPLANE TECHNIQUE 111
5. A SECOND-ORDER DIFFERENTIAL DYNAMIC  
PROGRAMMING ALGORITHM FOR MIXED STATE-  
CONTROL VARIABLE CONSTRAINTS  $g(x, u; t) \leq 0$  121
6. EXAMPLES 146
7. REFERENCES 183

PART 4 - OPTIMAL CONTROL OF A TRAVELLING OVERHEAD  
CRANE - A FEASIBILITY STUDY 187

1. INTRODUCTION 188
2. STATEMENT OF THE PROBLEM 190
3. MATHEMATICAL MODELS OF THE CRANE 192
4. OPTIMAL CONTROL OF MODEL 1 198
5. OPTIMAL CONTROL OF MODEL 2 218
6. REFERENCES 225

## INTRODUCTION

The mathematical theory of optimal control has developed very rapidly since the maximum principle and the method of dynamic programming were discovered in the late fifties. A very large number of contributions have been presented in different books and papers, and the theory for optimal control of deterministic processes has now reached a fairly satisfactory stage. In parallel with the development of the theory, computational methods have been continuously studied, and considerable progress have been made in this area too. However, there is still a remarkable gap between theory and practice, a fact which is emphasized by the small number of industrial or other applications that have been reported. There may be many reasons for this gap, but one certainly is that the existing computational methods are not efficient enough. However, except for such obvious reasons as too long execution times etc., it is in general impossible to isolate a few reasons that explain why the developed methods often are insufficient for problems with the complexity encountered in many industrial processes. It is rather a wide range of reasons that are accumulated, and which together make the development of efficient numerical methods both difficult and time-consuming.

In this thesis we consider different problems related to the numerical solution of optimal control problems. The thesis is divided into four separate parts, which in principle are independent of each other, but they all treat problems in the wide range between theory and applications. The titles of the different parts are:

- Part 1 - On the Matrix Riccati Equation. (In an earlier version published in "Information Sciences", vol. 3, 1971).
- Part 2 - A New Approach to Constrained Function Optimization. (Accepted for publication in the "Journal of Optimization Theory and Applications").
- Part 3 - A Constraining Hyperplane Technique for State Variable Constrained Optimal Control Problems.
- Part 4 - Optimal Control of a Travelling Overhead Crane - a feasibility study.

The four parts are organized independently of each other. Thus all equations are numbered afresh in the different parts, and the references are collected at the end of each part. Since the different parts were originally written as separate reports or papers, they are also sometimes referred to as papers. Hopefully this will not lead to any confusion.

Part 1 of the thesis treats the linear-quadratic optimal control problem. We thus consider the linear system

$$\frac{dx}{dt} = Ax + Bu \quad x(t_0) = x_0 \quad (1)$$

with the quadratic cost functional

$$J = x^T(t_f)Q_0x(t_f) + \int_{t_0}^{t_f} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\} ds \quad (2)$$

where  $Q_0$  and  $Q_1$  are nonnegative definite symmetric matrices and  $Q_2$  is a positive definite symmetric matrix. It is well known that the solution of (1)-(2) is given by the linear feedback

$$u(t) = -Q_2^{-1}B^T S(t)x(t) \quad (3)$$

where  $S(t)$  is the nonnegative definite symmetric solution of the Riccati differential equation

$$-\frac{dS}{dt} = A^T S + SA - SBQ_2^{-1}B^T S + Q_1 \quad S(t_f) = Q_0 \quad (4)$$

The problem formulation (1)-(2) becomes particularly useful by letting  $t_0$  tend to  $-\infty$ . It is then referred to as the regulator problem, and the solution (if it exists) is given by the stationary solution of (4), that is, by a solution of the algebraic equation

$$A^T X + XA - XBQ_2^{-1}B^T X + Q_1 = 0 \quad (5)$$

A thorough study of this equation will thus yield further insight into both the regulator problem and into properties of the optimal solution of (1)-(2) for large time intervals  $t_f - t_0$ .

The properties of (5) have previously been studied by Kalman ([7], part 1) and Wonham ([8], part 1). In [7] it was shown that if the system

$[A, B]$  is completely controllable and the pair  $[Q_1, A]$  completely observable, then there is only one nonnegative definite symmetric solution of (5). Moreover, this solution is positive definite. In [8] Wonham relaxed the assumptions to stabilizability and detectability, that is, modes of  $A$  with  $\text{Re}\{\lambda\} \geq 0$  are controllable in  $[A, B]$  and observable in  $[Q_1, A]$ . It can then be proved that there is still a unique nonnegative definite symmetric solution of (5), but this solution is not necessarily positive definite.

We make a further generalization, and consider properties of (5) for arbitrary nonnegative definite symmetric matrices  $Q_1$ , that is, the observability assumption is completely relaxed. It is then shown that every solution of (5) may be expressed in terms of the eigenvectors of the Euler matrix

$$E = \begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \quad (6)$$

and conditions for Hermitian, real symmetric and nonnegative definite solutions of (5) are established. These results are mainly generalizations of the results given by Potter ([1], part 1). In [1] it was assumed that  $E$  has a diagonal Jordan form. This restriction is shown to exclude many interesting cases, and our results hold for a general Jordan form of  $E$ .

We then prove that the eigenvectors of  $E$  have a particular structure in case any mode of  $A$  is noncontrollable or nonobservable, and it is shown that this implies that there may be several nonnegative definite solutions of (5) if  $A$  has nonobservable unstable modes. It is also possible to order these solutions.

There is a weakness in the algebraic results which are presented, namely that we have not been able to prove the existence of the different nonnegative definite solutions. The explicit expression for the solutions involves a matrix inverse, and the statements about the character of the solutions have been necessary to make under the assumption that this inverse exists. However, it is probably possible to give a proof also for the existence, and in all numerical problems considered these solutions have actually existed.

The algebraic results are then used to study the asymptotic properties of (4), and it is shown that the different nonnegative definite solutions of (5) can be given a nice interpretation. It is also shown that the existence

of several stationary solutions of (4) implies that a straightforward integration can be a numerically unstable procedure. This is illustrated by numerical examples, and it is briefly discussed how this property influences the choice of numerical methods. The results on numerical properties yield valuable insight also into numerical methods for non-linear problems, since these methods often are based on the integration of similar Riccati equations.

Finally, it is shown that the regulator problem can be generalized to arbitrary nonnegative definite matrices  $Q_1$ . However, notice that to prove this result it is necessary to assume the existence of a largest nonnegative definite solution of (5).

In part 2 we consider the problem of minimizing a real-valued function  $f(u)$  subject to the equality constraints  $g(u) = 0$ ,  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , and a new approach is presented. It is shown that the conventional Lagrange multipliers can be generalized, and the new concept "Lagrange multiplier function" is introduced. Thus the  $m$ -dimensional vector function  $\mu(u)$  is called a multiplier function of the problem if  $\mu$  satisfies some simple conditions for the optimal solution  $u^*$ . The basic condition is  $\mu(u^*) = \lambda^*$ , where  $\lambda^*$  are the corresponding optimal Lagrange multipliers. We also introduce the generalized Lagrangian

$$H(u, c) = f(u) + \mu^T(u)g(u) + c g^T(u)g(u) \quad (7)$$

and it is shown that there exists a finite real-valued parameter  $c_0$ , such that  $H(u, c)$  for nonsingular problems has an isolated local minimum for  $u = u^*$  if  $c > c_0$ . The problem is thus converted from a constrained optimization problem into an unconstrained problem. Notice that  $c_0$  in this case is finite, while ordinary penalty function methods require that the penalty parameter tends to infinity.

Compared with the earlier works by Hestenes ([3], part 2), Fletcher ([5], part 2) and Mårtensson ([4], part 2), the approach made in this thesis is more general, and it contains all these previously suggested methods as special cases. This is illustrated by examples.

The multiplier functions can be used in different ways to design computational methods. One possibility is to directly minimize the generalized Lagrangian  $H(u, c)$  with ordinary function minimization methods. This is a very straightforward way, and it is illustrated by an example that the multiplier function can be chosen so that  $H(u, c)$  becomes well conditioned for direct minimization. Another possibility is to use the multiplier functions to get successive estimates of the

optimal Lagrange multipliers  $\lambda^*$ . This can be done in many ways. We have here chosen to investigate the following basic scheme:

i) Let  $\mu_k$  be an estimate of  $\lambda^*$  and minimize

$$F(u, \mu_k) = f(u) + \mu_k^T g(u) + c g^T(u)g(u) \quad (8)$$

with an ordinary function minimization method.

ii) Compute a new estimate  $\mu_{k+1}$  on the basis of the minimizing solution of (8) and return to i).

The advantage with this kind of estimation methods compared with direct minimization of  $H(u, c)$  is illustrated, and different first and second order methods based on the outlined scheme are derived. We also prove the convergence properties of the algorithms.

There are many analogies between the finite-dimensional problems considered in part 2 of this thesis and optimal control problems for dynamic systems. For example, in optimal control problems the equality constraints

$$\frac{dx}{dt} - f(x, u; t) = 0 \quad \forall t \in [t_0, t_f] \quad (9)$$

can be handled through adjoining of (9) to the cost functional with the adjoint variables  $\lambda(t)$ . The numerical methods are then generally based on schemes similar to the one outlined above, and the optimal multipliers  $\lambda^*(t)$  are successively estimated. However, the cost functional  $J(u, \lambda_k)$  is generally not completely minimized before a new estimate is computed. Besides it is not necessary to include the quadratic term

$$c \left( \frac{dx}{dt} - f(x, u; t) \right)^T \left( \frac{dx}{dt} - f(x, u; t) \right) \text{ in the cost functional if the constraint (9)}$$

is always satisfied. Thus most of the proposed algorithms are based on forward integration of (9), and backward integration of the adjoint equations which gives an estimate of the optimal multipliers  $\lambda^*(t)$ . Also notice that the additional constraint  $x(t_0) = x_0$  is automatically satisfied when the boundary condition in the forward integration of (9) is equal to  $x_0$ .

Recently, a new algorithm, the  $\epsilon$ -technique has been proposed by Balakrishnan. The method is a straightforward generalization of the

ordinary penalty function method, and is based on the fact that the unconstrained solution of

$$\bar{J} = J + \frac{1}{\epsilon} \int_{t_0}^{t_f} \left( \frac{dx}{dt} - f(x, u; t) \right) \left( \frac{dx}{dt} - f(x, u; t) \right) dt \quad (10)$$

tends to the optimal solution of  $J$  subject to the constraints (9) as  $\epsilon$  tends to zero. It is thus interesting to notice that in analogy with finite-dimensional optimization, numerical methods based on both Lagrange multipliers (or adjoint variables) and penalty functions exist for the optimal control problems. However, the possibility to combine the multipliers and the penalty functions in the way it is done in this thesis for finite-dimensional problems has not yet been investigated for the control problems. The results presented here also suggest the further possibility to convert the control problem into a completely unconstrained problem.

In part 3 a new approach to the numerical solution of optimal control problems with state variable inequality constraints is presented. We thus consider the general problem of minimizing the cost functional

$$J = F(x(t_f); t_f) + \int_{t_0}^{t_f} L(x, u; t) dt \quad (11)$$

subject to the constraints

$$\begin{aligned} \frac{dx}{dt} &= f(x, u; t) & x(t_0) &= x_0 \\ \psi(x(t_f); t_f) &= 0 & & \\ S(x; t) &\leq 0 & \forall t &\in [t_0, t_f] \\ g(x, u; t) &\leq 0 & \forall t &\in [t_0, t_f] \end{aligned} \quad (12)$$

where the state variable inequality constraint  $S \leq 0$  is of arbitrary order  $q$ . That is, the  $q$ -th total time derivative of  $S$  is the lowest order total time derivative that explicitly contains the control variable  $u$ .

The appearance of pure state variable constraints  $S$  increases the



complexity of the problem in a drastic way, and considerable effort has been dedicated to develop numerical methods for this problem. In ([4], part 3) Denham and Bryson exploit necessary conditions at the entry times, that is, the times when the constraint becomes active. The main drawback of their method is that it is necessary to à priori know the number of entry points, and it may also be necessary to have good estimates of the different entry times. This à priori knowledge is seldom available which is illustrated by part 4 of this thesis. Different penalty function methods have been proposed by different authors, e.g. Kelley ([11], part 3). The basic idea in these methods is to convert the original problem into an unconstrained problem by adding penalty terms to the cost functional  $J$ , and the methods are thus straightforward generalizations of existing penalty function methods for finite dimensional problems. However, it turns out that these methods are extremely sensitive to numerical errors, and it is difficult to compute the optimal solution accurately. Recently an alternative method has been proposed by Jacobson and Lele ([9], part 3). This method is based on a slack variable technique, and by introducing a sufficient number of additional state variables, the problem is converted into an unconstrained problem of higher dimension. However, the transformed problem becomes a singular problem, and then new computational difficulties appear. In spite of this disadvantage, this slack variable technique has in general proved to be superior to penalty function methods.

The basic idea in the method presented in this thesis is to approximate the feasible region  $S(x;t) \leq 0$  in the state space with a region that can be expressed as an explicit function of the control variables, that is,  $g(x, u;t) \leq 0$ . The transformed problem can then be solved with the same technique as problems with pure control variable constraints  $g(u;t) \leq 0$ . The approximation can be done in many different ways, but a simple and natural approach is to choose  $g(x, u;t)$  as hyperplanes in

the  $(S, \frac{dS}{dt}, \dots, \frac{d^q S}{dt^q})$ -space, that is the new constraint is

$$g(x, u;t) = \frac{d^q S}{dt^q}(x, u;t) + a_1 \frac{d^{q-1} S}{dt^{q-1}}(x;t) + \dots + a_q S(x;t) \leq 0 \quad (13)$$

It is shown that this half-space tends to the half-space  $S \leq 0$  as the zeroes  $\xi_1, \dots, \xi_q$  of the polynomial

$$p^q + a_1 p^{q-1} + \dots + a_q = 0 \quad (14)$$

tend to  $-\infty$ . To simplify the analysis we have assumed that  $a_1, \dots, a_q$  are chosen so that  $\xi_1 < \dots < \xi_q < 0$ , but the generalization to more general parameters is straightforward.

To solve the mixed state-control variable constrained problem, a second order algorithm based on Differential Dynamic Programming is then derived. The multiplier function concept introduced in part 2 is used in the derivation, and is shown to be very simple compared with the ordinary Lagrange multiplier technique. We also indicate some possible generalizations of the algorithm on the basis of the results in part 2.

The efficiency and the accuracy of the constraining hyperplane technique are investigated on a number of different problems. Comparisons are made with the established methods mentioned above, and it is shown that the combination of constraining hyperplanes and the second order algorithm is much superior as far as both accuracy and efficiency is concerned. For example, a problem with a third order constraint is solved, and to our knowledge this has not been done before with neither the penalty function methods nor with the slack variable technique. It is also illustrated that it may be possible to derive the existing necessary conditions for state variable constrained problems by letting the zeroes of (14) tend to  $-\infty$ . It is thus believed that the constraining hyperplane technique could also contribute to a better understanding of these problems.

Part 4 consists of a feasibility study of optimal control of a travelling overhead crane. The problem originates from a container terminal, and the purpose of the optimization is to minimize the load transfer times. The transfer is performed by a travelling overhead crane, and can be described by a set of ordinary differential equations. Two different models corresponding to different properties of the crane are studied. In both cases six state variables and two control variables are necessary to describe the operation, and the magnitude of the problem is thus unusually large.

The study illustrates the present status of computational methods, and also a lot of problems that must be taken into account in a practical realization, e.g. disturbances, execution times and available computer programs. It is also shown that there are further interesting analogies between the finite-dimensional problem considered in part 2 and optimal control problems.

It should be emphasized, that the optimal solutions for the different

cases have been obtained with a pre-designed programming package. To our knowledge, this is the first attempt to reduce the necessary programming work and to make optimal control theory available as a standard tool for analysis and synthesis of control problems.

### Acknowledgements

I am glad to express my sincere appreciation to Professor Karl Johan Åström for his encouragement and invaluable help throughout this work.

The research presented in this thesis was made possible through the support extended by the Swedish board for Technical Development under Contract no. 71-50/U33. The container problem originates from the Swedish company ASEA, and I want to express my gratitude to this company and particularly to Ing. Åke Rullgård for the cooperation. I am also indebted to Professor Peter Falb, Brown University USA, for many valuable comments on parts of the thesis. I have also profited from the contributions of many people at the Division of Automatic Control. In particular I would like to thank Tekn. lic. Per Hagander and Civ. ing. Torkel Glad for suggested improvements, Mrs. Gudrun Christensen and Mrs. Christina Pira who typed the first manuscripts, and Mrs. Carina Bolinder and Mrs. Birgitta Tell who prepared the figures.

## 1. INTRODUCTION

The matrix Riccati equation appears in many optimal control and filtering problems. In this paper the Riccati equation is studied from an algebraic point of view, and the results are applied to optimal control of linear time invariant systems with quadratic loss. Consider the system

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t), \quad x(t_0) = x_0, \quad (1.1)$$

where  $x$  is the  $n$ -dimensional state vector,  $u$  is the  $r$ -dimensional control vector, and  $A$  and  $B$  are matrices of dimension  $n \times n$  and  $n \times r$ . It is desired to determine a control  $u(t)$ , so that the loss function

$$J = x^T(t_f)Q_0x(t_f) + \int_{t_0}^{t_f} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\} ds \quad (1.2)$$

is minimized.  $Q_0$  and  $Q_1$  are symmetric nonnegative definite  $n \times n$  matrices, and  $Q_2$  is a symmetric positive definite  $r \times r$  matrix. It is well known [5] that the optimal control is given as a linear feedback from the state of the system

$$u(t) = -L(t)x(t) \quad (1.3)$$

where

$$L(t) = Q_2^{-1} B^T S(t) \quad (1.4)$$

and  $S(t)$  is the solution of the Riccati equation

$$-\frac{dS(t)}{dt} = A^T S(t) + S(t)A - S(t)BQ_2^{-1}B^T S(t) + Q_1 \quad (1.5)$$

The boundary condition is given at  $t = t_f$  as

$$S(t_f) = Q_0. \quad (1.6)$$

A special case of great interest is what is called the regulator problem. The task of the control is then to minimize

$$J = \int_0^{\infty} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\} ds. \quad (1.7)$$

Introducing controllability or stabilizability conditions on the system  $[A, B]$ , this can be considered as the limit of (1.2) as  $t_0 \rightarrow -\infty$  [5, 8]. The optimal control then is a linear time invariant feedback

$$u(t) = -Lx(t) \quad (1.8)$$

where

$$L = Q_2^{-1} B^T S \quad (1.9)$$

and  $S$  is a symmetric nonnegative definite solution of the stationary Riccati equation

$$A^T S + SA - SBQ_2^{-1} B^T S + Q_1 = 0. \quad (1.10)$$

If an observability condition is imposed on the pair  $[C, A]$ , where  $Q_1 = C^T C$  and  $\text{rank } C = \text{rank } Q_1$ , there is a unique nonnegative definite solution of (1.10). Moreover, this solution is positive definite, and the optimal closed loop system

$$\frac{dx(t)}{dt} = (A-BL)x(t) \quad (1.11)$$

is asymptotic stable [5, 6]. If  $[C, A]$  is just detectable, that is, modes such that  $\text{Re } \{\lambda\} \geq 0$  are observable, the optimal system is still asymptotic stable, but the unique nonnegative definite solution of (1.10) is not necessarily strictly positive [8].

In this paper we will consider the Riccati equation and the optimal regulator under the more general assumption that  $Q_1$  is an arbitrary nonnegative definite symmetric matrix. It will be shown that the observability or detectability condition may be relaxed, and that the Riccati equation has some very nice unexploited properties.

In Section 2 the equation (1.10) is considered from an algebraic point of view. A general form of all possible matrix solutions is proved in 2.1, and in 2.2 the Hermitian and real symmetric solutions are sorted out. These sections are generalizations of the results presented by Potter [1]. In [1] the Euler matrix is assumed to have a diagonal Jordan form, while our results hold for a general Jordan form in the case of multiple eigenvalues of the Euler matrix. Thus the possible choices of the criteria matrices  $Q_1$  and  $Q_2$  will be less restricted. The effect of noncontrollable and nonobservable modes is considered in 2.3, and in 2.4 conditions for the existence of several nonnegative definite solutions are

given. Similar to Section 2.1, Theorems 8 and 9 in Section 2.4 are generalizations of [1] to the multiple eigenvalue case.

In Section 3 we return to the optimal regulator problem, and in 3.2 new upper and lower a priori bounds for (1.5) are given. In 3.3 convergence properties are discussed and proofs are given for some special cases. Although computational results indicate that convergence holds under more general assumptions about the criteria matrices  $Q_0$ ,  $Q_1$ , and  $Q_2$ , we have not succeeded in giving a general proof of convergence. That a straightforward integration of the Riccati equation may be an unstable procedure, even in what is considered as the stable direction, is illustrated in 3.4, and it is shown that only one of the stationary solutions is a numerical stable solution. Finally, in 3.5 and 3.6 the different nonnegative definite solutions are given a physical interpretation, and the optimal control theory for linear systems with quadratic loss is generalized to cover arbitrary nonnegative definite matrices  $Q_1$ .

## 2. THE ALGEBRAIC EQUATION $A^T X + XA - XBQ_2^{-1}B^T X + Q_1 = 0$

### 2.1. General Form of the Solutions

In this section we will consider explicit expressions for the solution of the quadratic matrix equation

$$A^T X + XA - XBQ_2^{-1}B^T X + Q_1 = 0 \quad (2.1)$$

In [1] it is shown that if the  $2n \times 2n$  matrix

$$E = \begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \quad (2.2)$$

has a diagonal Jordan form, it is possible to express  $X$  in terms of the eigenvectors of  $E$ . The restriction that  $E$  must have a diagonal Jordan form may be important from a pure computational point of view, but will be shown to be an unnecessary restriction for the result to hold. We will use the notation

$$a_i = \begin{bmatrix} b_i \\ c_i \end{bmatrix}$$

for the  $2n$ -dimensional eigenvector of  $E$  corresponding to the eigenvalue  $\lambda_i$ .  $a_i$  is partitioned into two  $n$ -dimensional vectors  $b_i$  and  $c_i$  which constitute the upper and lower parts of  $a_i$ . If  $\lambda_i$  is an eigenvalue of  $E$  of multiplicity  $k$ , it will be assumed that the corresponding Jordan block has a minimal polynomial of degree  $k$  unless otherwise explicitly stated. The corresponding eigenvectors are then defined as the nontrivial solutions of

$$\begin{aligned} (E - \lambda_i I)a_1 &= 0, \\ (E - \lambda_i I)a_2 &= a_1 \\ &\vdots \\ (E - \lambda_i I)a_k &= a_{k-1}. \end{aligned} \tag{2.3}$$

$a_1, a_2, \dots, a_k$  will be called the generalized eigenvectors [2], and  $a_j$  is the eigenvector of rank  $j$  corresponding to the multiple eigenvalue  $\lambda_i$ . The eigenvectors of  $E$ , if generated according to (2.3) in the case of multiple eigenvalues, span the space  $R^{2n}$ , and the transformation

$$T^{-1}ET$$

where

$$T = [a_1, \dots, a_{2n}]$$

will bring  $E$  on Jordan form.

In the sequel we will frequently consider a collection of  $n$  eigenvectors  $a_1, \dots, a_n$  of  $E$ . Assume that  $a_i$  of rank  $\ell$  (corresponding to  $\lambda_i$  of multiplicity greater than  $\ell$ ) belongs to this collection. To simplify the notations, it will then in the following be assumed, unless otherwise explicitly stated, that all the eigenvectors (corresponding to  $\lambda_i$ ) of rank less than  $\ell$  are also included in the collection  $a_1, \dots, a_n$ . Similar to [1] we then have

### Theorem 1

Each solution of (2.1) can be expressed as

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1} \quad (2.4)$$

where

$$a_i = \begin{bmatrix} b_i \\ c_i \end{bmatrix}$$

$i = 1, \dots, n$  are eigenvectors of  $E$ . Conversely, let  $a_1, \dots, a_n$  be eigenvectors of  $E$  such that  $[b_1 \dots b_n]$  is nonsingular. Then

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$$

is a solution of (2.1).

Proof: Suppose  $X$  is a solution of (2.1) and introduce

$$G = A - BQ_2^{-1}B^T X. \quad (2.5)$$

(In the optimal control problem,  $G$  is the closed loop system matrix.)  
Premultiply with  $X$

$$XG = XA - XBQ_2^{-1}B^T X \quad (2.6)$$

and substitute in (2.1). Then

$$XG = -A^T X - Q_1. \quad (2.7)$$

Let  $S$  be a nonsingular transformation that brings  $G$  on a Jordan form  $J$ , that is,

$$S^{-1}GS = J$$

Further, let

$$R = XS$$

Then

$$\begin{aligned} G &= SJS^{-1}, \\ X &= RS^{-1} \end{aligned} \quad (2.8)$$



Substitute into (2.6) and (2.7).

$$SJ = AS - BQ_2^{-1}B^T R,$$

$$RJ = -A^T R - Q_1 S,$$

or

$$\begin{bmatrix} S \\ R \end{bmatrix} [J] = \begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} S \\ R \end{bmatrix} = E \begin{bmatrix} S \\ R \end{bmatrix}. \quad (2.9)$$

Let  $a_1, \dots, a_n$  be the columns of the  $2n \times n$  matrix

$$\begin{bmatrix} S \\ R \end{bmatrix}.$$

$J$  consists of the eigenvalues of  $G$ , and if  $\lambda_i$  is an eigenvalue of rank one we then have

$$a_i \lambda_i = E a_i$$

and then  $\lambda_i$  is also an eigenvalue of  $E$ , and  $a_i$  is the corresponding eigenvector. Now let  $\lambda_i$  be of rank  $k > 1$ . (2.9) then yields

$$a_i \lambda_i = E a_i,$$

$$a_i + \lambda_i a_{i+1} = E a_{i+1}$$

$$\vdots$$

$$a_{i+k-2} + \lambda_i a_{i+k-1} = E a_{i+k-1},$$

or

$$(E - \lambda_i I) a_i = 0,$$

$$(E - \lambda_i I) a_{i+1} = a_i$$

$$\vdots$$

$$(E - \lambda_i I) a_{i+k-1} = a_{i+k-2}. \quad (2.10)$$

Since  $S$  is assumed nonsingular,  $a_i, i=1 \dots n$ , cannot be identical to the null vector, and thus the system (2.10) must have nontrivial solutions.

But this holds if and only if  $\lambda_i$  is an eigenvalue of multiplicity  $k$  to  $E$ , and then  $a_1, \dots, a_{i+k-1}$  are the corresponding generalized eigenvectors of  $E$  [2]. Then the columns of the composed matrix

$$\begin{bmatrix} S \\ R \end{bmatrix}$$

constitute the eigenvectors of  $E$ .

Finally from (2.8) follows

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}.$$

The extension to nondiagonal Jordan forms obviously restricts the possibilities for composing a solution out of  $2n$  arbitrary eigenvectors. Suppose  $\lambda_i$  is an eigenvalue of  $E$  with multiplicity  $k$ . If the generalized eigenvector  $a_{i+k-1}$  of rank  $k$  constitute one column in the matrix

$$\begin{bmatrix} S \\ R \end{bmatrix}$$

then the eigenvectors  $a_1, \dots, a_{i+k-2}$  with rank  $1, \dots, k-1$  must also be columns in

$$\begin{bmatrix} S \\ R \end{bmatrix}.$$

Consequently the a priori upper limit for the possible number of solutions of (2.1) is larger when  $E$  is assumed to have a diagonal Jordan form.

For the sake of simplicity we have assumed the eigenvectors in

$$\begin{bmatrix} S \\ R \end{bmatrix}$$

to appear in increasing rank. To prove that the order is nonessential, let the solution  $X$  be composed in the following way

$$X = [c_1 \dots c_i c_j \dots c_n][b_1 \dots b_i b_j \dots b_n]^{-1}$$

and assume that

$$[b_1 \dots b_i b_j \dots b_n]^{-1} = \begin{bmatrix} d_1 \\ \vdots \\ d_i \\ d_j \\ \vdots \\ d_n \end{bmatrix}$$

where  $d_k, k = 1 \dots n$ , are  $n$ -dimensional row vectors.

It is easy to verify that

$$[b_1 \dots b_j b_i \dots b_n]^{-1} = \begin{bmatrix} d_1 \\ \vdots \\ d_j \\ d_i \\ \vdots \\ d_n \end{bmatrix}$$

and the solutions will then be the same since

$$[c_1 \dots c_i c_j \dots c_n][b_1 \dots b_i b_j \dots b_n]^{-1} = [c_1 \dots c_j c_i \dots c_n][b_1 \dots b_j b_i \dots b_n]^{-1}.$$

The second half of the theorem is proved by carrying out the steps above in reverse order, which completes the proof of Theorem 1. ■

The restriction imposed by a nondiagonal Jordan form is illustrated in the following example: Let

$$A = \begin{pmatrix} -3 & 2 \\ -2 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad Q_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad Q_2 = (1).$$

The eigenvalues of  $E$  are  $+1, +1, -1$  and  $-1$ , and the corresponding eigenvectors are

$$a_{\lambda=1}^1 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ -2 \end{pmatrix}, \quad a_{\lambda=1}^2 = \begin{pmatrix} -1 \\ -3/2 \\ 1 \\ 0 \end{pmatrix}, \quad a_{\lambda=-1}^1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad a_{\lambda=-1}^2 = \begin{pmatrix} 1 \\ 3/2 \\ 0 \\ 0 \end{pmatrix}.$$

Suppose  $a_{\lambda=1}^1$  and  $a_{\lambda=-1}^2$  are combined Then

$$X = \begin{pmatrix} 2 & 0 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 3/2 \end{pmatrix}^{-1} = \begin{pmatrix} -6 & 4 \\ 6 & -4 \end{pmatrix}.$$

However,  $X$  does not satisfy the equation

$$A^T X + XA - XBQ_2^{-1} B^T X + Q_1 = 0$$

and thus is not a solution.

From the proof of Theorem 1 we extract the following properties of the closed loop system matrix  $G$ .

### Corollary 1

Let

$$a_i = \begin{bmatrix} b_i \\ c_i \end{bmatrix}, \quad i = 1 \dots n,$$

be eigenvectors of

$$E = \begin{bmatrix} A & -BQ_2^{-1} B^T \\ -Q_1 & -A^T \end{bmatrix}$$

corresponding to  $\lambda_1, \dots, \lambda_n$ . If  $X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$  is a solution of (2.1), then  $\lambda_1, \dots, \lambda_n$  are eigenvalues of  $A - BQ_2^{-1} B^T X$  and  $b_1, \dots, b_n$  are the corresponding eigenvectors.

Proof. The corollary follows immediately from the fact that  $J$  is the Jordan form of  $A - BQ_2^{-1}B^T X$  and  $S = [b_1 \dots b_n]$  is the transformation matrix. ■

Since the matrices  $A$ ,  $B$ ,  $Q_1$ , and  $Q_2$  are assumed to be real, it is trivial that the eigenvalues of  $E$  are symmetric with respect to the real axis. But it is easy to prove that they are symmetric with respect to the imaginary axis too [3].

Then, if  $\lambda$  is an eigenvalue of  $E$ ,  $\bar{\lambda}$  ( $\bar{\lambda}$  is the complex conjugate of  $\lambda$ ),  $-\lambda$ , and  $-\bar{\lambda}$  are eigenvalues of  $E$  too. If  $E$  has no pure imaginary eigenvalues, it is then possible to find  $n$  eigenvalues with negative real parts, and, provided that  $[b_1 \dots b_n]^{-1}$  exists, it is possible to find a solution  $X$  of (2.1) such that the closed loop system matrix  $A - BQ_2^{-1}B^T X$  is asymptotic stable.

## 2.2. Hermitian and Real Symmetric Solutions

Next we concentrate upon those solutions  $X$  of (2.1) which have the property that they are Hermitian. The following theorem is a generalization of the theorem given in [1] for the diagonal Jordan form.

### Theorem 2

Let  $a_1, \dots, a_n$  be eigenvectors of  $E$  corresponding to eigenvalues  $\lambda_1, \dots, \lambda_n$ , and assume that  $[b_1 \dots b_n]^{-1}$  exists. If  $\bar{\lambda}_j \neq -\lambda_k$ ,  $1 \leq j, k \leq n$ , then

$$X = [c_1 \dots c_n] [b_1 \dots b_n]^{-1}$$

is Hermitian.

Proof. The proof is a generalization of the proof in [1] to the non-diagonal Jordan case. Let

$$P = [b_1 \dots b_n]^* [c_1 \dots c_n] \quad (2.11)$$

where  $[b_1 \dots b_n]^*$  is the adjoint of  $[b_1 \dots b_n]$ . Then

$$X = \{[b_1 \dots b_n]^{-1}\}^* P \{[b_1 \dots b_n]^{-1}\}$$

and it remains to prove that  $P$  is Hermitian. Let  $T$  be the  $2n \times 2n$  matrix

$$T = \begin{bmatrix} O_n & I_n \\ -I_n & O_n \end{bmatrix},$$

where  $O_n$  is the  $n \times n$  null matrix. It is then easily verified that

$$E^T T + TE = 0.$$

From (2.11) we have

$$p_{jk} = b_j^* c_k$$

and

$$p_{jk} - \bar{p}_{kj} = b_j^* c_k - c_j^* b_k = a_j^* T a_k.$$

Assume that  $(\bar{\lambda}_j + \lambda_k) \neq 0$ . Then

$$p_{jk} - \bar{p}_{kj} = (\bar{\lambda}_j + \lambda_k)^{-1} (\bar{\lambda}_j a_j^* T a_k + \lambda_k a_j^* T a_k). \quad (2.12)$$

If  $E$  is assumed to have a general block diagonal Jordan form,  $\bar{\lambda}_j a_j^*$  does not necessarily equal  $a_j^* E^T$  since  $a_j$  may be of rank larger than one. Then consider the different possibilities that may occur:

A.  $\bar{\lambda}_j \neq -\lambda_k$  and  $E a_j = \lambda_j a_j$ ,  $E a_k = \lambda_k a_k$ . Then

$$\begin{aligned} p_{jk} - \bar{p}_{kj} &= (\bar{\lambda}_j + \lambda_k)^{-1} (a_j^* E^T T a_k + a_j^* T E a_k) \\ &= (\bar{\lambda}_j + \lambda_k)^{-1} a_j^* (E^T T + TE) a_k = 0 \end{aligned}$$

and thus  $p_{jk} = \bar{p}_{kj}$ .

B.  $\bar{\lambda}_j \neq -\lambda_k$  and  $Ea_j = \lambda_j a_j$ , but  $(E - \lambda_k I)a_k = a_{k-1}$ .  $\lambda_k$  then is a multiple eigenvalue, and a generalized eigenvector of rank larger than one is used to determine the solution X.

$$\begin{aligned} p_{jk} - \bar{p}_{kj} &= (\bar{\lambda}_j + \lambda_k)^{-1} (a_j^* E^T Ta_k + a_j^* TEa_k - a_j^* Ta_{k-1}) \\ &= (\bar{\lambda}_j + \lambda_k)^{-1} (a_j^* (E^T T + TE)a_k - a_j^* Ta_{k-1}) \\ &= -(\bar{\lambda}_j + \lambda_k)^{-1} a_j^* Ta_{k-1}. \end{aligned}$$

Analogous to (2.12) this is equivalent to

$$p_{jk} - \bar{p}_{kj} = -(\bar{\lambda}_j + \lambda_k)^{-2} (\bar{\lambda}_j a_j^* Ta_{k-1} + \lambda_k a_j^* Ta_{k-1}).$$

If  $a_{k-1}$  is of rank one, then, according to A,  $p_{jk} = \bar{p}_{kj}$ . If the rank is higher than one, the procedure above is repeated, say  $m$  times, until

$$p_{jk} - \bar{p}_{kj} = (-1)^m (\bar{\lambda}_j + \lambda_k)^{-m} a_j^* Ta_{k-m}$$

and  $a_{k-m}$  is of rank one. Then  $p_{jk} = \bar{p}_{kj}$  according to case A.

C.  $\bar{\lambda}_j \neq -\lambda_k$  and  $(E - \lambda_j I)a_j = a_{j-1}$ ,  $(E - \lambda_k I)a_k = a_{k-1}$ . Both  $\lambda_j$  and  $\lambda_k$  are assumed to be multiple eigenvalues, and  $a_j, a_k$  are generalized eigenvectors both of rank larger than one. Then

$$p_{jk} - \bar{p}_{kj} = (\bar{\lambda}_j + \lambda_k)^{-1} [(a_j^* E^T - a_{j-1}^*) Ta_k + a_j^* T(Ea_k - a_{k-1})],$$

which yields

$$p_{jk} - \bar{p}_{kj} = -(\bar{\lambda}_j + \lambda_k)^{-1} (a_{j-1}^* Ta_k + a_j^* Ta_{k-1}). \quad (2.13)$$

If  $a_{j-1}$  or  $a_{k-1}$  is of rank one, the corresponding term in (2.13) will vanish according to B or A. If both have larger rank, the procedure is repeated:

$$p_{jk} - \bar{p}_{kj} = (-1)^2 (\bar{\lambda}_j + \lambda_k)^{-2} (a_{j-2}^* Ta_k + a_{j-1}^* Ta_{k-1} + a_{j-1}^* Ta_{k-1} + a_j^* Ta_{k-2}).$$

The rank of one of the eigenvectors in the product  $a_{j-1}^* T a_{k-m}$  is lowered by one in each step, and finally a situation arises in which either A or B can be applied. Then  $p_{jk} = \bar{p}_{kj}$ , and this finally proves that X is Hermitian if  $\bar{\lambda}_j = -\lambda_k$ ,  $1 \leq j, k \leq n$ . ■

Now let  $\lambda_r$  be an eigenvalue of multiplicity  $r$ , and  $a_1, \dots, a_r$  the corresponding eigenvectors. Then any attempt to include  $a_1, \dots, a_k$  but not  $a_{k+1}, \dots, a_r$ ,  $1 \leq k < r$ , in the solution will violate the condition  $\bar{\lambda}_j = -\lambda_k$ . The reason for this is as follows: If we have selected  $a_1, \dots, a_k$  we cannot make use of any of the  $r$  eigenvectors corresponding to  $-\bar{\lambda}_r$ . From the remaining  $2n - 2r$  eigenvectors we must choose either the one corresponding to  $\lambda_1$  or the one corresponding to  $-\bar{\lambda}_1$ , but not both. Then it only remains  $(n - r)$  possible ways to choose  $n - k$  eigenvectors. But  $n - r < n - k$  since it was assumed that  $k < r$ .

Summarizing, we then conclude that the only possibility to satisfy the sufficient condition for X to be Hermitian is to include all eigenvectors corresponding to  $\lambda_r$  or all eigenvectors corresponding to  $-\bar{\lambda}_r$ .

In the next section, conditions will be given that allow both  $\lambda_j$  and  $-\bar{\lambda}_j$  to be included in a Hermitian solution.

If E has  $2n$  distinct eigenvalues, and if  $[b_1 \dots b_n]^{-1}$  exists for all combinations of eigenvectors the theorem states that among the

$$\binom{2n}{n}$$

possible solutions X, at least  $2^n$  are Hermitian. In the case of multiple eigenvalues of E, more complex combinatorial problems are obtained.

In the optimal control problem, only real solutions of (2.1) are of interest, since the system matrices A, B and criteria matrices  $Q_1, Q_2$  are assumed real. Moreover, since  $Q_1$  and  $Q_2$  are assumed

symmetric we will next concentrate upon real symmetric solutions of (2.1).



### Theorem 3

Let

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$$

be a solution of (2.1). Then  $X$  is real if and only if either

- i) all eigenvectors  $a_1, \dots, a_n$  are real, or
- ii) both  $a_i$  of rank  $k$  corresponding to  $\lambda_i$  and  $\bar{a}_i$  of rank  $k$  corresponding to  $\bar{\lambda}_i$  are included in the solution.

**Proof.** (i) is trivial. To prove (ii), let  $a_i$  and  $\bar{a}_i$  be included in the solution. Then

$$X = [c_1 \dots c_i \dots \bar{c}_i \dots c_n][b_1 \dots b_i \dots \bar{b}_i \dots b_n]^{-1}$$

and

$$\bar{X} = [c_1 \dots \bar{c}_i \dots c_i \dots c_n][b_1 \dots \bar{b}_i \dots b_i \dots b_n]^{-1}.$$

Since the order of the eigenvectors is immaterial, it follows that

$$X = \bar{X}$$

and thus  $X$  is a real solution. This proves the sufficiency. To prove the necessity, consider the closed loop system matrix

$$G = A - BQ_2^{-1}B^T X.$$

$G$  is real if  $X$  is real, and then the eigenvalues of  $G$  are real or complex conjugated. But according to Corollary 1 of Section 2.1, the eigenvalues of  $G$  will be those eigenvalues that correspond to the eigenvectors used in the solution  $X$ . This finally proves that a necessary condition for  $X$  to be real is that (ii) holds. ■

Combining Theorems 2 and 3 will finally give sufficient conditions for symmetry of a real solution  $X$  of (2.1).

### 2.3. Nonobservable and Noncontrollable Modes

Now consider the optimal control problem defined in Section 1. Since the criteria matrices  $Q_1$  and  $Q_2$  are symmetric nonnegative and symmetric positive definite, we must look for a symmetric and nonnegative definite solution of (2.1) [5]. It is well known [5, 6] that, if the pair  $[C, A]$ , where  $Q_1 = C^T C$ , is completely observable, the stationary solution of (1.5) will be positive definite and the optimal system is asymptotic stable. In that case, there is only one nonnegative definite solution of (2.1) [7]. In [8] Wonham makes a generalization, and proves that detectability of the pair  $[C, A]$  is sufficient for the optimal system to be asymptotic stable. In this case the stationary solution is no longer necessarily positive definite, but may only be nonnegative definite.

We will now generalize further, and consider arbitrary real, symmetric and nonnegative definite matrices  $Q_1$ . Thus  $A$  is allowed to have unstable modes nonobservable in  $[C, A]$ . (In the sequel we will use the notation  $[Q_1, A]$ .)

In particular we will assume that the eigenvalues of  $A$  are distinct. It is then possible to diagonalize  $A$ , and simple definitions of observability and controllability may be used.

It will also be assumed that the nonobservable modes of  $A$  are real. However, it is easily verified that the following statements will also hold for complex nonobservable modes. In that case a nonobservable mode must always be considered together with its complex conjugate mode, and thus the assumption is made to simplify the notations. However, we will allow for a distinct nonobservable mode  $\lambda_i = 0$ .

Then introduce the following definition of observability.

#### Definition

Let  $T$  be a nonsingular linear transformation such that

$$TAT^{-1} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix},$$

where  $\lambda_1, \dots, \lambda_n$  are distinct eigenvalues of  $A$ . The mode  $\lambda_i$  is then

an observable mode of the pair  $[Q_1, A]$  if and only if the  $i$ -th column of the matrix  $CT^{-1}$ , where  $Q_1 = C^T C$ , has at least one element not identical zero.

If  $\lambda_i$  is a nonobservable mode of  $[Q_1, A]$ , and  $x_i$  is the corresponding eigenvector of  $A$ , it follows from the definition that  $Cx_i = 0$  and  $Q_1 x_i = C^T Cx_i = 0$ . We then have the following theorem.

#### Theorem 4

If  $\lambda_i$  is a nonobservable mode of the pair  $[Q_1, A]$ , then  $\lambda_i$  is an eigenvalue of

$$E = \begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix}$$

and the corresponding eigenvector of rank one is

$$\begin{bmatrix} x_i \\ o_n \end{bmatrix}.$$

( $o_n$  is the  $n$ -dimensional null vector.)

**Proof.** The proof is a straightforward application of the definition of eigenvalues and eigenvectors.

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} x_i \\ o_n \end{bmatrix} = \begin{bmatrix} Ax_i \\ -Q_1 x_i \end{bmatrix} = \lambda_i \begin{bmatrix} x_i \\ o_n \end{bmatrix}. \quad \blacksquare$$

Controllability of the pair  $[A, B]$  is defined in a similar way.

### Definition

Let  $T$  be a nonsingular linear transformation such that

$$TAT^{-1} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

where  $\lambda_1, \dots, \lambda_n$  are distinct eigenvalues of  $A$ . The mode  $\lambda_i$  is then a controllable mode of the pair  $[A, B]$  if and only if the  $i$ -th row of the matrix  $TB$  has at least one element not identical zero.

If  $\lambda_i$  is a noncontrollable mode of  $[A, B]$  and  $y_i^T$  the corresponding lefthand eigenvector of  $A$ , then, analogous to nonobservability, the definition yields  $y_i^T B = 0$  or  $B^T y_i = 0$ . The following theorem, similar to Theorem 4, is then easy to prove.

### Theorem 5

If  $\lambda_i$  is a noncontrollable mode of the pair  $[A, B]$ , then  $-\lambda_i$  is an eigenvalue of

$$E = \begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix}$$

and the corresponding eigenvector of rank one is

$$\begin{bmatrix} o_n \\ y_i \end{bmatrix}.$$

Proof.

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} o_n \\ y_i \end{bmatrix} = \begin{bmatrix} -BQ_2^{-1}B^T y_i \\ -A^T y_i \end{bmatrix} = -\lambda_i \begin{bmatrix} o_n \\ y_i \end{bmatrix}. \quad \blacksquare$$

In the following it will prove to be simplifying to have conditions available for the existence of eigenvalues of  $E$  on the imaginary axis. This can be established from Theorems 4 and 5 and from the definitions of controllability and observability.

Theorem 6

$\lambda_j = i\alpha$ ,  $\alpha \in \mathbb{R}$ , is an eigenvalue of  $E$  if and only if  $\lambda_j$  and  $\bar{\lambda}_j$  are either nonobservable modes of  $[Q_1, A]$  or noncontrollable modes of  $[A, B]$  (or both).

Proof: The sufficiency is proved in theorems 4 and 5. To prove the necessity, consider the definition of eigenvectors. There is then at least one nontrivial eigenvector of rank one satisfying

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} x_j \\ y_j \end{bmatrix} = i\alpha \begin{bmatrix} x_j \\ y_j \end{bmatrix}$$

and

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} \bar{x}_j \\ \bar{y}_j \end{bmatrix} = -i\alpha \begin{bmatrix} \bar{x}_j \\ \bar{y}_j \end{bmatrix}$$

Multiply the first relation from the left with  $[y_j^*, x_j^*]$  and the second relation from the left with  $[y_j^T, x_j^T]$ . Then

$$y_j^*Ax_j - y_j^*BQ_2^{-1}B^T y_j - x_j^*Q_1 x_j - x_j^*A^T y_j = i\alpha(x_j^*x_j + y_j^*y_j)$$

and

$$y_j^T A \bar{x}_j - y_j^T B Q_2^{-1} B^T \bar{y}_j - x_j^T Q_1 \bar{x}_j - x_j^T A^T \bar{y}_j = -i\alpha(x_j^T \bar{x}_j + y_j^T \bar{y}_j)$$

A straightforward addition then yields

$$-2y_j^*BQ_2^{-1}B^T y_j - 2x_j^*Q_1 x_j = 0$$

or

$$-y_j^* B Q_2^{-1} B^T y_j - x_j^* C^T C x_j = 0$$

But  $Q_2$  is positive definite, and then  $x_j$  and  $y_j$  must satisfy

$$C x_j = 0$$

$$B^T y_j = 0$$

From the definition of eigenvectors of  $E$  then follows that  $x_j$  and  $y_j$  must also satisfy

$$A x_j = i\alpha x_j$$

$$-A^T y_j = i\alpha y_j$$

Since the eigenvector is not identically zero, either  $x_j \neq 0$  or  $y_j \neq 0$ .

If  $x_j \neq 0$ , the definition of observability implies that  $\lambda_j = i\alpha$  is a non-observable mode of  $[Q_1, A]$ , and then  $\bar{\lambda}_j = -i\alpha$  is also a nonobservable mode of  $[Q_1, A]$ . Similarly, if  $y_j \neq 0$ , the definition of controllability implies that  $\lambda_j$  and  $\bar{\lambda}_j$  are noncontrollable modes of  $[A, B]$ . ■

### Corollary 2

$\lambda_i = 0$  is an eigenvalue of  $E$  if and only if  $\lambda_i = 0$  is either a non-observable mode of  $[Q_1, A]$  or a noncontrollable mode of  $[A, B]$ .

### Corollary 3

Let  $\lambda_i = 0$  be a distinct noncontrollable (nonobservable) mode of  $[A, B]$  ( $[Q_1, A]$ ). Then there are two eigenvectors of  $E$  of rank one, corresponding to  $\lambda_i = 0$ , if and only if  $\lambda_i = 0$  is also a nonobservable (noncontrollable) mode of  $[Q_1, A]$  ( $[A, B]$ ).

**Proof:** In the proof of theorem 6, it was shown that a nontrivial eigenvector of  $E$ , corresponding to  $\lambda_i = 0$ , satisfies

$$A x_i = 0$$

$$C x_i = 0$$

and

$$A^T y_i = 0$$

$$B^T y_i = 0$$

Let

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

be distinct eigenvectors of  $E$  of rank one, corresponding to  $\lambda_i = 0$ . Since  $\lambda_i$  is assumed noncontrollable, one of the  $x_i$ 's, say  $x_1$ , satisfy  $x_1 = 0$  according to Theorem 5. But if  $\lambda_i$  is a distinct eigenvalue of  $A$  (and thus of  $A^T$ ), any two solutions  $y_i$  of  $A^T y_i = 0$  are linearly dependent, that is,  $y_2 = \alpha y_1$ . Then both eigenvectors are of rank one only if  $x_2 \neq 0$ , and thus  $\lambda_i = 0$  is also a nonobservable mode of  $[Q_1, A]$ . The alternative formulation is proved in the same way, and the reverse follows from Theorems 4 and 5. ■

It is interesting to notice, that the requirement of controllability of modes such that  $\text{Re}\{\lambda_i\} \geq 0$ , can now be justified with simple algebraic considerations.

Suppose there exists a noncontrollable mode of  $[A, B]$ , such that  $\text{Re}\{\lambda_i\} > 0$ . Then according to Theorem 5,  $-\lambda_i$  is an eigenvalue of  $E$ , and the corresponding eigenvector is

$$\begin{bmatrix} 0_n \\ c_i \end{bmatrix}$$

There is then no solution  $X$  of (2.1) such that  $A - BQ_2^{-1}B^T X$  is asymptotic stable, since this would require the inclusion of the eigenvector

$$\begin{bmatrix} 0_n \\ c_i \end{bmatrix}$$

in which case  $[b_1 \dots b_n]$  is singular.

If  $\lambda_i = 0$  is a noncontrollable mode, the situation is similar. Assume for the sake of simplicity that  $E$  has  $n-1$  eigenvalues satisfying  $\text{Re}\{\lambda_j\} < 0$ . Then  $\lambda_i = 0$  is an eigenvalue of  $E$  of multiplicity two.

If the  $\lambda_i$ -block of the Jordan form of  $E$  is nondiagonal, there is only one eigenvector of rank one corresponding to  $\lambda_i = 0$ , and then there is no solution  $X$  of (2.1) such that the eigenvalues of  $A - BQ_2^{-1} B^T X$  satisfy  $\text{Re}\{\lambda_j\} \leq 0$ . The only possibility to satisfy this condition is that  $\lambda_i = 0$  is also a nonobservable mode of  $[Q_1, A]$  (Corollary 3), in which case the eigenvector

$$\begin{bmatrix} b_i \\ 0 \\ \vdots \\ 0 \\ n \end{bmatrix}$$

may be used.

When formulating the optimization problem, the ultimate goal is in general an asymptotic stable closed loop system. It is thus natural to claim controllability of unstable or purely imaginary modes.

Having verified that noncontrollable and nonobservable modes imply structural properties of the corresponding eigenvectors, it is now possible to relax the sufficient conditions for symmetry proved in Section 2.2. As before it is sufficient to prove that

$$P = [b_1 \dots b_n]^* [c_1 \dots c_n]$$

is symmetric (Hermitian). For  $\lambda_j \neq -\bar{\lambda}_k$  it was proved in Theorem 2 that  $p_{jk} = \bar{p}_{kj}$ . We will now prove that although  $\lambda_j = -\bar{\lambda}_k$ , the solution may still be symmetric (Hermitian).

Consider first the case when  $\lambda_j = -\lambda_k$  and both  $\lambda_j$  and  $\lambda_k$  are real nonobservable modes of  $[Q_1, A]$ . Then

$$p_{jk} - \bar{p}_{kj} = b_j^T c_k - c_j^T b_k = 0$$

since  $c_j = c_k = 0_n$ , and thus  $P$  is still Hermitian.

Then consider the case when  $\lambda_j$  is a real nonobservable mode of  $[Q_1, A]$ , and the criteria matrices  $Q_1$  and  $Q_2$  are chosen so that



$\lambda_k = -\lambda_j$  is an eigenvalue of  $E$ , that is,  $\lambda_k$  is not due to the symmetry property of the eigenvalues of  $E$ . Since the eigenvalues of a matrix are continuous functions of the entries, it then follows that both  $\lambda_j$  and  $\lambda_k$  will be multiple. Since  $\lambda_j$  is nonobservable,  $c_j = o_n$ , and

$$p_{jk} - \bar{p}_{jk} = b_j^T c_k - c_j^T b_k = b_j^T c_k$$

With the assumption that there is only one eigenvector of rank one corresponding to  $\lambda_k$ , it can now be proved that  $b_j^T c_k = 0$ , and thus that  $P$  is still Hermitian.

Let  $T$  be a nonsingular linear transformation such that  $TAT^{-1}$  is diagonal. Then

$$T^{-1} = [x_1, \dots, b_j, \dots, x_n].$$

Introduce the  $2n \times 2n$  matrix

$$V = \begin{bmatrix} T & O_n \\ O_n & (T^{-1})^T \end{bmatrix}$$

where  $O_n$  denotes the  $n \times n$  null matrix. As  $V$  is nonsingular the eigenvalues of

$$\tilde{E} = VEV^{-1}$$

are the same as those of  $E$ , and the corresponding eigenvectors are

$$\tilde{a}_i = Va_i. \quad (2.14)$$

This holds for generalized eigenvectors too.

Carrying out the transformation  $VEV^{-1}$  we have

$$\tilde{E} = \begin{bmatrix} TAT^{-1} & -TBQ_2^{-1}B^T T^T \\ -(T^{-1})^T Q_1 T^{-1} & -(T^{-1})^T A^T T^T \end{bmatrix},$$

which reduces to

$$\tilde{E} = \left[ \begin{array}{c|c} \begin{matrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_j & \\ & & & \ddots \\ & & & & \lambda_n \end{matrix} & \\ \hline & R \\ \hline P & \begin{matrix} -\lambda_1 & & & \\ & \ddots & & \\ & & -\lambda_j & \\ & & & \ddots \\ & & & & -\lambda_n \end{matrix} \end{array} \right], \quad (2.15)$$

where  $R = -TBQ_2^{-1}B^T T^T$  and  $P = -(T^{-1})^T Q_1 T^{-1} = -(CT^{-1})^T (CT^{-1})$ . As  $\lambda_j$  is a nonobservable mode of  $[Q_1, A]$ , the  $j$ -th column of  $CT^{-1}$  equals zero, and hence both the  $j$ -th column and the  $j$ -th row of  $P$  equal zero.

Now consider that  $\lambda_k = -\lambda_j$  is a multiple eigenvalue of  $E$  and  $\tilde{E}$ , and introduce

$$\tilde{a}_k^1 = \begin{bmatrix} \tilde{b}_1 & 1 \\ \vdots & \vdots \\ \tilde{b}_n & 1 \\ \tilde{c}_1 & 1 \\ \vdots & \vdots \\ \tilde{c}_j & 1 \\ \vdots & \vdots \\ \tilde{c}_n & 1 \end{bmatrix}$$

as the corresponding eigenvector of rank one.

Then  $\tilde{E}\tilde{a}_k^1 = \lambda_k \tilde{a}_k^1 = -\lambda_j \tilde{a}_k^1$ . The eigenvector of rank two,  $\tilde{a}_k^2$ , is determined through



$$b_j^T c_k^{1-1} = \tilde{c}_j^{1-1} = 0.$$

Notice that we have only proved symmetry for the case where  $\lambda_j$  or  $\lambda_k$  are due to nonobservable modes of  $[Q_1, A]$ . For eigenvalues of E not due to nonobservable modes we can still not make any statements about the symmetry except for the sufficient conditions of Theorem 2 ( $\lambda_j \neq -\lambda_k$ ). Also notice that we proved symmetry under the assumption that there is only one eigenvector corresponding to  $\lambda_k$  of rank one.

The results are summarized in the following theorem.

### Theorem 7

Suppose  $\lambda_j$  is a nonobservable mode of  $[Q_1, A]$ . Let  $a_1, \dots, a_n$  be eigenvectors of E corresponding to  $\lambda_1, \dots, \lambda_j, \lambda_k, \dots, \lambda_n$  and assume that  $[b_1 \dots b_n]^{-1}$  exists. If the sufficient conditions of Theorem 2 are satisfied except for  $\lambda_k = -\lambda_j$ , then

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$$

is still Hermitian if either

- (i)  $\lambda_k$  is a nonobservable mode of  $[Q_1, A]$  and only the eigenvector of rank one corresponding to  $\lambda_k$  is included in the solution, or if
- (ii) there is only one eigenvector corresponding to  $\lambda_k$  of rank one, and the number of generalized eigenvectors corresponding to  $\lambda_k$  included in the solution is less than or equal to  $m-1$ , where  $m$  is the multiplicity of  $\lambda_k$ .

The theorem is illustrated in the following example:

$$A = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} 0 & 0 \\ 0 & 3 \end{pmatrix}, \quad Q_2 = (1),$$

The eigenvalues of E are

$$\lambda_1 = 2, \text{ nonobservable mode of } [Q_1, A];$$

$$\lambda_2 = -2, \lambda_2 = -\lambda_1;$$

$\lambda_3 = -2$ , due to the specific choice of  $Q_1$  and  $Q_2$ ;  $\lambda_3$  is a continuous function of the elements of  $Q_1$  and  $Q_2$ ;

$$\lambda_4 = 2, \lambda_4 = -\lambda_3.$$

Eigenvectors of rank one corresponding to  $+2$  and  $-2$  are

$$a_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad a_2 = \begin{pmatrix} 1 \\ 4 \\ 0 \\ 4 \end{pmatrix}.$$

Then

$$X = \begin{pmatrix} 0 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

which is a symmetric solution.

#### 2.4. Nonnegative Definite Solutions

Among the symmetric solutions we will now look for solutions with the property that they are nonnegative definite. Since the criteria matrices  $Q_1$  and  $Q_2$  are nonnegative respectively positive definite, this is a necessary condition for  $X$  to be a solution of the optimal control problem [5]. Then what choice of  $n$  eigenvectors  $a_1, \dots, a_n$  will cause  $X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$  to be nonnegative definite or positive definite?

The following theorem states necessary conditions for a positive definite solution.

##### Theorem 8

Let  $a_1, \dots, a_n$  be eigenvectors of  $E$  corresponding to  $\lambda_1, \dots, \lambda_n$ .

Assume that

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$$

is real, symmetric and positive definite. Then  $\text{Re}\{\lambda_i\} < 0, i = 1, \dots, n$ .

Proof. Consider the closed loop system matrix

$$G = A - BQ_2^{-1}B^T X.$$

As  $X$  is a solution of (2.1) it is easy to verify that

$$G^T X + XG = -(Q_1 + XBQ_2^{-1}B^T X).$$

From Lyapunov stability theory, it follows that  $\text{Re}\{\lambda_i\} \leq 0$ , and it then remains to prove that  $\text{Re}\{\lambda_i\} \neq 0$ . But according to Theorem 6,  $\text{Re}\{\lambda_i\} = 0$  implies that  $\lambda_i$  is either a noncontrollable or a nonobservable mode (or both). However, noncontrollability contradicts the assumption that  $X$  exists, since eigenvectors with the structure

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ c_i \end{bmatrix}$$

in that case should be included in the solution. For similar reasons, nonobservability contradicts the assumption that  $X$  is positive definite. The eigenvector then has the structure

$$\begin{bmatrix} b_i \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

and thus

$$X = [c_1 \dots 0 \dots c_n] [b_1 \dots b_i \dots b_n]^{-1}$$

is singular. ■

#### Corollary 4

Let  $\lambda_i \leq 0$  be a nonobservable mode of  $[Q_1, A]$ . Then there is no positive definite solution of (2.1).

Proof: If  $\lambda_i = 0$  is a nonobservable mode of  $[Q_1, A]$ ,  $\lambda_i = 0$  is also an eigenvalue of  $E$ . Thus there are at most  $n-1$  eigenvalues of  $E$  such that  $\text{Re}\{\lambda_j\} < 0$ , which excludes the possibility of a positive definite solution of (2.1). If  $\lambda_i < 0$ , a positive definite solution  $X$  must include the corresponding eigenvector  $a_i$  with the structure

$$\begin{bmatrix} b_i \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

and thus  $X$  is singular, which contradicts the assumption that  $X$  is positive definite. ■

Notice that since there is only one way to select  $n$  eigenvalues of  $E$  with  $\text{Re}\{\lambda_i\} < 0$ , Theorem 8 implies that (2.1) can never have more than one positive definite solution.

Sufficient conditions for the existence of a nonnegative definite solution of (2.1), are given in [1] for the case when  $E$  has a diagonal Jordan form. The following theorem is a generalization of [1], and covers nondiagonal Jordan forms of  $E$ .

### Theorem 9

Suppose that  $Q_1$  and  $Q_2$  are nonnegative definite respectively positive definite real symmetric matrices, and let  $a_1, \dots, a_n$  be eigenvectors of  $E$  corresponding to  $\lambda_1, \dots, \lambda_n$ . If  $\text{Re}\{\lambda_i\} < 0$ ,  $i = 1, \dots, n$ , and  $[b_1 \dots b_n]$  is nonsingular, then

$$X = [c_1 \dots c_n][b_1 \dots b_n]^{-1}$$

is real, symmetric and nonnegative definite.

Proof: That  $X$  is real and symmetric follows from Theorems 2 and 3. To prove that  $X$  is nonnegative definite, let

$$X = \{[b_1 \dots b_n]^{-1}\}^* P \{[b_1 \dots b_n]^{-1}\}$$

where

$$P = [b_1 \dots b_n]^* [c_1 \dots c_n].$$

Since  $[b_1 \dots b_n]$  is nonsingular, it is sufficient to prove that  $P$  is nonnegative definite. Introduce the  $2n \times n$  matrix  $U(t)$

$$U(t) = [e^{\lambda_1 t} a_1, \dots, e^{\lambda_n t} a_n].$$

If  $\lambda_k$  is a multiple eigenvalue of multiplicity  $r$ , then  $U(t)$  is defined as

$$\begin{aligned}
 U(t) = & \left[ e^{\lambda_1 t} a_1, \dots, e^{\lambda_k t} a_k, e^{\lambda_k t} (a_{k+1} + a_k t), \right. \\
 & \dots, e^{\lambda_k t} \left( a_{k+r} + a_{k+r-1} \cdot t + \dots + \frac{a_k \cdot t^{r-1}}{(r-1)!} \right), \\
 & \left. \dots, e^{\lambda_n t} a_n \right]
 \end{aligned} \tag{2.17}$$

It is easily verified that  $U(t)$  satisfies the differential equation

$$\frac{dU(t)}{dt} = EU(t), \quad U(0) = [a_1, \dots, a_n].$$

Let  $L$  be the  $2n \times 2n$  matrix

$$L = \begin{bmatrix} O_n & I_n \\ O_n & O_n \end{bmatrix}.$$

Then

$$P = U^*(0)LU(0)$$

Further introduce

$$S(t) = -U^*(t)LU(t) + U^*(0)LU(0). \tag{2.18}$$

Since  $\operatorname{Re}\{\lambda_i\} < 0$ ,  $i = 1 \dots n$ , the definition of  $U$  implies that

$$\lim_{t \rightarrow \infty} U(t) = O_{2n \times n}$$

and thus

$$\lim_{t \rightarrow \infty} S(t) = P.$$

(2.18) is equivalent to



$$\begin{aligned}
S(t) &= - \int_0^t \frac{d}{ds} [U^*(s)LU(s)] ds \\
&= - \int_0^t [U^*(s)E^T LU(s) + U^*(s)LEU(s)] ds \\
&= - \int_0^t [U^*(s)[E^T L + LE]U(s)] ds.
\end{aligned}$$

But

$$E^T L + LE = \begin{bmatrix} -Q_1 & O_n \\ O_n & -BQ_2^{-1}B^T \end{bmatrix}$$

and then  $S(t) \geq 0$ ,  $t \geq 0$ . When  $t \rightarrow \infty$ ,  $S(t) \rightarrow P$ , and thus  $P$  is non-negative definite. ■

It is possible to relax the assumptions of Theorem 9, and include one eigenvalue  $\lambda_i = 0$ . This is proved in the following Theorem.

### Theorem 10

Assume that the conditions of Theorem 9 are satisfied except for one eigenvalue  $\lambda_i = 0$  of  $E$ . Then  $X$  is still real, symmetric and non-negative definite.

Proof: Since  $\lambda_i$  is an eigenvalue of  $E$ , and  $X$  is assumed to exist, it follows that  $\lambda_i = 0$  is a nonobservable mode of  $[Q_1, A]$ , and the corresponding eigenvector (of rank one) has the structure

$$\begin{bmatrix} b_i \\ o_n \end{bmatrix}$$

From Theorems 2 and 3 then still follows that  $X$  is real and symmetric. To prove that  $X$  is nonnegative definite, consider the definition of  $U$ ,  $L$  and  $S$  in the proof of Theorem 9. It is easily verified that

$$\lim_{t \rightarrow \infty} U(t) \dagger O_{2n \times n}$$

but

$$\lim_{t \rightarrow \infty} U^*(t)LU(t) = O_n$$

since  $c_i = 0_n$ . Thus

$$\lim_{t \rightarrow \infty} S(t) = P$$

still holds, and the last part of the proof of Theorem 9 is then still applicable. ■

In [7] it is proved that, if  $[Q_1, A]$  is completely observable, then a unique nonnegative definite solution of (2.1) exists. Moreover, this solution is positive definite. However, this is no longer true if the observability criterium is relaxed. This is illustrated below with an example.

In Theorem 11 sufficient conditions for two real symmetric solutions of (2.1) to be nonnegative definite are given. Notice that we must assume that the solutions exist, since we have not given any sufficient conditions for the nonsingularity of  $[b_1 \dots b_n]$ . However, we have not found any example contradicting the assumption that both solutions exist.

### Theorem 11

Let  $\lambda_i > 0$  be a nonobservable mode of  $[Q_1, A]$ . Assume that  $X_1$  is a solution of (2.1) corresponding to  $-\lambda_i$  and  $n-1$  eigenvalues of  $E$  with  $\text{Re}\{\lambda_j\} < 0$ . Assume that  $X_2$  is another solution with the same eigenvectors except for the one corresponding to  $-\lambda_i$ . This eigenvector is replaced by the eigenvector corresponding to  $\lambda_i$ . Then both  $X_1$  and  $X_2$  are real, symmetric and nonnegative definite.

Proof: That  $X_1$  is real, symmetric and nonnegative definite follows from Theorem 9. According to Theorems 3 and 7  $X_2$  is real and symmetric, and then it remains to prove that  $X_2$  is also nonnegative definite. To simplify the proof, we assume that the eigenvalues of  $E$  are distinct. Connecting to the proof of Theorem 9, we will prove that  $U^*(t)LU(t) \rightarrow O_n$  as  $t \rightarrow \infty$ . From the definition of  $U$  then follows<sup>†</sup>)

†) In case of multiple eigenvalues of  $E$ ,  $U$  is defined according to (2.17). The theorem may then be proved in the same way, but the notations will be more involved.

$$U^*(t)LU(t) = \begin{bmatrix} b_1^* e^{\bar{\lambda}_1 t} & & c_1^* e^{\bar{\lambda}_1 t} \\ \vdots & & \vdots \\ b_i^* e^{\bar{\lambda}_i t} & & c_i^* e^{\bar{\lambda}_i t} \\ \vdots & & \vdots \\ b_n^* e^{\bar{\lambda}_n t} & & c_n^* e^{\bar{\lambda}_n t} \end{bmatrix} \begin{bmatrix} O_n & I_n \\ O_n & O_n \end{bmatrix} \begin{bmatrix} b_1 e^{\lambda_1 t} & \dots & b_i e^{\lambda_i t} & \dots & b_n e^{\lambda_n t} \\ c_1 e^{\lambda_1 t} & \dots & c_i e^{\lambda_i t} & \dots & c_n e^{\lambda_n t} \end{bmatrix}$$

$U^*(t)LU(t)$  is an  $n \times n$  matrix, and the elements are

$$[U^*(t)LU(t)]_{kl} = b_k^* c_l e^{\bar{\lambda}_k t} e^{\lambda_l t}.$$

Since

$$P = U^*(0)LU(0) = \begin{pmatrix} b_1^* c_1 & \dots & b_1^* c_n \\ \vdots & & \vdots \\ b_n^* c_1 & \dots & b_n^* c_n \end{pmatrix}$$

is symmetric, it follows that  $U^*(t)LU(t)$  is symmetric too. For  $k \neq l$  and  $l \neq i$  the elements

$$b_k^* c_l e^{\bar{\lambda}_k t} e^{\lambda_l t} \rightarrow 0$$

as  $t \rightarrow \infty$  since  $\text{Re}\{\bar{\lambda}_k\} < 0$  and  $\text{Re}\{\lambda_l\} < 0$ . But the  $i$ -th column of  $U^*(t)LU(t)$  is identical to the zero column vector since  $c_i = 0$ .

The symmetry then implies that the  $i$ -th row equals zero too. Thus  $U^*(t)LU(t) \rightarrow 0$  as  $t \rightarrow \infty$ , and

$$S(t) = -U^*(t)LU(t) + U^*(0)LU(0) \rightarrow P.$$

It then follows from Theorem 9 that  $X_2$  is nonnegative definite. The solutions  $X_1$  and  $X_2$  are not identical, since the eigenvalues of  $G_1 = A - BQ_2^{-1}B^T X_1$  and  $G_2 = A - BQ_2^{-1}B^T X_2$  are different. This completes the proof of two different nonnegative definite solutions of (2.1). ■

The theorem is easily generalized to multiple eigenvalues  $\lambda_k$ ,  $\text{Re}\{\lambda_k\} < 0$ , and to an arbitrary number of distinct nonobservable modes of  $[Q_1, A]$ . This is illustrated in the following example:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q_2 = (1).$$

The eigenvalues of E are

$\lambda_1 = 1$ , due to the nonobservable mode of  $[Q_1, A]$ ;

$\lambda_2 = -1$ ,  $\lambda_2 = -\lambda_1$ ;

$\lambda_3 = 2$ , nonobservable mode;

$\lambda_4 = -2$ ,  $\lambda_4 = -\lambda_3$ ;

$\lambda_5 = -3$ , nonobservable mode;

$\lambda_6 = 3$ ,  $\lambda_6 = -\lambda_5$ .

Since  $\lambda_5 = -3$  is a nonobservable mode, there is no positive definite solution (Corollary 4). The eigenvectors  $a_i$  corresponding to  $\lambda_i$  are

$$a_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad a_2 = \begin{pmatrix} 3 \\ 2 \\ -3 \\ 6 \\ 0 \\ 0 \end{pmatrix}, \quad a_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$a_4 = \begin{pmatrix} 4 \\ 3 \\ -12 \\ 0 \\ 12 \\ 0 \end{pmatrix}, \quad a_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad a_6 = \begin{pmatrix} -3 \\ -6 \\ -1 \\ 0 \\ 0 \\ 6 \end{pmatrix}$$

In this case there are four different nonnegative definite symmetric solutions:

$$(i) \quad a_1, a_3, a_5; \quad X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$(ii) \quad a_2, a_3, a_5; \quad X_2 = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$(iii) \quad a_1, a_4, a_5; \quad X_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 4 & 0 \\ 0 & 3 & 0 \\ 0 & -12 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$(iv) \quad a_2, a_4, a_5; \quad X_4 = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 & 4 & 0 \\ 2 & 3 & 0 \\ -3 & -12 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 18 & -24 & 0 \\ -24 & 35 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In Section 3 the different solutions will be discussed from an optimal control point of view. It is shown that they all in some sense can be considered as solutions of the optimal control problem.

In the general case, assume that  $\lambda_1, \dots, \lambda_m$  are  $m$  distinct nonobservable modes of  $[Q_1, A]$  such that  $\text{Re}\{\lambda_i\} > 0, i = 1 \dots m$ . Using the result of Theorem 11 in a combinatorial way, it is possible to prove that there are at least

$$\sum_{j=0}^m \binom{m}{j} = 2^m$$

nonnegative definite solutions, provided that  $[b_1 \dots b_n]^{-1}$  exist. It is also possible to get some kind of order between the different solutions in the sense that there is always one largest and one smallest solution.

### Theorem 12

Let  $\lambda_1, \dots, \lambda_m$  be distinct nonobservable modes of  $[Q_1, A]$  such that  $\text{Re}\{\lambda_i\} > 0, i = 1 \dots m$ . Assume that  $X_1$  is the nonnegative definite solution obtained by the eigenvectors corresponding to the eigenvalues of  $E$  with  $\text{Re}\{\lambda\} < 0$ . If  $X_2$  is another nonnegative definite solution of (2.1), then  $X_1 \geq X_2$ .

Proof. Both  $X_1$  and  $X_2$  satisfy (2.1). Then

$$A^T X_1 + X_1 A - X_1 B Q_2^{-1} B^T X_1 + Q_1 = 0,$$

$$A^T X_2 + X_2 A - X_2 B Q_2^{-1} B^T X_2 + Q_1 = 0.$$

Subtracting the second equation from the first and reordering the terms yields

$$\begin{aligned} (A - B Q_2^{-1} B^T X_1)^T (X_1 - X_2) + (X_1 - X_2) (A - B Q_2^{-1} B^T X_1) \\ = -(X_1 - X_2) B Q_2^{-1} B^T (X_1 - X_2). \end{aligned}$$

Since  $\tilde{A} = (A - B Q_2^{-1} B^T X_1)$  is asymptotic stable, it follows from the Lyapunov stability theory that the symmetric solution  $Y$  of

$$\tilde{A}^T Y + Y \tilde{A} = -Y B Q_2^{-1} B^T Y$$

is nonnegative definite. Then  $Y = X_1 - X_2 \geq 0$ , which finally proves that  $X_1 \geq X_2$ . ■

In the previous example  $X_4$  is the largest solution. Using a similar technique it can be shown that among all nonnegative definite solutions there is a smallest solution. This solution is obtained if the eigenvectors corresponding to  $\lambda_1, \dots, \lambda_m$  all are included. In the example above  $X_1$  is the smallest solution.

Notice that Theorems 11 and 12 can not be extended to cases where  $\lambda_i = 0$  is a distinct nonobservable mode of  $[Q_1, A]$ . The corresponding eigenvector of rank one must then be included in the solution, and the only possibility to vary the solution with respect to  $\lambda_i = 0$  is to include also the eigenvector of rank two. But it is easily verified that the structure

$$\begin{bmatrix} b_i \\ 0 \\ n \end{bmatrix}$$

is not preserved for eigenvectors of  $E$  of rank greater than one. To prove this, assume that there is one eigenvector

$$\begin{bmatrix} x_1 \\ 0_n \end{bmatrix}$$

of  $E$  of rank one, and one eigenvector

$$\begin{bmatrix} x_2 \\ 0_n \end{bmatrix}$$

of  $E$  of rank two, both corresponding to  $\lambda_i = 0$ . Then

$$\begin{bmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{bmatrix} \begin{bmatrix} x_2 \\ 0_n \end{bmatrix} = \begin{bmatrix} x_1 \\ 0_n \end{bmatrix}$$

and thus

$$Ax_2 = x_1$$

Since  $x_1 \neq 0$  and  $x_2 \neq 0$ , this implies that  $A$  has an eigenvalue  $\lambda_i = 0$  of multiplicity two, which contradicts the previous assumption that  $\lambda_i = 0$  is a distinct eigenvalue of  $A$ . The eigenvector of rank two then must have the structure

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

where  $y_2 \neq 0$ . Then nothing can be stated about symmetry and existence of nonnegative definite solutions.

### 3. THE RICCATI EQUATION IN OPTIMAL CONTROL PROBLEMS

#### 3.1. The Optimal Control Problem

Consider the linear time-invariant system

$$\frac{dx}{dt} = Ax + Bu, \quad x(t_0) = x_0, \quad (3.1)$$

with the criteria

$$J = x^T(t_f)Q_0x(t_f) + \int_{t_0}^{t_f} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\} ds, \quad (3.2)$$

where  $Q_0$  and  $Q_1$  are nonnegative definite symmetric matrices and  $Q_2$  is a positive definite symmetric matrix. The minimum value of (3.2) is known to be [5]

$$J^0(x; t_0) = x^T(t_0)S(t_0)x(t_0),$$

where  $S(t)$  is a nonnegative definite symmetric solution of the matrix Riccati equation

$$-\frac{dS}{dt} = A^T S + SA - SBQ_2^{-1}B^T S + Q_1, \quad S(t_f) = Q_0, \quad (3.3)$$

The optimal control  $u(t)$ ,  $t_0 \leq t \leq t_f$ , is a linear time-varying feedback from the state of the system

$$u(t) = -L(t)x(t),$$

where

$$L(t) = Q_2^{-1}B^T S(t).$$

In particular we are interested in the optimal regulator problem, that is, we look for a time-invariant linear feedback

$$u(t) = -Lx(t)$$

such that



$$J = \int_0^{\infty} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\}ds \quad (3.4)$$

is minimized. This problem is generally solved by a straightforward integration of (3.3) until a stationary solution is reached.

Existence and uniqueness of solutions of (3.3) is proved in [5] and [6]. It is also shown that, with the assumptions made about  $Q_1$  and  $Q_2$ , the solution  $S(t)$  is nonnegative definite and symmetric. If the pair  $[A, B]$  is completely controllable and the pair  $[Q_1, A]$  is completely observable, it is shown in [5], [6], and [7] that  $S(t)$  tends to a unique positive definite solution  $S$  of the algebraic equation

$$A^T S + SA - SBQ_2^{-1}B^T S + Q_1 = 0. \quad (3.5)$$

Then  $L = Q_2^{-1}B^T S$  will be the solution of the optimal regulator problem, and the optimal closed loop system  $A - BQ_2^{-1}B^T S$  is asymptotic stable (c.f. Theorem 8). It is also shown that the boundary condition  $Q_0$  is arbitrary.

In [8] Wonham generalizes to  $[A, B]$  being stabilizable and  $[Q_1, A]$  being detectable. Then  $S(t)$  converges toward a unique nonnegative definite solution  $S$  of (3.5). The closed loop system  $A - BQ_2^{-1}B^T S$  is then asymptotic stable, and the boundary condition  $Q_0$  is still arbitrary.

In this section we will consider the optimal control problem under the assumption that  $Q_1$  is an arbitrary symmetric nonnegative definite matrix. Detectability of  $[Q_1, A]$  is thus no longer assumed. Existence, uniqueness and symmetry of the solution  $S(t)$  of (3.3) then still hold [5], but according to section 2 there may be more than one nonnegative definite solution of the stationary Riccati equation (3.5). Then the boundary condition  $Q_0$  can not be chosen arbitrarily, but will determine to what stationary solution  $S(t)$  converges. As for the numerical solution of the optimization problem, this implies that a straightforward integration of the Riccati equation may be an unstable procedure.

The asymptotic dependence on  $Q_0$  will be shown to have a nice physical interpretation, and this will finally lead to a generalization of optimal control theory for linear systems with quadratic loss functions.

To simplify the analysis we will throughout this section assume that the nonobservable modes of  $A$  are real, distinct and not identical to zero. Since it was shown in Section 2 that nonobservable modes  $\lambda_i = 0$  will not cause several nonnegative definite solutions of (2.1), this restriction is of no importance for the purpose of this section.

### 3.2. Upper and Lower à priori Bounds of S(t)

Suppose that the control variable  $u(t)$  is given through an arbitrary linear feedback from the state of the system.

$$u(t) = -\tilde{L}x(t).$$

Since  $[A, B]$  is assumed stabilizable, it is always possible to choose  $\tilde{L}$  so that the closed loop system matrix  $A - B\tilde{L}$  is asymptotic stable [9]. Introduce the fundamental matrix  $\tilde{\Psi}(t; s)$  associated with  $A - B\tilde{L}$ :

$$\frac{\partial \tilde{\Psi}(t; s)}{\partial t} = (A - B\tilde{L})\tilde{\Psi}(t; s),$$

$$\tilde{\Psi}(t; t) = I.$$

The corresponding cost is

$$\tilde{J} = x^T(t_f)\tilde{\Psi}^T(t_f; t)Q_0\tilde{\Psi}(t_f; t)x(t) + \int_t^{t_f} x^T(t)\tilde{\Psi}^T(s; t)\{Q_1 + \tilde{L}^T Q_2 \tilde{L}\}\tilde{\Psi}(s; t)x(t)ds$$

or

$$\tilde{J} = x^T(t)\tilde{S}(t)x(t),$$

where

$$\tilde{S}(t) = \tilde{\Psi}^T(t_f; t)Q_0\tilde{\Psi}(t_f; t) + \int_t^{t_f} \tilde{\Psi}^T(s; t)\{Q_1 + \tilde{L}^T Q_2 \tilde{L}\}\tilde{\Psi}(s; t)ds. \quad (3.6)$$

( $A - B\tilde{L}$ ) being asymptotic stable,  $\tilde{S}(t)$  tends toward a nonnegative definite matrix  $\tilde{S}$  as  $t \rightarrow -\infty$ .  $\tilde{S}$  is the unique solution of the algebraic equation

$$(A - B\tilde{L})^T \tilde{S} + \tilde{S}(A - B\tilde{L}) + Q_1 + \tilde{L}^T Q_2 \tilde{L} = 0. \quad (3.7)$$

Obviously  $J^0 \leq \tilde{J}$ , and then  $S(t) \leq \tilde{S}(t)$ ,  $t \leq t_f$ . Then any linear feedback  $\tilde{L}$  such that  $A - B\tilde{L}$  is asymptotic stable yields an upper bound for  $S(t)$ ,  $t \leq t_f$ . This is a very rough bound, and we will show that there exists a smaller à priori bound.

Let  $S_1$  be the solution of the stationary Riccati equation corresponding to  $\text{Re}\{\lambda_i\} < 0$ ,  $i = 1 \dots n$ . (Notice that we will throughout Section 3 assume that the different nonnegative definite stationary solutions exist.) Then

$$A^T S_1 + S_1 A - S_1 B Q_2^{-1} B^T S_1 + Q_1 = 0,$$

and the closed loop system matrix  $(A - B Q_2^{-1} B^T S_1)$  is asymptotic stable. Further, assume that  $S_2(t)$  is the solution of

$$-\frac{dS_2}{dt} = A^T S_2 + S_2 A - S_2 B Q_2^{-1} B^T S_2 + Q_1,$$

with boundary condition

$$S_2(t_f) = \alpha I.$$

$I$  is the identity matrix and  $\alpha$  is a positive scalar.

Then  $(S_2 - S_1)$  satisfies the differential equation

$$-\frac{d}{dt} (S_2 - S_1) = (A - B Q_2^{-1} B^T S_1)^T (S_2 - S_1) + (S_2 - S_1) \cdot (A - B Q_2^{-1} B^T S_1) - (S_2 - S_1) B Q_2^{-1} B^T (S_2 - S_1) \quad (3.8)$$

with boundary condition

$$(S_2 - S_1)(t_f) = \alpha I - S_1.$$

Now choose

$$\alpha > \|S_1\| \quad (\alpha > \max_i \lambda_i, \text{ where } \lambda_i \text{ are eigenvalues of } S_1).$$

Then  $\alpha I - S_1$  is positive definite, and the solution of (3.8) exists and is unique. It is also nonnegative definite for  $t \leq t_f$ . Let  $\Psi(t;s)$  be the fundamental matrix associated with  $(A - B Q_2^{-1} B^T S_1)$ . Then

$$\frac{\partial}{\partial t} \Psi(t;s) = (A - B Q_2^{-1} B^T S_1) \Psi(t;s),$$

$$\Psi(t;t) = I,$$

$$\frac{\partial}{\partial s} \Psi(t;s) = -\Psi(t;s) (A - B Q_2^{-1} B^T S_1),$$

and (3.8) is equivalent to the integral equation

$$(S_2 - S_1)(t) = \Psi^T(t_f;t) \left\{ (\alpha I - S_1)^{-1} + \int_t^{t_f} \Psi^T(t_f;s) B Q_2^{-1} B^T \Psi(t_f;s) ds \right\}^{-1} \Psi(t_f;t). \quad (3.9)$$

$(\alpha I - S_1)^{-1}$  exists since  $\alpha > \|S_1\|$ , and then

$$\left\{ (\alpha I - S_1)^{-1} + \int_t^{t_f} \Psi^T(t_f; s) B Q_2^{-1} B^T \Psi(t_f; s) ds \right\}^{-1}$$

exists and is positive definite. If  $P_1$  and  $P_2$  are two arbitrary positive definite matrices, the inequality  $P_1 \leq P_2$  implies that  $P_1^{-1} \geq P_2^{-1}$  also holds [10]. Then

$$\left\{ (\alpha I - S_1)^{-1} + \int_t^{t_f} \Psi^T(t_f; s) B Q_2^{-1} B^T \Psi(t_f; s) ds \right\}^{-1} \leq (\alpha I - S_1)$$

and

$$(S_2 - S_1)(t) \leq \Psi^T(t_f; t) (\alpha I - S_1) \Psi(t_f; t).$$

The fundamental matrix  $\Psi(t_f; t) \rightarrow 0$  as  $t \rightarrow -\infty$  since  $(A - B Q_2^{-1} B^T S_1)$  is asymptotic stable, and then  $(S_2 - S_1)(t) \rightarrow 0$  as  $t \rightarrow -\infty$ .

The solution of (3.3) with boundary condition

$$S(t_f) = \alpha I; \quad \alpha > \|S_1\|$$

then converges to the largest solution  $S_1$  of (3.5). Now let  $Q_0$  be an arbitrary nonnegative definite symmetric matrix, and assume that  $Q_0$  is the boundary condition of

$$-\frac{dS_1}{dt} = A^T S_1 + S_1 A - S_1 B Q_2^{-1} B^T S_1 + Q_1,$$

$$S_1(t_f) = Q_0.$$

Further let  $S_2(t)$  be the solution of

$$-\frac{dS_2}{dt} = A^T S_2 + S_2 A - S_2 B Q_2^{-1} B^T S_2 + Q_1,$$

$$S_2(t_f) = \beta I.$$

As before the difference  $(S_2 - S_1)(t)$  satisfies

$$\begin{aligned}
-\frac{d}{dt}(S_2 - S_1) &= (A - BQ_2^{-1}B^T S_1)^T (S_2 - S_1) + (S_2 - S_1)(A - BQ_2^{-1}B^T S_1) \\
&\quad - (S_2 - S_1)BQ_2^{-1}B^T (S_2 - S_1), \quad (3.10) \\
(S_2 - S_1)(t_f) &= \beta I - Q_0.
\end{aligned}$$

With  $\Psi(t;s)$  being the fundamental matrix associated with  $(A - BQ_2^{-1}B^T S_1(t))$ , (3.10) is equivalent to

$$\begin{aligned}
(S_2 - S_1)(t) &= \Psi^T(t_f;t)(\beta I - Q_0)\Psi(t_f;t) \\
&\quad + \int_t^{t_f} \Psi^T(s;t)(S_2(s) - S_1(s))BQ_2^{-1}B^T(S_2(s) - S_1(s))\Psi(s;t) ds \quad (3.11)
\end{aligned}$$

$(S_2 - S_1)(t)$  is then nonnegative definite, and

$$S_2(t) \geq S_1(t)$$

for

$$\beta \geq \|Q_0\|.$$

For a solution  $S(t)$  of (3.3) with an arbitrary nonnegative definite boundary condition  $S(t_f) = Q_0$ , it is then always possible to find an upper a priori bound  $\bar{S}(t)$ , such that  $S(t) \leq \bar{S}(t)$ ,  $t \leq t_f$ , and  $\bar{S}(t) \rightarrow S_m$  as  $t \rightarrow -\infty$ .  $\bar{S}(t)$  can be chosen as the solution of (3.3) with boundary condition  $\bar{S}(t_f) = \gamma I$  where  $\gamma > \max\{\|S_m\|, \|Q_0\|\}$ , and  $S_m$  is the largest solution of the algebraic equation (3.5).

In a similar way it is easy to give a priori lower bounds for the solutions of (3.3). Let  $S_1(t)$  and  $S_2(t)$  be solutions corresponding to the boundary conditions  $S_1(t_f) = 0$  and  $S_2(t_f) = Q_0$ ,  $Q_0 \geq 0$ . From (3.11) it then follows that  $S_2(t) \geq S_1(t)$ ,  $t \leq t_f$ . The smallest solution  $S(t)$  ( $= S_1(t)$ ) of (3.3), will then correspond to the boundary condition  $S(t_f) = 0$ .  $S(t)$  is the solution of the integral equation

$$S(t) = \int_t^{t_f} \Psi^T(s;t)\{S(s)BQ_2^{-1}B^T S(s) + Q_1\}\Psi(s;t)ds, \quad (3.12)$$

where

$$\frac{\partial}{\partial t} \Psi(t;s) = (A - BQ_2^{-1}B^T S(t)) \Psi(t;s).$$

From (3.12) follows that  $S(t)$  is monotonic nondecreasing as  $t \rightarrow -\infty$ , and since the solutions are bounded,  $S(t)$  converges toward a solution of the stationary Riccati equation (3.5). This obviously is the smallest solution  $S'$ , because assume that  $S(t)$  converges toward the solution  $S''$  of (3.5), and  $S'' \geq S'$ . This contradicts the fact that  $S(t) \leq S'$ ,  $t \leq t_f$ , unless  $S' = S''$ . Thus the solution  $S(t)$  of (3.3) with boundary condition  $S(t_f) = 0$  converges to the smallest solution of the algebraic equation (3.5).

When the pair  $[Q_1, A]$  is completely observable [7] or detectable [8], there is a unique positive definite or nonnegative definite solution of (3.5). The upper and lower a priori bounds for  $S(t)$  are then identical, and then convergence of  $S(t)$  follows. In the case of nonobservable unstable modes of  $[Q_1, A]$ , however, these bounds do not coincide, and it then remains to prove convergence of  $S(t)$  toward a stationary solution of (3.5) as  $t \rightarrow -\infty$ , for arbitrary nonnegative definite boundary conditions  $Q_0$ .

### 3.3. Convergence Properties

The convergence of  $S(t)$  toward a stationary solution is proved in [6] for the pair  $[Q_1, A]$  completely observable, and in [8] for the case of  $[Q_1, A]$  completely detectable. In this section we will prove convergence for the particular case  $Q_1 = 0$ ,  $Q_0 > 0$ , and all modes of  $A$  unstable. It may then be possible to combine this result with those of [6] and [8] to prove convergence for arbitrary nonnegative definite matrices  $Q_1$  and  $Q_0$ .

As before, the differential equation

$$-\frac{dS}{dt} = A^T S + SA - SBQ_2^{-1}B^T S + Q_1, \quad S(t_f) = Q_0,$$

is transformed into an integral equation

$$S(t) = \Psi^T(t_f; t) \left\{ Q_0^{-1} + \int_t^{t_f} \Psi^T(t_f; s) B Q_2^{-1} B^T \Psi(t_f; s) ds \right\}^{-1} \Psi(t_f; t),$$

where

$$\frac{\partial}{\partial t} \Psi(t;s) = (A - BQ_2^{-1}B^T S(t))\Psi(t;s),$$

$$\Psi(t;t) = I.$$

Since  $Q_0 > 0$ , and  $\Psi(t_f;t)$  has full rank for  $t \leq t_f$ ,  $S(t)$  is positive definite and hence invertible for  $t \leq t_f$ . Then consider  $S^{-1}(t)$ ,

$$-\frac{dS^{-1}}{dt} = -S^{-1}A^T - AS^{-1} + BQ_2^{-1}B^T \quad (3.13)$$

$$S^{-1}(t_f) = Q_0^{-1}.$$

Let  $\phi(t;s)$  be the fundamental matrix associated with  $-A$ .

$$\frac{\partial}{\partial t} \phi(t;s) = -A\phi(t;s),$$

$$\phi(t;t) = I.$$

It is then possible to give an explicit expression for the solution of (3.13):

$$S^{-1}(t) = \phi^T(t_f;t) \left\{ Q_0^{-1} + \int_t^{t_f} \phi^T(s;t_f) BQ_2^{-1}B^T \phi(s;t_f) ds \right\} \phi(t_f;t), \quad (3.14)$$

which reduces to

$$S^{-1}(t) = \phi^T(t_f;t) Q_0^{-1} \phi(t_f;t) + \int_t^{t_f} \phi^T(s,t) BQ_2^{-1}B^T \phi(s,t) ds.$$

$\{-A\}$  being asymptotic stable implies that  $\phi(t_f,t) \rightarrow 0$  as  $t \rightarrow -\infty$ , and

$$S^{-1}(t) \rightarrow \int_t^{t_f} \phi^T(s,t) BQ_2^{-1}B^T \phi(s,t) ds. \quad (3.15)$$

The pair  $[Q_1, A]$  having just nonobservable unstable modes implies that stabilizability is equivalent to complete controllability, and thus (3.15) is positive definite for  $t < t_f$ .  $S^{-1}(t)$  then converges toward the unique positive definite solution of

$$AS^{-1} + S^{-1}A^T - BQ_2^{-1}B^T = 0$$

as  $t \rightarrow -\infty$ , and thus  $S(t)$  converges toward a positive definite solution of

$$A^T S + SA - SBQ_2^{-1}B^T S = 0$$

as  $t \rightarrow -\infty$ . This completes the proof of convergence for the special case  $Q_1 = 0$  and  $Q_0 > 0$ .

Now assume that convergence holds for arbitrary  $Q_0$  and  $Q_1$ , symmetric and nonnegative definite. It is then of interest to examine to what stationary solution  $S(t)$  converges as  $t \rightarrow -\infty$ . Consider the equivalent integral equation

$$S(t) = \Psi^T(t_f; t) Q_0 \Psi(t_f; t) + \int_t^{t_f} \Psi^T(t_f; s) [Q_1 + S(s) B Q_2^{-1} B^T S(s)] \Psi(t_f; s) ds$$

where  $\Psi(t; s)$  is the fundamental matrix associated with the closed loop system matrix  $[A - B Q_2^{-1} B^T S(t)]$ . When  $t, s \rightarrow -\infty$ ,

$$\Psi^T(t_f; t) Q_0 \Psi(t_f; t) \rightarrow 0$$

and

$$\Psi^T(t_f; s) Q_1 \Psi(t_f; s) \rightarrow 0.$$

since  $S(t)$  is bounded, and since it is assumed that the eigenvalues  $\lambda_i$  of the stationary closed loop system satisfy  $\text{Re}\{\lambda_i\} \neq 0$ .

Now let  $\lambda_i > 0$  be a nonobservable mode of  $[Q_1, A]$ , and assume that two stationary solutions  $S_1$  and  $S_2$  of (2.1) exist, such that  $\lambda_i > 0$  is an eigenvalue of the closed loop system  $A - B Q_2^{-1} B^T S_1$  and  $-\lambda_i < 0$  an eigenvalue of  $A - B Q_2^{-1} B^T S_2$ . From section 2.4 then follows that  $S_2 \geq S_1$ . If  $\lambda_i$  also is a nonobservable mode of  $[Q_0, A]$ ,  $\lambda_i$  will have no influence on the cost functional, and then it is not necessary to stabilize this mode. Then  $S_1$  will be the solution of the optimization problem, and the optimal closed loop system will contain an unstable mode. However, if  $\lambda_i$  is observable through  $Q_0$ ,  $S_1$  can not be the optimal solution, since the unstable mode of the closed loop system will then result in a large cost due to the term  $x^T(t_f) Q_0 x(t_f)$ . When  $t \rightarrow -\infty$ , this contribution to the cost tends to infinity, and then it follows that  $S(t)$  converges to  $S_2$  as  $t \rightarrow -\infty$ . The boundary condition  $Q_0$  will thus have the same influence on the stationary solution as  $Q_1$ .

In the general case, assume that  $A$  has  $r$  unstable eigenvalues  $\lambda_1, \dots, \lambda_r$ , nonobservable in  $[Q_1, A]$ . If  $\lambda_1, \dots, \lambda_k$ ,  $k < r$ , are observable in  $[Q_0, A]$ ,  $S(t)$  must converge toward a stationary solution  $S$  of (3.5), such that the optimal system  $A - B Q_2^{-1} B^T S$  has eigenvalues  $-\lambda_1, \dots, -\lambda_k, \lambda_{k+1}, \dots, \lambda_r$ .



### 3.4. Numerical Instability

The optimal regulator problem is generally solved by straightforward integration of (3.3) until a stationary solution is reached with desired accuracy. In the case of complete detectability of the pair  $[Q_1, A]$ , this is a stable procedure when (3.3) is integrated backward in time. However, the existence of several stationary solutions may cause even the backward integration to be an unstable process. This is illustrated in the following example [7]:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad Q_2 = (1).$$

The unstable mode  $\lambda = 1$  is nonobservable in  $Q_1$ , and there are two nonnegative definite solutions of (3.5).

$$S_1 = \begin{pmatrix} 3 + \sqrt{2} & 1 + \sqrt{2} \\ 1 + \sqrt{2} & 1 + \sqrt{2} \end{pmatrix}, \quad S_2 = \begin{pmatrix} \sqrt{2} - 1 & -\sqrt{2} + 1 \\ -\sqrt{2} + 1 & \sqrt{2} - 1 \end{pmatrix}.$$

$S_1$  (positive definite) yields the closed loop mode  $\lambda = -1$ , while  $S_2$ , which is the solution of the optimal regulator problem, leaves  $\lambda = 1$  unchanged. It is easily verified that  $S_1 \geq S_2$ .

If the boundary condition  $S(t_f) = 0$  is chosen,  $S(t)$  converges toward  $S_2$  according to Section 3.2. From (3.3) then follows that  $S(t)$  will have the structure

$$S(t) = \begin{pmatrix} \alpha(t) & -\alpha(t) \\ -\alpha(t) & \alpha(t) \end{pmatrix},$$

where  $\alpha(t) > 0$ ,  $t < t_f$ . Depending on how  $\frac{dS}{dt}$  is computed, numerical inaccuracies may occur in different ways. Suppose that at time  $t_1$ ,  $t_1 < t_f$ , the computed solution is

$$S(t_1) = \begin{pmatrix} \alpha(t_1) + \epsilon & -\alpha(t_1) \\ -\alpha(t_1) & \alpha(t_1) \end{pmatrix},$$

where  $\epsilon > 0$  is a small quantity.  $S(t_1)$  is then positive definite, and can be considered as boundary condition for further computation of  $S(t)$ ,  $t < t_1 < t_f$ .

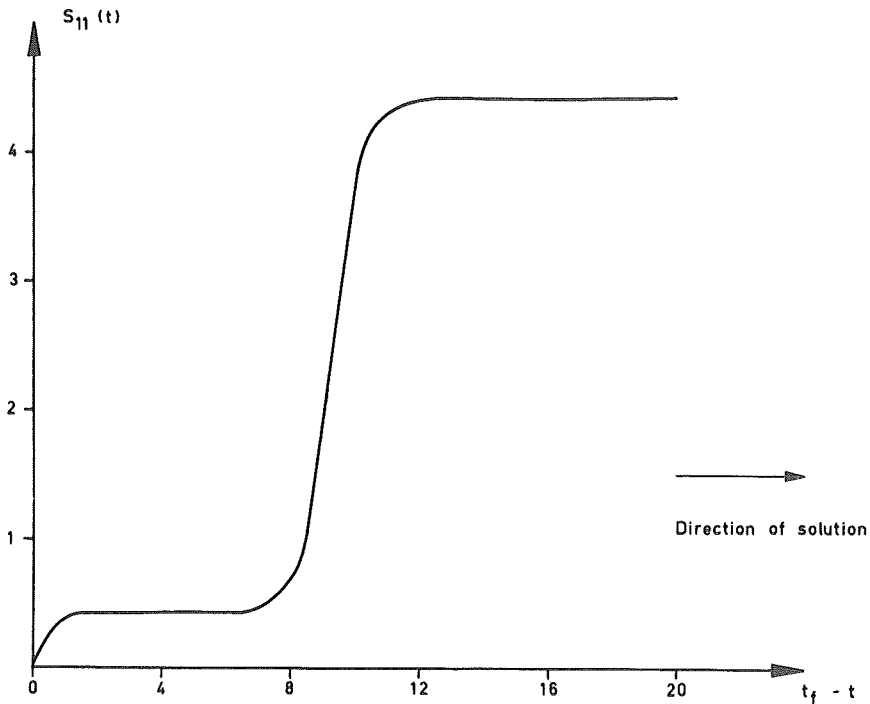


Fig. 1. -  $S_{11}(t)$  computed with a fourth-order Runge- Kutta method,

$$Q_0 = \begin{pmatrix} 10^{-7} & 0 \\ 0 & 0 \end{pmatrix}.$$

But  $[S(t_1), A]$  is completely observable and the solution will converge toward the largest solution  $S_1$ . This is illustrated in Figure 1, where the computed 1-1 element of  $S(t)$  is shown. A disturbance  $\epsilon = 10^{-7}$  was introduced in the 1-1 element of  $Q_0$ , and a fourth-order Runge-Kutta method was used.

The same situation arises if the errors are equal in all elements of  $S(t_1)$ :

$$S(t_1) = \begin{pmatrix} \alpha(t_1) + \epsilon & -\alpha(t_1) + \epsilon \\ -\alpha(t_1) + \epsilon & \alpha(t_1) + \epsilon \end{pmatrix}.$$

For  $\alpha(t_1) > 0$  and  $\epsilon > 0$ ,  $S(t_1)$  is positive definite, and  $S(t)$  will converge toward  $S_1$  as  $t \rightarrow -\infty$ .

Another way to solve (3.3) is the fundamental matrix approach [4, 11]. With the algorithm proposed in [11], the numerical errors entered in the following way:

$$S(t_1) = \begin{pmatrix} \alpha(t_1) & -\alpha(t_1) - \epsilon \\ -\alpha(t_1) - \epsilon & \alpha(t_1) \end{pmatrix}.$$

For  $\epsilon > 0$ ,  $S(t_1)$  is indefinite, and can no longer be considered as a new boundary condition for further computation. However, computational experiments show that  $S(t)$  still converges toward  $S_1$ , and the fundamental matrix method can then be considered as a stable method. The 1-1 element of the computed solution  $S(t)$  is shown in Figure 2 for different values of  $Q_2$ . Notice that the differences for small values of  $t_f - t$  is slightly exaggerated.

With the same errors introduced, the Runge-Kutta method was applied. Due to large values of  $\frac{dS}{dt}$  exponent overflow occurred and the stationary solution  $S_1$  was never reached.

### 3.5. Generalization of Optimal Control Theory for Linear Systems with Quadratic Loss

The preceding section indicates a possible generalization of optimal control theory for linear systems with a quadratic loss function. We then drop the requirement that  $[Q_1, A]$  is detectable (except for modes  $\lambda_i = 0$ ), but it is still assumed that  $[A, B]$  is stabilizable. Since asymptotic stability of the optimal system is a desired property, we will look for the minimizing control in the class of asymptotic stable linear feedback controls.

#### Theorem 13

Consider the stabilizable system

$$\frac{dx}{dt} = Ax + Bu$$

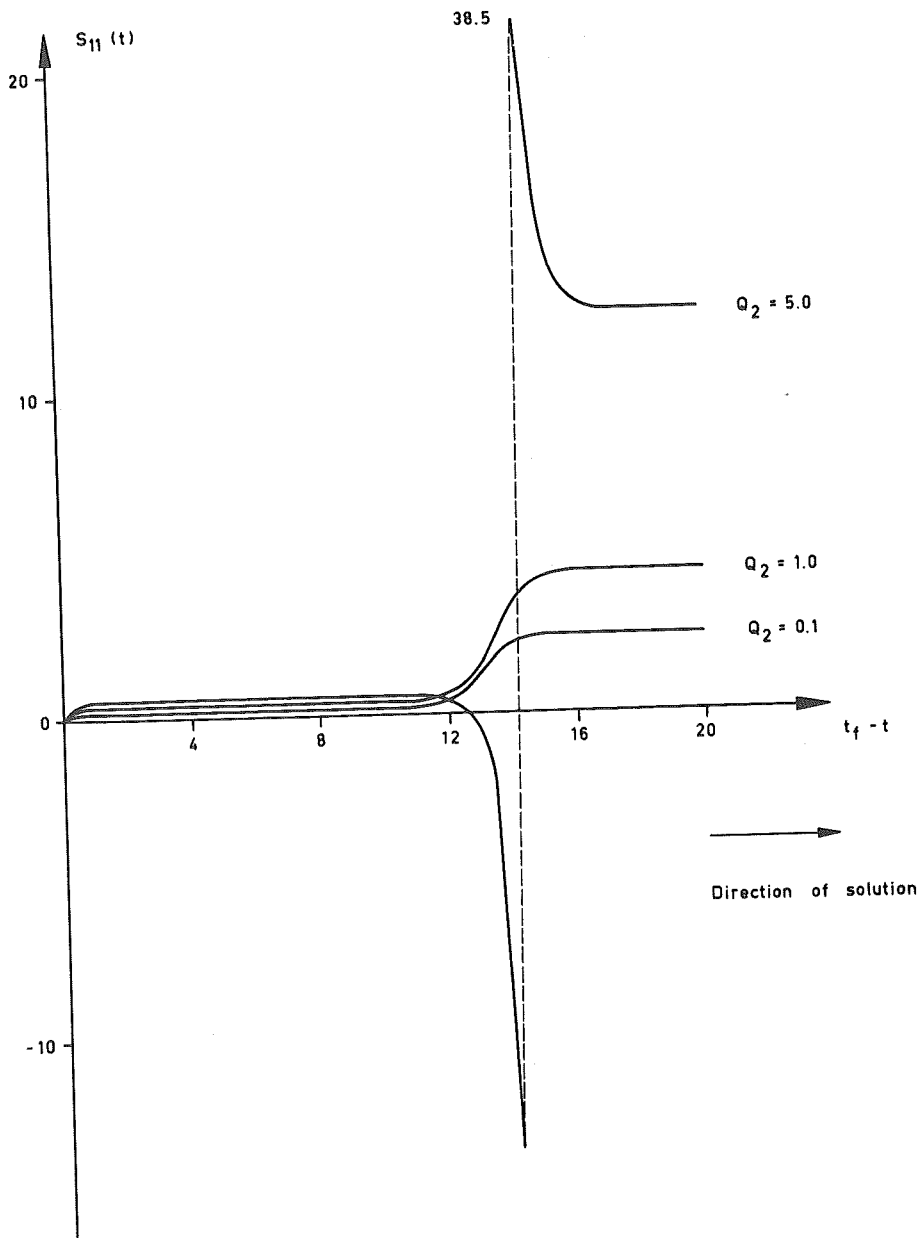


Fig. 2. -  $S_{11}(t)$  computed by the fundamental matrix method for various  $Q_2$ .  $S_{11}(t)$  is plotted versus the time difference  $t_f - t$ .

with the loss function

$$J = \int_0^{\infty} \{x^T(s)Q_1x(s) + u^T(s)Q_2u(s)\} ds,$$

where  $Q_1$  is nonnegative definite symmetric, and  $Q_2$  positive definite symmetric. Assume that there is no nonobservable mode  $\lambda_i = 0$  of  $[Q_1, A]$ . In the class of asymptotic stable linear feedback controls, the minimizing control is given by

$$u = -Q_2^{-1}B^T Sx$$

where

$$S = \lim_{t \rightarrow -\infty} S^*(t).$$

$S^*(t)$  is the solution of

$$-\frac{dS^*}{dt} = A^T S^* + S^* A - S^* B Q_2^{-1} B^T S^* + Q_1$$

with boundary condition

$$S^*(t_f) = \gamma I \quad (\gamma > \|S_m\|, t_f \text{ is arbitrary})^1)$$

Proof: From Section 3.2 follows that  $S^*(t)$  converges toward the largest stationary solution  $S_m$  of (3.5). But, if there are several nonnegative definite solutions of (3.5),  $S_m$  is not the solution of the optimal control problem, and it remains to prove that in the class of stable linear feedbacks  $u = -Lx$ ,  $L_m = Q_2^{-1}B^T S_m$  yields the minimum value of the loss function  $V$ .

Consider an arbitrary stable linear feedback  $u = -L_1x$ . The corresponding value of  $J$  is

$$J = x^T(0)S_1x(0).$$

where

$$S_1 = \int_0^{\infty} e^{(A-BL_1)^T s} \{Q_1 + L_1^T Q_2 L_1\} e^{(A-BL_1)s} ds \quad (3.16)$$

1) From Section 3.3 follows that probably any  $\gamma > 0$  will do since  $Q_0$  has the same influence on the stationary solution as  $Q_1$ .  $S_m$  denotes the largest stationary solution of (3.5).

is nonnegative definite symmetric. Since  $(A - BL_1)$  is asymptotic stable,  $S_1$  satisfies the algebraic equation

$$(A - BL_1)^T S_1 + S_1 (A - BL_1) + Q_1 + L_1^T Q_2 L_1 = 0. \quad (3.17)$$

The corresponding equation for  $L_m = Q_2^{-1} B^T S_m$  is

$$(A - BL_m)^T S_m + S_m (A - BL_m) + Q_1 + L_m^T Q_2 L_m = 0. \quad (3.18)$$

(3.18) is equivalent to

$$(A - BL_1)^T S_m + S_m (A - BL_1) + Q_1 + L_1^T B^T S_m + S_m BL_1 - L_m^T B^T S_m - S_m BL_m + L_m^T Q_2 L_m = 0. \quad (3.19)$$

Subtract (3.19) from (3.17):

$$(A - BL_1)^T (S_1 - S_m) + (S_1 - S_m)(A - BL_1) + L_1^T Q_2 L_1 - L_1^T B^T S_m - S_m BL_1 + L_m^T B^T S_m + S_m BL_m - L_m^T Q_2 L_m = 0$$

Since  $Q_2 L_m = B^T S_m$ , this equation reduces to

$$(A - BL_1)^T (S_1 - S_m) + (S_1 - S_m)(A - BL_1) + L_1^T Q_2 L_1 - L_1^T Q_2 L_m - L_m^T Q_2 L_1 + L_m^T Q_2 L_m = 0 \quad (3.20)$$

or

$$(A - BL_1)^T (S_1 - S_m) + (S_1 - S_m)(A - BL_1) + (L_1 - L_m)^T Q_2 (L_1 - L_m) = 0.$$

Since  $(A - BL_1)$  is asymptotic stable, the solution  $(S_1 - S_m)$  is nonnegative definite, and is zero if and only if  $L_1 = L_m$ . This completes the proof. ■

It is now possible to give a physical interpretation of the different stationary nonnegative definite solutions of (3.5). Suppose that  $[Q_1, A]$  has some unstable nonobservable modes. Then the smallest solution, which

is the solution of the optimal control problem, leaves these modes unchanged, and the closed loop system is unstable. If it is desired to stabilize just one mode, the best linear feedback is given by the stationary solution which corresponds to that mode stabilized. Naturally this requires more energy, and thus the term

$$\int_0^{\infty} u^T(s) Q_2 u(s) ds$$

becomes larger. The most expensive case is of course when all modes are stabilized, which corresponds to the largest solution of (3.5).

The stationary Riccati equation then has the nice property that it contains the optimal solutions for all degrees of stability.

### 3.6. Minimum Energy Regulator

As an interesting special case, consider the problem of finding an asymptotic stable linear feedback  $u = -Lx$  for the system

$$\frac{dx}{dt} = Ax + Bu$$

with the criteria

$$J = \int_0^{\infty} u^T(s) Q_2 u(s) ds.$$

$Q_2$  is positive definite symmetric, and  $J$  can then be interpreted as the total energy required. If  $A$  already is asymptotic stable, the problem has the trivial solution  $u(t) \equiv 0$ .

Then assume that  $A$  has eigenvalues  $\lambda_1, \dots, \lambda_k$  such that  $\text{Re}\{\lambda\} > 0$ , and  $\lambda_{k+1}, \dots, \lambda_n$  with  $\text{Re}\{\lambda\} < 0$ . Since  $Q_1 = 0$ ,  $\lambda_1, \dots, \lambda_k$  are nondetectable, and then  $E$  has the eigenvalues  $\pm\lambda_1, \dots, \pm\lambda_k$ ,

$\pm\lambda_{k+1}, \dots, \pm\lambda_n$ , independent of  $Q_2$ . The optimal stable system thus has the eigenvalues  $-\lambda_1, \dots, -\lambda_k, \lambda_{k+1}, \dots, \lambda_n$ . This can be formulated as some kind of minimum energy principle.

### Theorem 14

Consider the system

$$\frac{dx}{dt} = Ax + Bu$$

where  $A$  has eigenvalues  $\lambda_1, \dots, \lambda_k$  such that  $\text{Re}\{\lambda\} > 0$ , and

$\lambda_{k+1}, \dots, \lambda_n$  with  $\text{Re}\{\lambda\} < 0$ . The minimum energy regulator

$u = -Lx$  then has the property that the eigenvalues of the closed loop system are  $-\lambda_1, \dots, -\lambda_k, \lambda_{k+1}, \dots, \lambda_n$ .

Notice that in case of a multidimensional control vector  $u(t)$ , the feedback  $L$  will in general depend on the particular choice of  $Q_2$ . However, in the single control variable case,  $L$  is independent of  $Q_2$ .

### 4. REFERENCES

- [1] J. E. Potter, "Matrix Quadratic Solutions", *SIAM J. Appl. Math.*, vol. 14, 1966, 496-501.
- [2] B. Friedman, "Principles and Techniques of Applied Mathematics", John Wiley, New York, 1965.
- [3] K. J. Åström, "Lecture Notes on Optimal Control Theory", Lund Institute of Technology, Division of Automatic Control, 1965.
- [4] R. E. Kalman and T. Englar, "A User's Manual for the Automatic Synthesis Program", NASA Report CR-475, 1966, 170-221.
- [5] R. E. Kalman, "Contributions to the Theory of Optimal Control", *Bol. Soc. Mat. Mexicana*, vol. 5, 1960, 102-119.
- [6] R. S. Bucy, "Global Theory of the Riccati Equation", *J. Comput. System Sci.*, vol. 1, 1967, 349-361.
- [7] R. E. Kalman, "When is a Linear Control System Optimal?", *J. Basic Eng.*, vol. 86, 1967, 51-60.
- [8] W. M. Wonham, "On Matrix Quadratic Equations and Matrix Riccati Equations", Technical Report 65-5, Division of Applied Mathematics, Brown University, 1967.

[9]

[10]

[11]



- [9] W. M. Wonham, "On Pole Assignment in Multi-Input Controllable Linear Systems", IEEE Trans. Automatic Control, vol. 12, 1967, 660-665.
- [10] E. F. Beckenbach and R. E. Bellman, "Inequalities", Springer-Verlag, New York, 1965.
- [11] K. Mårtensson, "Linear Quadratic Control Package. Part I, The Continuous Problem", Research Report 6802, Lund Institute of Technology, Division of Automatic Control, 1968.

Part

A N  
OPT

ABS

A no  
pre:  
and  
the  
then  
pro  
alg

Part 2

A NEW APPROACH TO CONSTRAINED FUNCTION  
OPTIMIZATION

ABSTRACT

A new approach to the constrained function optimization problem is presented. It is shown that the ordinary Lagrange multiplier method and the penalty function method may be generalized and combined, and the new concept "multiplier function" is introduced. The problem may then be converted into an unconstrained well-conditioned optimization problem. Methods for numerical solution are discussed, and new algorithms are derived.

## 1. INTRODUCTION

In this paper new methods and algorithms for constrained function optimization are presented. The problem considered is minimization of a function  $f(u)$  subject to the equality constraint  $g(u) = 0$ , where  $g(u)$  is an  $m$ -dimensional vector with components  $g^i(u)$ . (The superscript will be used to denote components of a vector. This will simplify the notations later.)

Many methods for solving this problem have been published, but generally they are based on one of two main ideas.

One is the Lagrange multiplier technique, where the constraints are adjoined to the function by means of multipliers  $\lambda$ , to form a new function

$$L(u, \lambda) = f(u) + \sum_{i=1}^m \lambda^i g^i(u)$$

generally called the Lagrangian of the problem. The problem is then reduced to finding a saddle-point of  $L(u, \lambda)$  in the  $u$ - $\lambda$  space, and thus the dimension of the problem is increased from  $n$  to  $n+m$ .

The other basic approach is the penalty function method. A function including the constraints in a proper manner is adjoined to the original function  $f(u)$ , e. g.

$$f(u) + cg^T(u)g(u)$$

where  $c$  is a positive real-valued parameter. Under very mild conditions, the solution of

$$\min_u \{f(u) + cg^T(u)g(u)\}$$

tends to the solution of

$$\min_u \{f(u)\}$$

subject to  $g(u) = 0$ , as  $c$  tends to infinity. However, the penalty function method is not very attractive from a numerical point of view, since the functions created become very badly conditioned for numerical optimization. Different ways to overcome this difficulty have

been suggested, e. g. by Fiacco and McCormick [1] and by Powell [2]. The basic idea in these papers is to change the penalty function in an iterative way, so as to make the optimum of the penalty function agree with the optimal solution of the problem. However, this requires introduction of a new set of parameters to be iterated on, again increasing the dimension of the minimization problem.

These two basic ideas are combined by Hestenes in [3]. The function

$$F(u, \lambda) = f(u) + \sum_{i=1}^m \lambda_i g_i(u) + c g^T(u)g(u)$$

is introduced, and it is shown that for nonsingular problems,  $F(u, \lambda^*)$ , where  $\lambda^*$  are the optimal multipliers, has a local minimum for  $u=u^*$ , provided that  $c > c_0$ .  $c_0$  is a finite real-valued parameter. This is a considerable improvement over both the original Lagrange multiplier technique, and the penalty function methods. The reasons are that  $c_0$  is finite in contrast to the penalty function methods, and that  $u^*$  constitutes a minimum of  $F(u, \lambda^*)$ , while the extremum of  $L(u, \lambda^*)$  could have any character. However, it still remains to determine the optimal multipliers  $\lambda^*$ .

The method presented in this paper is a generalization of Hestenes' method. We will introduce the concept "multiplier function", and the  $m$ -dimensional vector function  $\mu(u)$  is called an admissible multiplier function if it satisfies some simple conditions. The basic condition is that  $\mu(u^*) = \lambda^*$ . We also define a "generalized Lagrangian" as

$$H(u, c) = f(u) + \mu^T(u)g(u) + c g^T(u)g(u)$$

Using wellknown results, which are briefly stated as lemmas in Section 2, properties of  $H(u, c)$  are established in Section 3. It is shown that  $H(u, c)$  has an extremum at  $u=u^*$ , and that there is a finite real-valued  $c_0$ , such that  $H(u, c)$  for nonsingular problems has an isolated local minimum for  $u=u^*$  if  $c > c_0$ . Properties of  $H(u, c)$  for singular problems are also investigated in Section 3.

In Section 4 the multiplier function concept is illustrated with some simple examples. The choice of  $\mu(u)$  is discussed, and it is shown that the particular multiplier functions  $\mu(u) = \lambda^*$ , which is chosen by Hestenes, and  $\mu(u) = -[g_u(u)g_u^T(u)]^{-1}g_u(u)g_u^T(u)$ , which has been investigated by Mårtensson [4] and Fletcher [5], may be considered as special cases of this general approach.

Numerical methods for the minimization of  $H(u, c)$  are considered in Section 5. Straightforward minimization of the generalized Lagrangian with ordinary function minimization methods is compared with new algorithms based on properties of the multiplier function  $\mu(u)$ .

## 2. NECESSARY AND SUFFICIENT CONDITIONS FOR A CONSTRAINED LOCAL MINIMUM

In this section we state the necessary and sufficient conditions for a local isolated minimum. For proofs and a more detailed treatment, see e.g. [1], [6].

Introduce the Lagrangian  $L(u, \lambda)$  associated with the minimization problem formulated in Section 1.

$$L(u, \lambda) = f(u) + \sum_{i=1}^m \lambda^i g^i(u)$$

$\lambda^i$  are components of the  $m$ -dimensional vector  $\lambda$ , generally called the Lagrange multipliers.

We then have

Lemma 1 (First order necessary condition)

If

- i)  $f$  has a local minimum at  $u^*$  subject to the constraints  $g(u) = 0$ ,
- ii)  $f$  and  $g$  are once differentiable at  $u^*$ ,
- iii)  $g_u^i$ ,  $i = 1, \dots, m$ , are linearly independent at  $u^*$ ,

then there exists a unique  $m$ -dimensional vector  $\lambda^*$ , such that

$$L_u(u^*, \lambda^*) = 0$$

Notice that i) - iii) are sufficient conditions for the existence of finite

Lagrange multipliers  $\lambda^*$ . The constraint qualification iii) may for some problems be replaced by weaker conditions that are sufficient for the existence of  $\lambda^*$ . However, iii) is very useful from a computational point of view, and is assumed to hold in the sequel.

A stronger necessary condition for a minimum is given by the following second-order condition.

### Lemma 2

If  $f$  and  $g$  are twice continuously differentiable at  $u^*$ , and if the constraint qualification of Lemma 1 holds at  $u^*$ , then a necessary condition for  $u^*$  to be a local minimum, is the existence of a vector  $\lambda^*$ , such that

$$g(u^*) = 0$$

$$L_u(u^*, \lambda^*) = 0$$

Further, for every  $n$ -dimensional vector  $y$  such that  $g_u(u^*)y = 0$ ,

$$y^T L_{uu}(u^*, \lambda^*)y \geq 0$$

This can be strengthened to second-order sufficient conditions.

### Lemma 3

Sufficient conditions for  $u^*$  to be an isolated local minimum, are that

- i) the necessary conditions of Lemma 2 hold,
- ii) for every non-zero vector  $y$  such that  $g_u(u^*)y = 0$ ,
 
$$y^T L_{uu}(u^*, \lambda^*)y > 0$$

### 3. LAGRANGE MULTIPLIER FUNCTIONS

We now introduce the concept "Lagrange multiplier function".

#### Definition 1

Let  $\mu(u)$  be a real-valued  $m$ -dimensional vector defined on  $R^n$ . Then  $\mu(u)$  is a Lagrange multiplier function for the minimization problem if and only if

- i)  $\mu(u)$  exists and is twice differentiable in a neighbourhood of  $u^*$ ,
- ii)  $\mu(u^*) = \lambda^*$ ,
- iii) for every  $y \in R^n$ , such that  $y \neq 0$ ,  $g_u(u^*)y = 0$ ,  
and  $y^T L_{uu}(u^*, \lambda^*) y = 0$ ,  $\mu(u)$  satisfies

$$\{g_u(u^*)L_{uu}(u^*, \lambda^*) + g_u(u^*)g_u^T(u^*)\mu(u^*)\}y = 0$$

Condition iii) will prove to be necessary to handle singular problems. In iii) it is also assumed that  $f(u)$  and  $g(u)$  are at least twice differentiable at  $u=u^*$ . We assume throughout the paper that this holds in a neighbourhood of  $u^*$ .

With the properties of  $\mu(u)$  established, we define a "generalized Lagrangian"  $H(u, c)$  as follows.

#### Definition 2

The generalized Lagrangian  $H(u, c)$  associated with the minimization problem, is defined as

$$H(u, c) = f(u) + \mu^T(u)g(u) + cg^T(u)g(u)$$

where  $\mu(u)$  is an arbitrary multiplier function (Definition 1) and  $c$  is a real-valued parameter.

With the assumptions made about  $f(u)$ ,  $g(u)$  and  $\mu(u)$ ,  $H(u, c)$  exists and is twice differentiable in a neighbourhood of  $u^*$ .

In the following theorems we will establish some important properties of  $H$ .



### Theorem 1

For any value of the parameter  $c$ , the generalized Lagrangian  $H(u, c)$  has a stationary point at  $u=u^*$ .

Proof: A straightforward differentiation yields

$$H_u = f_u + \mu^T g_u + g^T \mu_u + 2cg^T g_u$$

Since  $g(u^*) = 0$ , and, according to Lemma 1 and to the definition of  $\mu(u)$ ,

$$f_u(u^*) + \mu^T(u^*)g_u(u^*) = f_u(u^*) + (\lambda^*)^T g_u(u^*) = 0$$

it follows that  $H_u(u^*, c) = 0, \forall c$  ■

Intuitively it now seems reasonable that the stationary point  $u=u^*$  can be made a minimum point by choosing the parameter  $c$  large enough. To prove this, we have to distinguish between nonsingular and singular problems.

### Theorem 2

Let  $u^*$  be a local isolated minimum of  $f(u)$  subject to the constraints  $g(u) = 0$ , and assume that the sufficient conditions of Lemma 3 are satisfied. Then there exists a real-valued parameter  $c_0$ , such that  $H_u(u^*, c) = 0$  and  $H_{uu}(u^*, c) > 0$  for  $c > c_0$ .

Proof: In Theorem 1 it was shown that  $H_u(u^*, c) = 0$  independent of  $c$ . Then consider  $H_{uu}(u, c)$ .

$$H_{uu}(u, c) = f_{uu} + \sum_{i=1}^m \mu^i g_{uu}^i + \mu_u^T g_u + g_u^T \mu_u + \sum_{i=1}^m g^i \mu_{uu}^i + 2cg_u^T g_u + 2c \sum_{i=1}^m g^i g_{uu}^i$$

For  $u=u^*$ , this reduces to

$$H_{uu}(u^*, c) = f_{uu} + \sum_{i=1}^m \mu_i^i g_{uu}^T + \mu_u^T g_u + g_u^T \mu_u + 2c g_u^T g_u$$

or

$$H_{uu}(u^*, c) = L_{uu}(u^*, \lambda^*) + \mu_u^T g_u + g_u^T \mu_u + 2c g_u^T g_u$$

Now let  $Q$  be the subspace of  $R^n$  spanned by the rows of  $g_u(u^*)$ , and let  $Q^\perp$  be the orthogonal complement. Since the constraint qualifications are assumed to hold at  $u^*$ ,  $Q$  has dimension  $m$ . If  $y_1 \in Q$  and  $y_2 \in Q^\perp$ , we then have  $y_1^T y_2 = 0$ ,  $g_u y_2 = 0$  and  $y_1 = g_u^T \alpha$ , where  $\alpha \in R^m$  is uniquely determined by  $y_1$ . Similarly, we can choose an arbitrary basis  $e_1, \dots, e_{n-m}$  in  $Q^\perp$ . Then any  $y_2 \in Q^\perp$  may be written

$$y_2 = \sum_{i=1}^{n-m} \beta_i e_i \quad \text{or} \quad y_2 = G\beta$$

where  $\beta \in R^{n-m}$  and  $G$  is an  $n \times (n-m)$ -dimensional matrix of rank  $n-m$ . Conversely,  $y_2 = G\beta$  lies in  $Q^\perp$  for any  $\beta \in R^{n-m}$ . Then we may write an arbitrary vector  $y \in R^n$  in the form

$$y = g_u^T \alpha + G\beta$$

Now consider  $y^T H_{uu}(u^*, c)y$ .

$$\begin{aligned} y^T H_{uu}(u^*, c)y &= (g_u^T \alpha + G\beta)^T H_{uu}(u^*, c)(g_u^T \alpha + G\beta) = \\ &= \alpha^T \{2c g_u^T g_u g_u^T + g_u^T L_{uu} g_u^T + \\ &\quad + g_u^T \mu_u g_u^T + g_u^T \mu_u g_u^T\} \alpha + \\ &\quad + \alpha^T \{g_u^T L_{uu} G + g_u^T \mu_u G\} \beta + \\ &\quad + \beta^T \{G^T L_{uu} g_u^T + G^T \mu_u g_u^T\} \alpha + \\ &\quad + \beta^T \{G^T L_{uu} G\} \beta \end{aligned}$$

where all quantities are evaluated at  $u=u^*$ .

To get a better survey, we introduce

$$A(c) = 2c g_u^T g_u^T + g_u^T L_{uu} g_u^T + g_u^T \mu_u^T g_u^T + g_u^T \mu_u^T g_u^T$$

$$B = g_u^T L_{uu} G + g_u^T \mu_u^T G$$

$$D = G^T L_{uu} G$$

Then

$$y^T H_{uu}(u^*, c) y = \begin{bmatrix} \alpha^T & \beta^T \end{bmatrix} \begin{bmatrix} A(c) & B \\ B^T & D \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

It now remains to prove the existence of a parameter  $c_0$ , such that

$$\begin{bmatrix} A(c) & B \\ B^T & D \end{bmatrix} > 0$$

for  $c > c_0$ .

This will be done in three steps. First  $A(c)$  is considered, and it is shown that for every real  $k > 0$ , there exists a  $c_0(k)$ , such that  $A(c) > kI_m$  for  $c > c_0(k)$ . Then we will prove that  $D > 0$ , and finally it will be shown that

$$\begin{bmatrix} kI_m & B \\ B^T & D \end{bmatrix} > 0$$

for  $k$  large enough.

Consider

$$\begin{aligned} A(c) - kI_m &= 2c(g_u^T g_u^T)(g_u^T g_u^T) + g_u^T L_{uu} g_u^T + \\ &+ g_u^T \mu_u^T g_u^T + g_u^T \mu_u^T g_u^T - kI_m \end{aligned}$$

Since the constraint qualifications hold at  $u^*$ ,  $g_u g_u^T$  is nonsingular, and  $(g_u g_u^T)(g_u g_u^T)$  is positive definite symmetric. Then there exists a nonsingular transformation  $S(k)$  such that [7]

$$2(g_u g_u^T)(g_u g_u^T) = S^T(k)S(k)$$

and

$$\begin{aligned} g_u L_{uu} g_u^T + g_u \mu_u^T g_u g_u^T + g_u g_u^T \mu_u g_u^T - kI_m = \\ = S^T(k) \begin{bmatrix} a_1(k) & & 0 \\ & \ddots & \\ 0 & & a_m(k) \end{bmatrix} S(k) \end{aligned}$$

This yields

$$A(c) - kI_m = S^T(k) \begin{bmatrix} c+a_1(k) & & 0 \\ & \ddots & \\ 0 & & c+a_m(k) \end{bmatrix} S(k)$$

and thus  $A(c) - kI_m > 0$  for

$$c > -\min_i a_i(k).$$

Next consider the matrix  $D$ . For every  $y_2 \in Q^\perp$ ,  $y_2 \neq 0$ , we have

$$y_2^T L_{uu}(u^*, \lambda^*) y_2 > 0$$

according to Lemma 3. But any  $y_2$  in  $Q^\perp$  may be written  $y_2 = G\beta$ , and any vector  $G\beta$  lies in  $Q^\perp$ . Then

$$\beta^T G^T L_{uu}(u^*, \lambda^*) G \beta > 0$$

for  $\beta \neq 0$ , which proves that  $D = G^T L_{uu}(u^*, \lambda^*) G$  is positive definite.

Finally we will consider the matrix

$$\begin{bmatrix} kI_m & B \\ B^T & D \end{bmatrix}$$

Introduce the nonsingular transformation

$$L(k) = \begin{bmatrix} I_m & -\frac{1}{k}B \\ 0 & I_{n-m} \end{bmatrix}$$

Then

$$L^T(k) \begin{bmatrix} kI_m & B \\ B^T & D \end{bmatrix} L(k) = \begin{bmatrix} kI_m & 0 \\ 0 & \frac{1}{k}(kD - B^T B) \end{bmatrix}$$

and thus it is sufficient to prove that  $kD - B^T B > 0$  for  $k$  large enough. But  $D$  is positive definite, and then there exists a nonsingular transformation  $T$ , such that

$$D = T^T T$$

and

$$B^T B = T^T \begin{bmatrix} b_1 & & 0 \\ & \ddots & \\ 0 & & b_m \end{bmatrix} T$$

Thus we have  $kD - B^T B > 0$  for

$$k > \max_i b_i$$

This completes the proof of the existence of a finite  $c_0$ , such that  $H_{uu}(u^*, c) > 0$  for  $c > c_0$ . ■

The theorem evidently breaks down if the problem is singular, i. e.  $D$  is only nonnegative definite. To be able to extend the multiplier function

concept to this case, it is natural to require that  $H(u, c)$  should have the properties

$$H_u(u^*, c) = 0$$

$$H_{uu}(u^*, c) \geq 0$$

independent of the character of the optimal solution. We will now prove, that this is attained by including the condition iii) in the definition of the multiplier functions.

### Theorem 3

Let  $u^*$  be a local minimum of  $f(u)$  subject to the constraints  $g(u) = 0$ . Then there exists a  $c_0$ , such that  $H_u(u^*, c) = 0$  and  $H_{uu}(u^*, c) \geq 0$  for  $c > c_0$ .

Proof: It is necessary and sufficient to prove that

$$kD - B^T B \geq 0$$

for  $k$  large enough. The theorem will then follow from the proof of Theorem 2.

Since  $D$  may be singular, a necessary condition obviously is

$$B\beta = 0$$

for every  $\beta \in R^{n-m}$ , such that

$$\beta^T D \beta = 0$$

But from the definition of  $B$ ,  $D$  and  $\beta$  follows that this is equivalent to

$$(g_u^T L_{uu} + g_u^T g_u^T \mu_u) y = 0$$

for every  $y \in R^n$ , such that

$$g_u y = 0$$

and

$$y^T L_{uu} y = 0$$

Thus condition iii) in Definition 1 is a necessary condition for  $H_{uu}(u^*, c) \geq 0$ . To prove that iii) is a sufficient condition (together with i) and ii), and to get a measure of  $k$ , assume that  $\text{rank } D = r$ ,  $0 < r < (n-m)$ . Then  $D$  may be written

$$D = D_1^T D_1$$

where  $D_1$  is an  $r \times (n-m)$  matrix of rank  $r$ . Let  $P$  be the subspace of  $R^{n-m}$  spanned by the rows of  $D_1$ , and let  $P^\perp$  be the orthogonal complement. Then every  $\beta \in R^{n-m}$  may be uniquely decomposed into

$$\beta = D_1^T \gamma + \beta_2$$

where  $D_1^T \gamma \in P$  and  $\beta_2 \in P^\perp$ . Since  $D_1 \beta_2 = 0$ ,  $\forall \beta_2 \in P^\perp$ , it follows that  $\beta_2^T D_1^T D_1 \beta_2 = \beta_2^T D \beta_2 = 0$ , and thus  $B \beta_2 = 0$  according to condition iii). Then

$$\begin{aligned} \beta^T (kD - B^T B) \beta &= (D_1^T \gamma + \beta_2)^T (kD_1^T D_1 - B^T B) (D_1^T \gamma + \beta_2) = \\ &= \gamma^T \left[ k(D_1 D_1^T) (D_1 D_1^T) - D_1 B^T B D_1^T \right] \gamma \end{aligned}$$

But  $(D_1 D_1^T) (D_1 D_1^T)$  is positive definite symmetric and then there exists a nonsingular transformation  $V$  such that

$$(D_1 D_1^T) (D_1 D_1^T) = V^T V$$

and

$$D_1 B^T B D_1^T = V^T \begin{bmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_r \end{bmatrix} V$$

Thus

$$\beta^T (kD - B^T B) \beta \geq 0, \quad \beta \in R^{n-m}$$

for

$$k \geq \max_i s_i.$$

We have then proved the existence of a finite  $c_0$ , such that  $H_{uu}(u^*, c) \geq 0$  for  $c > c_0$ . ■

#### 4. EXAMPLES

To illustrate the multiplier function concept, some particular choices of  $\mu(u)$  are investigated in this section. We will also try to make clear by examples, how different choices of  $\mu(u)$  may result in different properties of the generalized Lagrangian. The possibility to generate different generalized Lagrangians is of great importance for the numerical solution of the optimization problem.

##### Example 1

Assume that the optimal multipliers  $\lambda^*$  are à priori known. A simple choice of the multiplier function might then be

$$\mu(u) = \lambda^* = \text{const.}$$

This special case has been considered by Hestenes [3]. It may seem strange to assume the optimal multipliers to be à priori known, when the optimal solution is not known. The reason for this will become clear in the next section, where computational methods based on successive estimations of  $\lambda^*$  are considered. It is then important to establish properties of

$$H_1(u, c) = f(u) + \lambda_1^T g(u) + c g^T(u) g(u)$$

as  $\lambda_1$  tends to the optimal multipliers  $\lambda^*$ .

For nonsingular problems,  $\mu(u) = \lambda^*$  obviously is an admissible multiplier function since it trivially satisfies conditions i) and ii) of the multiplier function definition. In the singular case, it depends on the particular problem whether  $\mu(u) = \lambda^*$  satisfies condition iii) or not. This condition becomes particularly simple for this multiplier function. Consider

$$(g_u^T L_{uu} + g_u^T g_u^T \mu_u) y_2 = g_u^T L_{uu} y_2 = 0$$



Thus  $L_{uu} y_2 \in Q^\perp$ , where  $Q^\perp$  is the orthogonal complement of the subspace spanned by the rows of  $g_u$ . But in the singular case, there exists  $y_2 \in Q^\perp$ ,  $y_2 \neq 0$ , such that

$$y_2^T L_{uu} y_2 = 0$$

and thus  $L_{uu} y_2 \in Q$ . Since  $Q \cap Q^\perp = \{0\}$ , the condition iii) then reduces to  $L_{uu} y_2 = 0$ .

Consider the following simple singular problem. Minimize

$$f(u) = (u_1 - u_2)^2$$

subject to the constraint

$$g(u) = u_1 - u_2 = 0 \quad (\text{Problem 1})$$

Choosing  $y_2^T = (\alpha, \alpha)$ ,  $g_u y_2 = 0$  and  $y_2^T L_{uu} y_2 = 0$  for any value of  $\alpha$ .

But  $L_{uu} y_2 = 0$  and thus  $\mu(u) = \lambda^* = 0$  is an admissible multiplier function. In this case the generalized Lagrangian is  $H(u, c) = (1+c)(u_1 - u_2)^2$ .

For the following problem  $\mu(u) = \lambda^*$  is not a multiplier function. Minimize

$$f(u) = u_1^2 - u_2^2$$

subject to

$$g(u) = u_1 + u_2 = 0 \quad (\text{Problem 2})$$

Choosing  $y_2^T = (\alpha, -\alpha)$ ,  $g_u y_2 = 0$ ,  $y_2^T L_{uu} y_2 = 0$  but  $L_{uu} y_2 \neq 0$ ,  $\alpha \neq 0$ .

The generalized Lagrangian becomes  $H(u, c) = u_1^2 - u_2^2 + \lambda^*(u_1 + u_2) + c(u_1 + u_2)^2$ , or

$$H(u, c) = \frac{1}{2} (u - u^*)^T \begin{bmatrix} 2c+2 & 2c \\ 2c & 2c-2 \end{bmatrix} (u - u^*)$$

where  $u^*$  is the optimal solution corresponding to the particular choice of  $\lambda^*$  ( $\lambda^*$  turns out to be arbitrary). It is easily verified, that for this problem, the optimal solution  $u^*$  cannot be made a minimum of  $H(u, c)$  by choosing  $c$  large enough. To prove this, choose

$$\bar{u}^T = \{u_1^* - 2c\alpha/(2c+2), u_2^* + \alpha\}. \text{ Then}$$

$$H(\bar{u}, c) - H(u^*, c) = -4\alpha^2$$

for any finite  $c$ , and thus  $H(u^*, c) > H(\bar{u}, c)$  if  $\alpha \neq 0$ .

### Example 2

To avoid the trouble associated with singular problems, one obviously should look for a multiplier function that satisfies the conditions i) - iii) for any character of the problem. One possibility to achieve this is to choose

$$\mu(u) = - \{g_u(u)g_u^T(u)\}^{-1} g_u(u)f_u^T(u)$$

This multiplier function has been investigated by Mårtensson [4] and Fletcher [5]. Assuming  $f(u)$  and  $g(u)$  three times differentiable in a neighbourhood of  $u^*$ , it can be shown [4], that  $\mu(u)$  exists and is twice differentiable in a neighbourhood of  $u^*$ , that  $\mu(u^*) = \lambda^*$ , and that  $g_u(u^*)L_{uu}(u^*, \lambda^*) + g_u(u^*)g_u^T(u^*)\mu_u(u^*) = 0$ . Thus  $\mu(u)$  is an admissible multiplier function for both singular and nonsingular problems.

With this choice of multiplier function, we get the same generalized Lagrangian for Problem 1 as in the previous example, namely  $H(u, c) = (1+c)(u_1 - u_2)^2$ . However, for Problem 2, the generalized Lagrangian now becomes  $H(u, c) = c(u_1 + u_2)^2$ , and this possesses all the desired properties.

We will illustrate the multiplier function concept with one more simple problem. This also illustrates the possibility of handling inequality constraints by means of slack variables. Minimize

$$f(u) = -\frac{16}{3}u_1^3 - 2u_1^2 + 2u_1$$

subject to the constraint

$$u_1 - 1 \leq 0$$

(Problem 3)

The problem has two local isolated minima, one at the constraint  $u_1=1$ , and one at  $u_1=-0.5$ . The inequality constraint may be transformed into an equality constraint by adding a slack variable  $u_2$  in such a way that the constraint qualifications of Lemma 1 are satisfied, e.g.

$$g(u) = u_1 - 1 + u_2^2 = 0$$

In Fig. 1, contour levels of  $H(u, c)$  are drawn for  $c=0, 1.0$  and  $5.0$ . Since  $H$  is symmetric with respect to  $u_2$ , the contour levels are drawn only for  $u_2 \geq 0$ . From the figure it is clear that any minimization method should be able to reach one of the minima if we choose  $c$  large enough, and if the initial guess is not too far from the curve representing the equality constraint  $g(u) = 0$ . For further examples of the slack variable technique, we refer to [4].

### Example 3

In this example we will indicate some obvious generalizations of the preceding example.

One drawback of the multiplier function

$$\mu(u) = - (g_u^T g_u)^{-1} g_u^T f_u$$

is that  $(g_u^T g_u)$  may be singular for some  $u$  outside the equality constraint. A possible way to overcome this, is to choose

$$\mu(u) = - (g_u^T g_u + g^T g I_m)^{-1} g_u^T f_u$$

It is easily verified that this is an admissible multiplier function for both singular and nonsingular problems.

It is also clear from Fig. 1, that one may get into trouble if the initial guess of the optimal solution is too far away from the curve representing the constraint. To get the right slope of  $H(u, c)$ , but preserving its smooth character around the optimal solution, one could select

$$\mu(u) = - (g_u^T g_u)^{-1} g_u^T f_u + (g^T g)^{-1} g$$

If  $m \geq 0$   
 In Fig  
 illustr  
 the m

to get  
 in Fig  
 is co

$u_2$   
 1.5

1.0

0.5

Fig

The  
 fun  
 wh  
 not  
 fur

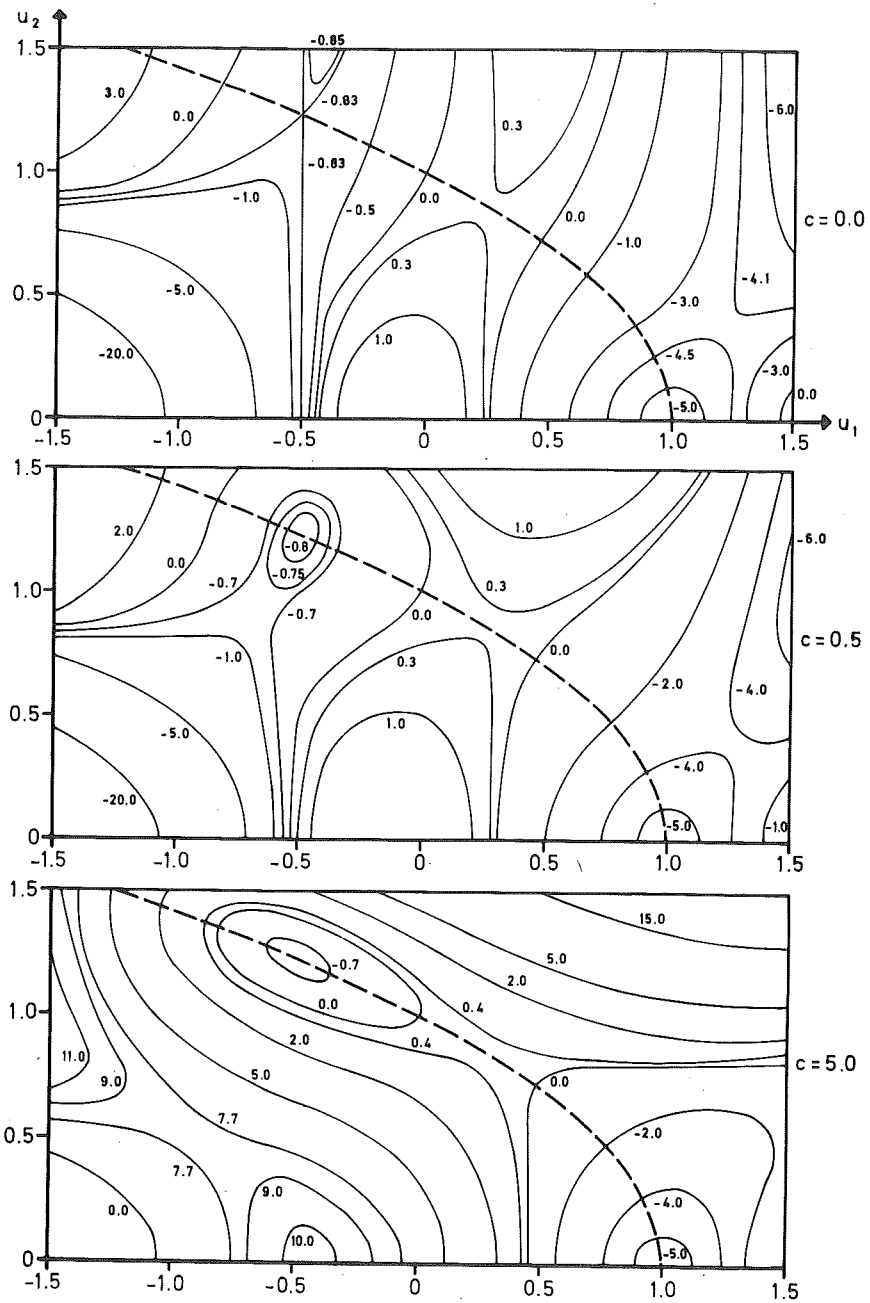


Fig. 1 - Contour levels of the generalized Lagrangian for Problem 3:

$$\mu(u) = - (g_u^T g_u^T)^{-1} g_u^T f_u^T$$

If  $\mu \geq 0$ ,  $\mu(u)$  is an admissible multiplier function.

In Fig. 2, the possibility to improve the properties of  $H(u, c)$  is illustrated. For Problem 3 we have now (rather arbitrarily) chosen the multiplier function

$$\mu(u) = - (g_u^T g_u + 5g^T g)^{-1} g_u^T f_u^T + (g^T g)g + 0.1f^2 g$$

to get the right slope of  $H(u, c)$ . Contour levels of  $H(u, c)$  are drawn in Fig. 2 for  $c = 0.5$ , and comparing with Fig. 1, the improvement is considerable.

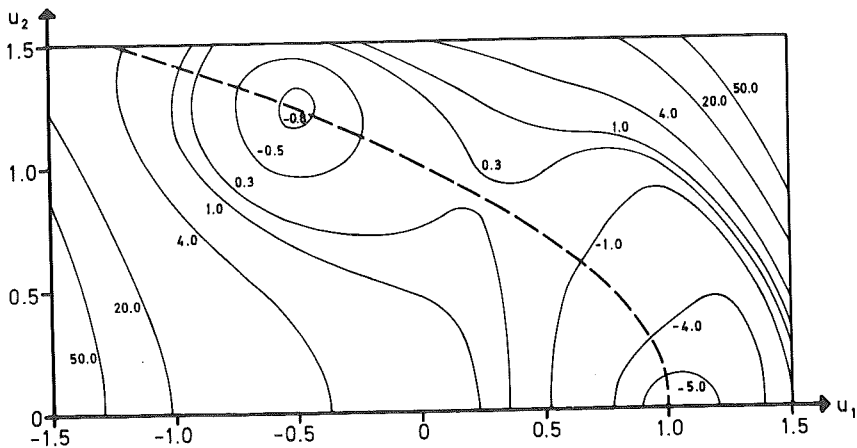


Fig. 2 - Contour levels of the generalized Lagrangian for Problem 3:

$$\mu(u) = - (g_u^T g_u + 5g^T g)^{-1} g_u^T f_u^T + (g^T g)g + 0.1f^2 g, \quad c = 0.5.$$

The multiplier functions considered in this section have been explicit functions with similar basic structure. An interesting problem is then, whether there exist multiplier functions with different structures or not. It may also be of great interest to try to define the multiplier function implicitly. These problems have not yet been investigated.

## 5. ALGORITHMS AND COMPUTATIONAL ASPECTS

A number of different algorithms can be designed for the minimization of the generalized Lagrangian  $H(u, c)$ . Roughly they can be separated into two major classes, direct minimization of  $H(u, c)$  with ordinary function minimization methods, and iterative estimation of the multiplier function  $\mu(u)$ . Although the latter methods require iteration in a larger space, they can be made very efficient by using the function properties of the multiplier  $\mu(u)$ .

### 5.1. Direct Minimization Methods

A straightforward way to minimize  $H(u, c)$  is to use a minimization method where only function value evaluations are required, e.g. the methods of Powell [8] and Stewart [9]. Stewart's method, which is a modification of Davidon's method [10], is generally considered to be somewhat more efficient since difference approximations of the derivatives can be used. However, as was illustrated in Section 4, the multiplier functions must be chosen carefully since we do not have any a priori estimate of the parameter  $c$ .

This problem can to some extent be avoided if it is possible to evaluate  $\mu_u$ . Then the derivative

$$H_u = f_u + \mu^T g_u + g^T \mu_u + 2cg^T g_u$$

of  $H(u, c)$  can be computed, and a more efficient minimization method can be used, e.g. the method of Fletcher and Powell [11]. But it will also be possible to get an a priori estimate of  $c$ , and thereby make the minimization of  $H(u, c)$  less sensitive to the particular choice of  $\mu(u)$ .

Consider the quantity  $g^T(u)g(u)$ , which equals zero if and only if the constraints are satisfied. If it is required that  $H(u, c)$  has the property

$$\left[ \frac{d}{du} (g^T g) \right] H_u^T > 0$$

then the magnitude of  $g^T g$  can be decreased by moving in the direction opposite to  $H_u$ , that is, in the steepest descent direction. But

$$\left[ \frac{d}{du} (g^T g) \right] H_u^T = 2g_u^T (f_u + \mu g_u + g \mu_u + 2cg_u^T g_u)^T$$

and then this condition is satisfied if

$$c > - \frac{g_u^T g_u (f_u^T + g_u^T \mu + \mu_u^T g)}{2g_u^T g_u g_u^T}$$

for  $g(u) \neq 0$ . Further properties of this measure are discussed in [4].

## 5.2. Multiplier Estimation Methods

An obvious disadvantage of direct minimization methods is the time-consuming function and gradient evaluations that have to be carried out for every step. It then seems reasonable that methods based on iterative estimation of the optimal multipliers could be made more efficient than the direct minimization methods.

In this section, different estimation algorithms are derived, and they will be classified according to their convergence properties for quadratic functions with linear constraints.

Consider the following simple iteration scheme:

Make an initial guess of  $\lambda^*$ , say  $\mu_k$  (subscripts will be used to denote the iteration step). Then minimize  $F(u, \mu_k) = f(u) + \mu_k^T g(u) + cg^T(u)g(u)$  with an efficient minimization method. Assume that the minimum occurs for  $u = u_{k+1}$ . If  $g(u_{k+1}) \leq \delta$ , where  $\delta$  is a small quantity, then  $u_{k+1}$  is the optimal solution. Otherwise compute a new estimate  $\mu_{k+1} = \mu(u_{k+1})$  and repeat the procedure.

Notice that the algorithm does not depend on any particular choice of the multiplier function  $\mu(u)$ . It is then natural to examine if  $\mu(u)$  can be selected so that the algorithm is further simplified.

Assume that  $u_{k+1}$  minimizes  $F(u, \mu_k)$ . Then

$$f_u(u_{k+1}) + \mu_k^T g_u(u_{k+1}) + 2cg^T(u_{k+1})g_u(u_{k+1}) = 0$$

and post-multiplying by  $g_u^T(u_{k+1}) [g_u(u_{k+1}) g_u^T(u_{k+1})]^{-1}$ , we get

$$f_u(u_{k+1}) g_u^T(u_{k+1}) [g_u(u_{k+1}) g_u^T(u_{k+1})]^{-1} + \mu_k^T + 2cg^T(u_{k+1}) = 0$$

From this we conclude that  $\mu(u) = - (g_u g_u^T)^{-1} g_u^T f_u^T$  is a suitable choice of the multiplier function, since the new estimate  $\mu_{k+1}$  then satisfies

$$-\mu_{k+1}^T + \mu_k^T + 2cg^T(u_{k+1}) = 0$$

or

$$\mu_{k+1} = \mu_k + 2cg(u_{k+1})$$

This will drastically reduce the computations involved. It is also interesting to notice, that this recursive relation, which also has been suggested by Hestenes [3], can be considered as a special case of a more general estimation algorithm.

The algorithm can then be summarized as follows:

First order algorithm:

- a) Set  $\mu_k = 0$ .
- b) Minimize  $F(u, \mu_k) = f(u) + \mu_k^T g(u) + cg^T(u)g(u)$  with an ordinary function minimization algorithm, e.g. Fletcher-Powell. Notice that the evaluations of the function value and of the gradient are very simple. Assume that the minimum occurs for  $u = u_{k+1}$ .
- c) If  $\|g(u_{k+1})\| \leq \delta$ , where  $\delta$  is a small quantity, then  $u^* = u_{k+1}$ .
- d) If  $\|g(u_{k+1})\| > \delta$ , set  $\mu_{k+1} = \mu_k + 2cg(u_{k+1})$  and return to b).

It is possible to establish convergence properties of the algorithm for quadratic functions with linear constraints.



#### Theorem 4

Let  $f(u)$  be quadratic, and assume that the constraint  $g(u)$  is linear. If  $\mu(u) = \lambda^*$  is an admissible multiplier function for the problem, the algorithm converges to the optimal solution for  $c > \max(0, 2c_0)$ , where  $c_0$  is defined in theorem 2, and refers to the multiplier function  $\mu(u) = \lambda^*$ .

Proof: Consider the situation at stage  $k$ . Since  $u_{k+1}$  minimizes  $F(u, \mu_k)$ , we have

$$f_u(u_{k+1}) + \mu_k^T g_u(u_{k+1}) + 2cg^T(u_{k+1})g_u(u_{k+1}) = 0$$

or

$$f_u(u_{k+1}) + \mu_{k+1}^T g_u(u_{k+1}) = 0$$

At stage  $k+1$ ,  $u_{k+2}$  minimizes  $F(u, \mu_{k+1})$ , and

$$f_u(u_{k+2}) + \mu_{k+1}^T g_u(u_{k+2}) + 2cg^T(u_{k+2})g_u(u_{k+2}) = 0$$

Combining these conditions, and expanding  $f(u)$  and  $g(u)$  yields

$$2cg^T g_u = (u_{k+1} - u_{k+2})^T f_{uu} - 2cu_{k+2}^T g_u^T g_u$$

where  $g$  is evaluated at  $u=0$ . Then consider the identity

$$\begin{aligned} c \left[ g^T(u_{k+2})g(u_{k+2}) - g^T(u_{k+1})g(u_{k+1}) \right] &= \\ &= cu_{k+2}^T g_u^T g_u u_{k+2} - cu_{k+1}^T g_u^T g_u u_{k+1} + 2cg^T g_u [u_{k+2} - u_{k+1}] \end{aligned}$$

Insert the expression for  $2cg^T g_u$  and rearrange the terms to get

$$\begin{aligned} c \left[ g^T(u_{k+2})g(u_{k+2}) - g^T(u_{k+1})g(u_{k+1}) \right] &= \\ &= - (u_{k+2} - u_{k+1})^T (f_{uu} + cg_u^T g_u) (u_{k+2} - u_{k+1}) \end{aligned}$$

Since  $\mu(u) = \lambda^*$  is assumed to be an admissible multiplier function,

$$f_{uu} + cg_u^T g_u \geq 0$$

for  $c > 2c_0$  according to Theorem 3, and we then have to investigate two different cases separately.

Assume that  $f_{uu} + cg_u^T g_u > 0$  for  $c > 2c_0$ . Then

$$g^T(u_{k+2})g(u_{k+2}) < g^T(u_{k+1})g(u_{k+1})$$

for  $c > \max(0, 2c_0)$  provided that  $u_{k+2} \neq u_{k+1}$ . Since

$\|g(\cdot)\| \geq 0$ ,  $g^T(u_i)g(u_i)$  will converge either to zero or to a

finite  $G > 0$ . In the latter case we get  $u_{i+2} = u_{i+1}$  and  $\mu_{i+2} = \mu_{i+1}$ .

But  $\mu_{i+2} = \mu_{i+1} + 2cg(u_{i+2})$ , which proves that  $u_{i+2} = u_{i+1}$  if and only if  $g(u_{i+2}) = g(u_{i+1}) = 0$ . Thus the algorithm converges for non-singular problems.

Then consider the singular case, that is, there exists  $y_2 \neq 0$ ,

such that  $y_2^T (f_{uu} + cg_u^T g_u) y_2 = 0$ ,  $c > \max(0, 2c_0)$ . Then, according

to the multiplier function definition and to Theorem 3,  $g_u y_2 = 0$

and  $L_{uu} y_2 = f_{uu} y_2 = 0$ . This implies that

$$2cg_u^T g_u = (u_{k+1} - u_{k+2})^T f_{uu} - 2cu_{k+2}^T g_u^T g_u$$

reduces to

$$2cg_u^T(u_{k+2})g_u = 0$$

and then  $g(u_{k+2}) = 0$ ,  $c > 0$ , since  $g_u$  is assumed to satisfy the constraint qualifications, i. e. to have full rank. This completes the convergence proof of algorithm. ■

It is instructive to verify the convergence for the following simple example. Minimize

$$f(u) = u_1^2 - u_2^2$$

subject to

$$u_1 - 2u_2 - 2 = 0$$

For this problem  $c_0 = \frac{1}{3}$ , and for  $c > \frac{2}{3}$  the algorithm converges to the optimal solution  $u_1 = -\frac{2}{3}$ ,  $u_2 = -\frac{4}{3}$ . For  $c < \frac{2}{3}$  the algorithm diverges, while for  $c = \frac{2}{3}$  the quantity  $g^T(u_1)g(u_1)$  is constant.

The convergence rate depends on the choice of  $c$ . Introduce  $\epsilon_k = \|g(u_k)\|$  and choose  $c=1$ . The following residuals are then obtained:

$$\epsilon_{k+1} = \frac{1}{2} \epsilon_k \qquad \epsilon_1 = 1$$

Increasing  $c$  to  $c=5$ , convergence is improved considerably, and the residuals are

$$\epsilon_{k+1} = \frac{1}{14} \epsilon_k \qquad \epsilon_1 = \frac{1}{7}$$

So far, the multiplier function concept has been used only for estimation of the optimal multipliers. The properties of  $\mu(u)$  will now be exploited to develop second order algorithms, that is, algorithms with one-step convergence for linear-quadratic problems. The following theorem will be required.

### Theorem 5

Let  $u^*$  minimize  $f(u)$  subject to the constraints  $g(u) = 0$ , and assume that the sufficient conditions of Lemma 3 are satisfied. Define the projection  $P(u)$  as

$$P(u) = I_n - g_u^T(u) \left[ g_u(u) g_u^T(u) \right]^{-1} g_u(u)$$

Then

$$P(u^*) L_{uu}(u^*, \lambda^*) + 2c g_u^T(u^*) g_u(u^*)$$

is nonsingular for  $c \neq 0$ .

Proof: The theorem is proved by contradiction. Assume that

$PL_{uu} + 2cg_u^T g_u$  is singular, that is, there exists  $z \neq 0$ , such that

$z^T (PL_{uu} + 2cg_u^T g_u) = 0$ . Decompose  $z$  into  $z = g_u^T \alpha + z_2$ , where

$g_u^T \alpha \in Q$ , the space spanned by the rows of  $g_u$ , and  $z_2 \in Q^\perp$ . Then

$z^T (PL_{uu} + 2cg_u^T g_u) = 0$  is equivalent to

$$2c\alpha^T (g_u g_u^T) g_u + z_2^T L_{uu} = 0$$

since  $g_u P = 0$  and  $z_2^T P = z_2^T$ . Now assume that there exists  $z_2 \neq 0$ ,

such that  $2c\alpha^T (g_u g_u^T) g_u + z_2^T L_{uu} = 0$ . Postmultiplying by  $z_2$  then

yields

$$z_2^T L_{uu} z_2 = 0$$

which contradicts the assumption that  $z_2^T L_{uu} z_2 > 0$  for  $z_2 \neq 0$  (Lemma 3).

If  $z_2 = 0$ ,  $\alpha \neq 0$ , then  $\alpha^T (g_u g_u^T) g_u = 0$  for  $c \neq 0$ , which contradicts the constraint qualification, i. e. the linear independence of the rows of  $g_u$ .

Thus, for  $c \neq 0$ , there is no nonzero solution, and  $P(u^*)L_{uu}(u^*, \lambda^*) + 2cg_u^T(u^*)g_u(u^*)$  is nonsingular. ■

### Corollary

If  $f(u)$  is quadratic and the constraints  $g(u)$  are linear, then

$$Pf_{uu} + 2cg_u^T g_u$$

is nonsingular for  $c \neq 0$ .

Proof:  $Pf_{uu} + 2cg_u^T g_u$  is independent of  $u$  and is nonsingular for  $u = u^*$ . ■

Using these results, we will now design a second order estimation method. Assume that an estimate  $\mu_k$  of the optimal multipliers

$\lambda^* = \mu_k + \delta\mu_k$  is available. Let  $u^* = u_{k+1} + \delta u_{k+1}$ , where  $u_{k+1}$

minimized  $F(u, \mu_k)$ , be the optimal solution. Approximate  $F_u(u^*, \lambda^*)$  with a first order series expansion about  $u_{k+1}, \mu_k$ . Then

$$F_u(u^*, \lambda^*) = F_u(u_{k+1}, \mu_k) + F_{uu}(u_{k+1}, \mu_k) \delta u_{k+1} + F_{u\mu}(u_{k+1}, \mu_k) \delta \mu_k = 0$$

Since  $F_u(u_{k+1}, \mu_k) = 0$ , this reduces to

$$F_{uu}(u_{k+1}, \mu_k) \delta u_{k+1} + F_{u\mu}(u_{k+1}, \mu_k) \delta \mu_k = 0$$

In contrast to the first order algorithm, the quantity  $\delta \mu_k$  is now unknown. But  $\mu(u)$  is a function of  $u$ , and so we make the following approximation of  $\delta \mu_k$

$$\delta \mu_k = \mu(u_{k+1}) - \mu(u_k) + \mu_u(u_{k+1}) \delta u_{k+1}$$

Now choose the particular multiplier function  $\mu(u) = - (g_u^T g_u^T)^{-1} g_u^T f_u^T$ . Then

$$\delta \mu_k = 2c g(u_{k+1}) + \mu_u(u_{k+1}) \delta u_{k+1}$$

Inserting this into the series expansion, we get

$$F_{uu} \delta u_{k+1} + F_{u\mu} \delta \mu_k = F_{uu} \delta u_{k+1} + 2c F_{u\mu} g + F_{u\mu} \mu_u \delta u_{k+1} = 0$$

where all quantities are evaluated at  $u_{k+1}, \mu_k$ . Noticing that

$F_{u\mu} = g_u^T$ , this is equivalent to

$$\left[ f_{uu} + \sum_{i=1}^m \mu_k^i g_{uu}^i + 2c g_u^T g_u + 2c \sum_{i=1}^m g_u^i g_{uu}^i + g_u^T \mu_u \right] \delta u_{k+1} + 2c g_u^T g = 0$$

or

$$\left[ f_{uu} + \sum_{i=1}^m \mu_{k+1}^i g_{uu}^i + 2cg_u^T g_u + g_u^T \mu_u \right] \delta u_{k+1} + 2cg_u^T g = 0$$

To evaluate  $\mu_u(u_{k+1})$ , we have to differentiate the multiplier function

$\mu(u) = - (g_u g_u^T)^{-1} g_u^T f_u^T$ . It is then easily verified that

$$\mu_u(u_{k+1}) = - (g_u g_u^T)^{-1} g_u \left[ f_{uu} + \sum_{i=1}^m \mu_{k+1}^i g_{uu}^i \right]$$

and thus

$$\left[ P \left[ f_{uu} + \sum_{i=1}^m \mu_{k+1}^i g_{uu}^i \right] + 2cg_u^T g_u \right] \delta u_{k+1} + 2cg_u^T g = 0$$

where  $P$  is the projection matrix.

$$P = I_n - g_u^T (g_u g_u^T)^{-1} g_u$$

In Theorem 5 it was shown that  $P(u^*)L_{uu}(u^*, \lambda^*) + 2cg_u^T(u^*)g_u(u^*)$  is nonsingular for  $c \neq 0$ , so that for  $u_{k+1}$  sufficiently close to  $u^*$ , we get

$$\delta u_{k+1} = - [P(u_{k+1})L_{uu}(u_{k+1}, \mu_{k+1}) + 2cg_u^T(u_{k+1})g_u(u_{k+1})]^{-1} 2cg_u^T(u_{k+1})g(u_{k+1})$$

For the linear-quadratic problem, this will yield the optimal solution  $u^* = u_{k+1} + \delta u_{k+1}$  in one step. Also notice, that in case  $g_u$  is nonsingular,  $\delta u_{k+1} = -g_u^{-1}(u_{k+1})g(u_{k+1})$ , that is,  $u^*$  is determined by the condition  $g(u^*) = 0$ .

Summarizing, we get the following algorithm:

Second order algorithm I:

- a) Select  $\mu_k = 0$ .
- b) Minimize  $F(u, \mu_k)$  with an ordinary function minimization algorithm. Assume that the minimum occurs for  $u = u_{k+1}$ .
- c) If  $\|g(u_{k+1})\| < \delta$ , where  $\delta$  is a small quantity, then  $u^* = u_{k+1}$ .
- d) Compute

$$\mu_{k+1} = \mu_k + 2cg(u_{k+1})$$

$$G = P(u_{k+1})L_{uu}(u_{k+1}, \mu_{k+1}) + 2cg_u^T(u_{k+1})g_u(u_{k+1})$$

and

$$\delta u_{k+1} = -2cG^{-1}g_u^T(u_{k+1})g(u_{k+1})$$

If  $G$  is singular, return to b) and minimize  $F(u, \mu_{k+1})$ .

- e) Estimate  $\mu_{k+2} = \mu(u_{k+1} + \delta u_{k+1})$  and return to b).

Notice that this algorithm depends heavily on the particular choice of  $\mu(u)$ . To allow for arbitrary multiplier functions, one possibility is to simply approximate  $\mu(u)$  by the series expansion

$$\mu(u) = \mu(u_k) + \mu_u(u_k)(u - u_k)$$

We then have

Second order algorithm II:

- a) Set  $\mu(u_k) = 0$  and  $\mu_u(u_k) = 0$ .
- b) Minimize  $f(u) + [\mu(u_k) + \mu_u(u_k)(u - u_k)]^T g(u) + 2cg^T(u)g(u)$  with an ordinary minimization algorithm. Assume that the minimum occurs for  $u = u_{k+1}$ .
- c) If  $\|g(u_{k+1})\| < \delta$ , then  $u^* = u_{k+1}$ .

d) Compute  $\mu(u_{k+1})$  and  $\mu_u(u_{k+1})$  and return to b).

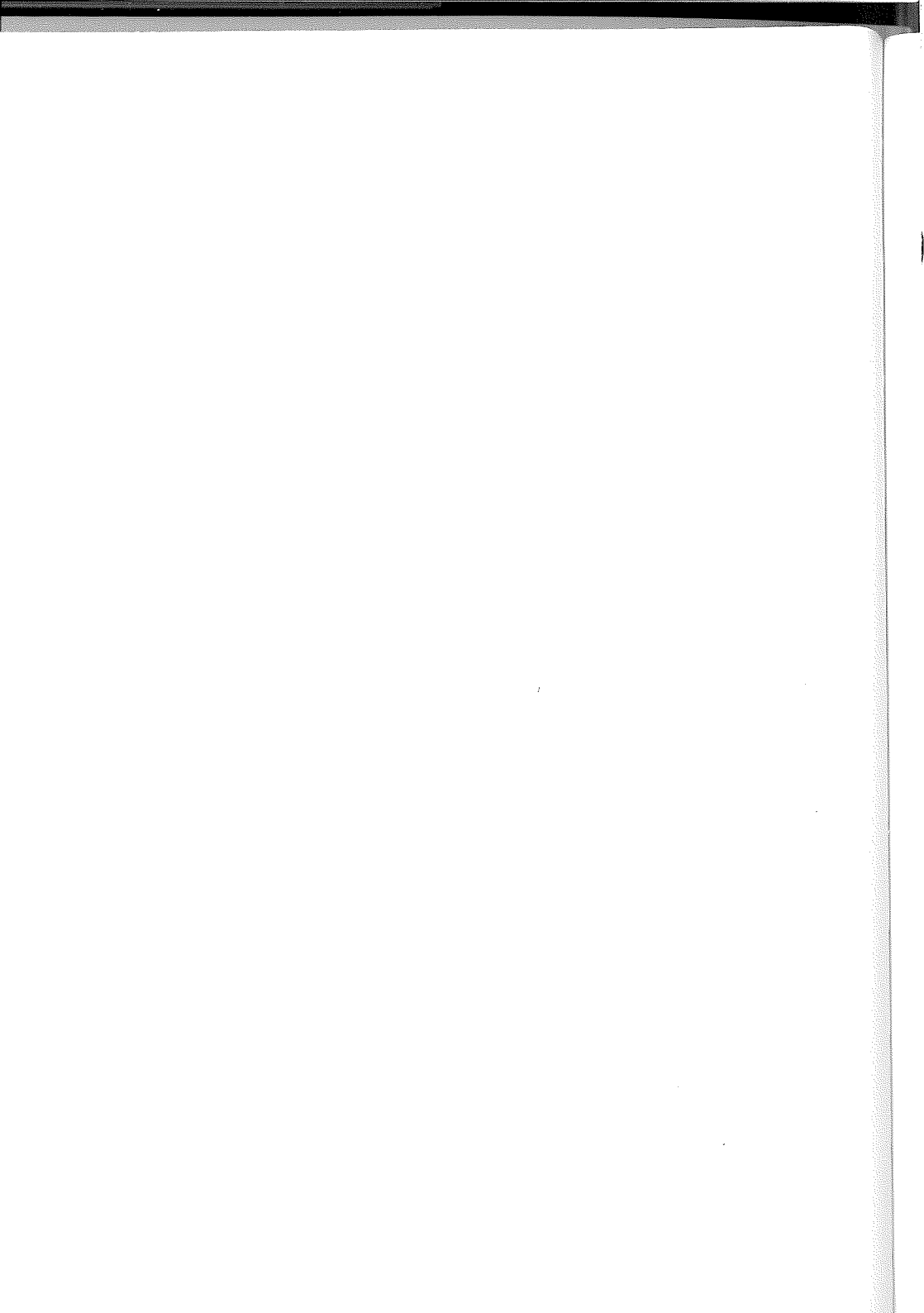
The function and gradient evaluations at stage b) are still very simple, since  $\mu(u_k)$  and  $\mu_u(u_k)$  are evaluated only at the minimizing point  $u_k$ . The convergence properties of the algorithm depend on the choice of  $\mu(u)$ . In particular, if  $\mu(u) = - (g_u g_u^T)^{-1} g_u^T f_u^T$  the algorithm has one-step convergence for linear-quadratic problems for  $c > c_0$ . In this case the approximation of  $\mu(u)$  is exact.

## 6. REFERENCES

- [1] A. V. Fiacco and G. P. McCormick: "Nonlinear Programming: Sequential Unconstrained Minimization Techniques", J. Wiley and Sons, New York, 1968.
- [2] M. J. D. Powell: "A Method for Nonlinear Constraints in Minimization Problems", in R. Fletcher, ed.: "Optimization", Academic Press, London, 1969.
- [3] M. R. Hestenes: "Multiplier and Gradient Methods", J. Opt. Theory Appl., Vol. 4, 1969, 303 - 320.
- [4] K. Mårtensson: "Methods for Constrained Function Minimization", Research Report 7101, March, 1971, Lund Institute of Technology, Division of Automatic Control.
- [5] R. Fletcher: "Methods for Nonlinear Programming", in J. Abadie, ed.: "Integer and Nonlinear Programming", North-Holland Publishing Company, 1970.
- [6] O. L. Mangasarian: "Nonlinear Programming", McGraw-Hill, 1968.
- [7] F. R. Gantmacher: "The Theory of Matrices", Chelsea Publishing Company, New York, 1960.
- [8] M. J. D. Powell: "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives", Computer Journal, Vol. 7, 1964, 155 - 162.
- [9] G. W. Stewart III: "A Modification of Davidon's Minimization Method to Accept Difference Approximations to Derivatives", J. Ass. Comput. Mach., Vol. 14, 1967, 72 - 83.



- [10] W.C. Davidon: "Variable Metric Method für Minimization",  
AEC Research and Development Report ANL 5990, 1959.
- [11] R. Fletcher and M.J.D. Powell: "A Rapidly Convergent Descent  
Method for Minimization", Computer Journal, Vol. 6,  
1963, 163 - 168.



A CONSTRAINING HYPERPLANE TECHNIQUE FOR  
STATE VARIABLE CONSTRAINED OPTIMAL CONTROL  
PROBLEMS

ABSTRACT

A new approach to the numerical solution of optimal control problems with state variable inequality constraints is presented. It is shown that simple approximations of admissible state variable regions by means of constraining hyperplanes transforms the problem into a mixed state-control variable constrained problem. A second order Differential Dynamic Programming algorithm is derived for the transformed problem. The efficiency and accuracy of the combination of constraining hyperplanes and the second order algorithm are investigated on problems of different complexity. Comparisons are made with the slack variable technique and with penalty function methods.

## 1. INTRODUCTION

Optimal control problems with state variable inequality constraints have for a long time been subject to a great interest. Research efforts have been concentrated on computational methods as well as conditions for optimality, and considerable progress has been made during the last years.

Necessary conditions for optimality have been investigated by several authors. Of particular interest are the conditions given by Bryson, Denham and Dreyfus [1], Speyer and Bryson [20], and Jacobson, Lele and Speyer [8]. These conditions are quite different. In [1] and [20] conditions at entry points, that is, the times when the constraint becomes active, are considered. This results in discontinuities of the adjoint variables at entry times, and thus an à priori unknown number of discontinuities (Lagrange multipliers) must be determined. In [8] the constraint is directly adjoined to the cost functional by means of a Lagrange multiplier function. It then turns out that the adjoint variables in general are discontinuous at both entry and exit times. However, the number of discontinuities is still à priori unknown. It has been shown by examples [8], that these conditions are somewhat stronger than the necessary conditions based on entry point considerations.

Similarly, different computational methods have been presented. A suitable way to classify these, is to relate them to either of three basic ideas; entry point conditions, penalty functions and slack variables.

Entry point conditions have been exploited by Denham and Bryson [4], and by Dreyfus [5]. Different modifications have also been reported, e.g. [22]. The basic idea of these methods is to iterate on the unknown adjoint variable discontinuities and on the unknown entry times. Thus it is necessary to à priori know the number of entry points, which for some problems is very difficult to predict (see e.g. [17]), and it may also be necessary to have good estimates of the different entry times. In general this approach is thus restricted to rather simple problems, e.g. problems with only one entry point and some knowledge about the entry time. However, for such problems very efficient algorithms may be constructed from the necessary entry point conditions.

Penalty function methods have been presented by Kelley [11] and Lasdon, Waren and Rice [13]. The main idea is to convert the original problem into an unconstrained problem by adding penalty terms to the cost functional. These methods can be considered as straightforward genera-

lizations of existing penalty function methods for finite dimensional problems. Thus the unconstrained problem is solved for successively increasing penalty weights, and it has been shown [9] that the solutions tend to satisfy the necessary conditions given by Jacobson et al., as the penalty tends to infinity. In penalty function methods it is thus not necessary to guess the number of entry points, and this is an obvious advantage compared with the methods mentioned above. However, it turns out that these methods are extremely sensitive to numerical errors, and it is difficult to reach solutions close to the optimal solution.

An alternative method has been proposed by Jacobson and Lele [9]. The method is based on a slack variable technique in function space, originally introduced by Valentine [21]. By introducing a sufficient number of additional state variables, the problem is converted into an unconstrained problem of higher dimension. Thus the number of entry points need not be a priori known. However, by introducing slack variables, the original problem is converted into a singular problem, and then new computational difficulties appear. In spite of this disadvantage, the slack variable technique has proved superior to penalty function methods [9].

In this paper a new approach is presented. The basic idea is to approximate the feasible region  $S(x;t) \leq 0$  with a region possible to express as an explicit function of the control variables, that is,  $g(x,u;t) \leq 0$ . Computational methods to handle the mixed state-control variable constrained problem may then be derived with e.g. a Differential Dynamic Programming technique [10]. It will be shown that a simple and natural approach

is to construct  $g(x,u;t)$  as hyperplanes in the  $(S, \frac{dS}{dt}, \dots, \frac{d^q S}{dt^q})$ -space,

where  $q$  is the order of  $S(x;t)$ . The half-space  $g(x,u;t) \leq 0$ , that is, the accuracy of the approximation is then determined by the choice of the hyperplane. The method can be considered as a penalty function method in the sense that the slope of the hyperplane tends to infinity as the half-space  $g(x,u;t) \leq 0$  tends to  $S(x;t) \leq 0$ . However, it will be shown that from a numerical point of view there are considerable advantages compared with ordinary penalty function methods.

In section 2 the problem is stated, and a brief survey of necessary conditions for optimality is given in section 3. The constraining hyperplane technique is presented in section 4, and it is shown that the approximating constraints always yield feasible solutions to the original problem. The necessary equations for a second order Differential Dynamic Programming algorithm are derived in section 5, and an algorithm is outlined. In section 6 the efficiency and the accuracy of the constraining

hyperplane technique are investigated on problems of different complexity. Comparisons are made with the penalty function methods and with the slack variable method, and it will be shown that the combination of constraining hyperplanes and a second order algorithm is superior to these methods. It will also become clear from the solved problems, that the constraining hyperplane technique may contribute to the understanding of the nature of state-variable constrained problems.

## 2. STATEMENT OF THE PROBLEM

Consider a dynamic system described by the following set of ordinary nonlinear differential equations

$$\frac{dx}{dt} = f(x, u; t) \quad x(t_0) = x_0 \quad (2.1)$$

where  $x(t)$  is the  $n$ -dimensional state vector and  $u(t)$  the  $m$ -dimensional control vector. We wish to determine a control history  $u(t)$ ,  $t \in [t_0, t_f]$ , such that the cost functional

$$J = F(x(t_f); t_f) + \int_{t_0}^{t_f} L(x, u; t) dt \quad (2.2)$$

is minimized. The terminal time  $t_f$  may be given either explicitly or implicitly. The minimization shall be carried out subject to the following constraints:

$$\begin{aligned} \psi(x(t_f); t_f) &= 0 \\ S(x; t) &\leq 0 \quad \forall t \in [t_0, t_f] \\ g(x, u; t) &\leq 0 \quad \forall t \in [t_0, t_f] \end{aligned} \quad (2.3)$$

The terminal constraint  $\psi$  is an  $s$ -dimensional ( $s \leq n$ ) nonlinear vector function of the state at time  $t_f$ , and it is assumed that the optimal solution satisfies  $\text{rank } \{\psi_x\} = s$ .

The state variable inequality constraint  $S$  is for simplicity assumed to

be a scalar, the generalization to the vector case being straightforward. We also assume that  $S$  is of order  $q$ , that is, the  $q$ -th total time derivative of  $S$  is the first to explicitly contain the control variable  $u$ .

The mixed state-control variable constraint  $g$  is a  $p$ -dimensional nonlinear vector function.  $g$  is an explicit function of the control variable  $u$ , but the explicit dependence on  $x$  is arbitrary.

The dimension  $p$  of  $g(x, u; t)$  is also arbitrary, but at any time  $t \in [t_0, t_f]$ , the number  $\hat{p}$  of active components of  $g$  must not exceed the number of control variables  $m$ . Moreover, if  $\hat{g}(x, u; t)$  denotes the  $\hat{p}$  active constraints at time  $t$ , it will be required that  $\text{rank} \{ \hat{g}_u \} = \hat{p} \leq m$ . If in addition the state variable inequality constraint  $S$  is active at time  $t$ , the rank of the composed  $(\hat{p} + 1) \times m$ -matrix

$$\begin{bmatrix} \hat{g}_u \\ \frac{\partial}{\partial u} \left\{ \frac{d^q S}{dt^q} \right\} \end{bmatrix} \quad (2.4)$$

must equal  $\hat{p} + 1$ , where  $\hat{p} + 1 \leq m$ .

### 3. SURVEY OF NECESSARY CONDITIONS FOR OPTIMALITY

Different necessary conditions for optimality can be derived for the problem stated in the preceding section. Particularly useful from a computational point of view are the conditions given by Bryson, Denham, Dreyfus [1], and the conditions recently given by Jacobson, Lele and Speyer [8]. In this section we will briefly summarize these two sets of necessary conditions.

#### 3.1. Necessary conditions according to Bryson, Denham and Dreyfus

Assume that  $t_1 \in (t_0, t_f)$  is an intermediate time when the state variable constraint  $S$  becomes active, and assume that  $t_2 \in (t_1, t_f)$  is the first time when  $S$  becomes inactive again.  $t_1$  and  $t_2$  are not a priori known.

The basic idea is now to transform the state variable constraint  $S$  in the interval  $[t_1, t_2]$  into a control variable constraint. Since  $S$  is identically zero on  $[t_1, t_2]$ , all its total derivatives with respect to  $t$  must be zero

and in particular  $\frac{d^q S}{dt^q}(x, u; t)$ , the lowest order derivative that contains  $u$  explicitly, must be zero. The interior point constraint  $N(x(t_1); t_1) = 0$ , where

$$N(x; t) = \begin{bmatrix} S(x; t) \\ \frac{dS}{dt}(x; t) \\ \vdots \\ \frac{d^{q-1} S}{dt^{q-1}}(x; t) \end{bmatrix} \quad (3.1)$$

is then imposed at the entry time, and the control variable constraint in  $[t_1, t_2]$  is given by

$$\frac{d^q S}{dt^q}(x, u; t) \leq 0 \quad (3.2)$$

The exit time  $t_2$  is then determined by (3.2) becoming inactive.

The transformed problem then is to minimize the cost functional

$$J = F(x(t_f); t_f) + \int_{t_0}^{t_f} L(x, u; t) dt$$

subject to the system equations

$$\frac{dx}{dt} = f(x, u; t) \quad x(t_0) = x_0$$

and the constraints

$$\psi(x(t_f); t_f) = 0$$

$$N(x(t_1); t_1) = 0$$



$$\frac{d^q S}{dt^q}(x, u; t) \leq 0 \quad \forall t \in [t_1, t_2]$$

$$g(x, u; t) \leq 0 \quad \forall t \in [t_0, t_f]$$

Following [1], we then adjoin the control variable constraints by means of Lagrange multipliers  $\mu(t)$  and  $\gamma(t)$  to the cost functional. The Hamiltonian  $H$  of the problem is then defined as

$$H = L(x, u; t) + \lambda^T(t)f(x, u; t) + \mu^T(t)g(x, u; t) + \gamma(t)\frac{d^q S}{dt^q}(x, u; t)$$

where

$$\mu_i(t) = \begin{cases} \geq 0 & \text{if } g_i(x, u; t) = 0 \\ 0 & \text{if } g_i(x, u; t) < 0 \end{cases} \quad (3.3)$$

and

$$\gamma(t) = \begin{cases} \geq 0 & \text{if } S(x; t) = 0 \\ 0 & \text{if } S(x; t) < 0 \end{cases}$$

Necessary conditions for optimality then are

$$\frac{\partial H}{\partial u} = 0 \quad t \in [t_0, t_f]$$

$$\frac{d\lambda^T}{dt} = -\frac{\partial H}{\partial x} \quad (3.4)$$

$$\lambda^T(t_f) = (\nu^T \psi_x + F_x^T)_{t=t_f}$$

where  $\nu$  is an  $s$ -dimensional vector of multipliers.  $\mu(t)$ ,  $\gamma(t)$  and  $\nu$  are determined by the conditions  $g(x, u; t) \leq 0$ ,  $\frac{d^q S}{dt^q}(x, u; t) \leq 0$  and

$\psi(x(t_f); t_f) = 0$ . If the terminal time  $t_f$  is given implicitly, the additional condition is

$$(F_t + \nu^T \psi_t + H)_t = t_f = 0 \quad (3.5)$$

At the entry point  $t_1$ , the adjoint variables  $\lambda(t)$  may be discontinuous, and

$$\lambda^T(t_1^-) = \lambda^T(t_1^+) + \pi^T \frac{\partial N}{\partial x}(x(t_1); t_1) \quad (3.6)$$

where  $\pi$  is a  $q$ -dimensional vector of multipliers determined by the condition  $N(x(t_1); t_1) = 0$ . Also the Hamiltonian may suffer discontinuities at  $t_1$  according to

$$H(t_1^-) = H(t_1^+) - \pi^T \frac{\partial N}{\partial t}(x(t_1); t_1)$$

It should be noticed that the interior point constraint (3.2) could equally well be imposed at the exit time  $t_2$ . The discontinuities in  $\lambda$  and  $H$  will then appear at  $t_2$  instead of  $t_1$ .

### 3.2. Necessary conditions according to Jacobson, Lele and Speyer<sup>+</sup>

An alternative approach, originally introduced by Chang [3], is to directly adjoin the state variable inequality constraint  $S \leq 0$  to the cost functional by means of a multiplier function  $\eta(t)$ , where

$$\eta(t) = \begin{cases} \geq 0 & \text{if } S = 0 \\ 0 & \text{if } S < 0 \end{cases}$$

Defining the Hamiltonian  $H$  as

$$H = L(x, u; t) + \lambda^T(t)f(x, u; t) + \mu^T(t)g(x, u; t) + \eta(t)S(x; t)$$

<sup>+</sup> These conditions will for brevity in the sequel be referred to as Speyer's necessary conditions, since they first (in a weaker version) were presented in [20]. We will also for brevity refer to the previous conditions as Bryson's conditions.

where  $\mu(t)$  is defined through (3.3), the necessary conditions are given by (3.4) and (3.5). However, the adjoint variables  $\lambda$  may now suffer discontinuities at both the entry and the exit points, and

$$\lambda^T(t_1^-) = \lambda^T(t_1^+) + \nu_1 S_x(x(t_1); t_1)$$

$$\lambda^T(t_2^-) = \lambda^T(t_2^+) + \nu_2 S_x(x(t_2); t_2)$$
(3.7)

The multipliers  $\nu_i \geq 0$  are determined by the conditions

$$S(x(t_i); t_i) = 0 \quad i = 1, 2$$
(3.8)

By inspection of (3.7) and (3.6), it is also clear that the directions of the discontinuities at the entry point are different.

These necessary conditions are of particular interest in the following sections. It will be illustrated by examples in section 6, that the computational method based on constraining hyperplanes converges to these necessary conditions, and not to the conditions given by Bryson et al.

#### 4. THE CONSTRAINING HYPERPLANE TECHNIQUE

The basic idea in the constraining hyperplane technique, is to approximate the admissible region  $S(x;t) \leq 0$  with an appropriate region in the

$(S, \frac{dS}{dt}, \dots, \frac{d^q S}{dt^q})$ -space. Thus the problem will be converted into a mixed state-control variable constrained problem.

We will first consider the approximation technique applied to first and second order constraints ( $q = 1$  and  $2$ ). Conditions for the application to constraints of general order are then established.

#### 4.1. First order constraints

Let  $S(x;t)$  be of first order, that is,  $\frac{dS}{dt} = \frac{dS}{dt}(x,u;t)$ . Approximate the feasible region  $S \leq 0$  in the  $(S, \frac{dS}{dt})$ -space with the half-space

$$\frac{dS}{dt} + a_1 S \leq 0 \quad (4.1)$$

generated by the straight line (hyperplane)

$$\pi_1 = \left\{ \left( S, \frac{dS}{dt} \right) \mid \frac{dS}{dt} + a_1 S = 0, \quad a_1 > 0 \right\} \quad (4.2)$$

This is illustrated in Fig. 1. Now suppose that the solution is computed

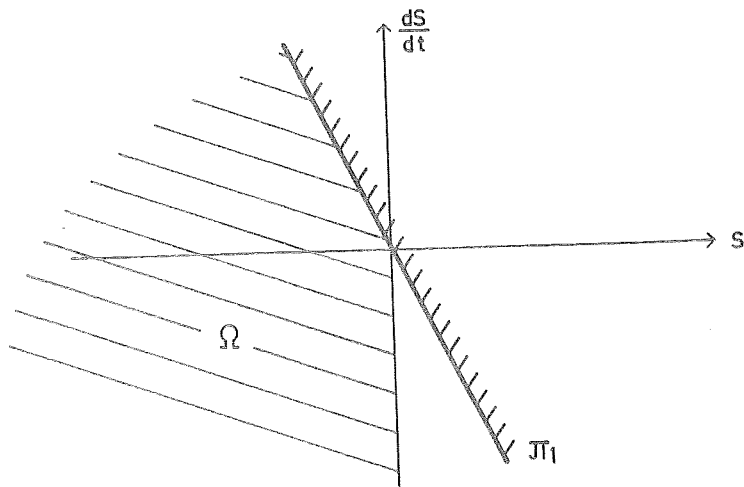


Fig. 1. - Constraining hyperplane  $\pi_1$  for first order constraints.

subject to the mixed state-control variable inequality constraint

$$\frac{dS}{dt}(x, u; t) + a_1 S(x; t) \leq 0$$

It is then easily verified that  $S(x(t); t) \leq 0, \forall t \geq t_0$ , if and only if

$S(x(t_0); t_0) \leq 0$ . Thus  $(S, \frac{dS}{dt}) \in \Omega, \forall t \geq t_0$ , if and only if  $(S, \frac{dS}{dt}) \in \Omega, t = t_0$ .

The slope of  $\pi_1$  obviously determines the accuracy of the approximation, and consequently it should be favourable to choose  $a_1$  as large as possible. In section 6 the computational aspects on the approximation accuracy are discussed and illustrated with solved problems.

#### 4.2. Second order constraints

The extension to second order constraints is straightforward.

Assuming that

$$\frac{dS}{dt} = \frac{dS}{dt}(x;t)$$

$$\frac{d^2S}{dt^2} = \frac{d^2S}{dt^2}(x, u;t)$$

we approximate the admissible region  $S \leq 0$  in the  $(S, \frac{dS}{dt}, \frac{d^2S}{dt^2})$ -space with the half-space

$$\frac{d^2S}{dt^2} + a_1 \frac{dS}{dt} + a_2 S \leq 0 \quad (4.3)$$

The half-space is generated by the plane (hyperplane)

$$\pi_2 = \left\{ \left( S, \frac{dS}{dt}, \frac{d^2S}{dt^2} \right) \mid \frac{d^2S}{dt^2} + a_1 \frac{dS}{dt} + a_2 S = 0 \right\} \quad (4.4)$$

Furthermore, it is postulated that the parameters  $a_i$  are real, and that the zeroes  $\xi_1, \xi_2$  of

$$p^2 + a_1 p + a_2 = 0$$

are real with  $\xi_1 < \xi_2 < 0$ . (These conditions may be somewhat relaxed, as will become clear from the following).

We will now prove that  $S \leq 0, \forall t \geq t_0$ , if (4.3) plus some simple conditions at  $t_0$  are satisfied. Then rewrite the differential inequality (4.3) as the inhomogeneous differential equation

$$\frac{d^2 S}{dt^2} + a_1 \frac{dS}{dt} + a_2 S + \epsilon^2(t) = 0 \quad (4.5)$$

introduce the state variables  $z_1 = \frac{dS}{dt}$ ,  $z_2 = S$ . Then (4.5) is equivalent to

$$\frac{dz}{dt} = \begin{bmatrix} -a_1 & -a_2 \\ 1 & 0 \end{bmatrix} z + \begin{bmatrix} -1 \\ 0 \end{bmatrix} \epsilon^2(t) = Az + B\epsilon^2(t) \quad (4.6)$$

to be able to establish the relevant properties of the system (4.6), we define new state variables  $w_1$  and  $w_2$  through the linear transformation  $w = Tz$ . Then

$$\frac{dw}{dt} = TAT^{-1}w + TB\epsilon^2(t) \quad (4.7)$$

choose  $T$  so that  $TAT^{-1}$  is diagonal. Then  $T^{-1}$  is given by the Vandermonde matrix

$$T^{-1} = \begin{bmatrix} \xi_1 & \xi_2 \\ 1 & 1 \end{bmatrix}$$

and (4.7) becomes

$$\frac{dw}{dt} = \begin{bmatrix} \xi_1 & 0 \\ 0 & \xi_2 \end{bmatrix} w + \begin{bmatrix} -1/(\xi_1 - \xi_2) \\ 1/(\xi_1 - \xi_2) \end{bmatrix} \epsilon^2(t) \quad (4.8)$$

From the explicit solution

$$w_1(t) = e^{\xi_1(t-t_0)} w_1(t_0) + \int_{t_0}^t e^{\xi_1(t-s)} [-1/(\xi_1 - \xi_2)] \epsilon^2(s) ds$$

$$w_2(t) = e^{\xi_2(t-t_0)} w_2(t_0) + \int_{t_0}^t e^{\xi_2(t-s)} [1/(\xi_1 - \xi_2)] \epsilon^2(s) ds$$

of (4.8) and from the assumption  $\xi_1 < \xi_2 < 0$ , it is obvious that

$$w_1(t_0) \geq 0 \Leftrightarrow w_1(t) \geq 0, \forall t \geq t_0 \quad (4.9)$$

$$w_2(t_0) \leq 0 \Leftrightarrow w_2(t) \leq 0, \forall t \geq t_0$$

Since  $w = Tz$ , (4.9) is equivalent to

$$(z_1(t_0) - \xi_2 z_2(t_0))/(\xi_1 - \xi_2) \geq 0 \Leftrightarrow (z_1(t) - \xi_2 z_2(t))/(\xi_1 - \xi_2) \geq 0, \forall t \geq t_0$$

$$(-z_1(t_0) + \xi_1 z_2(t_0))/(\xi_1 - \xi_2) \leq 0 \Leftrightarrow (-z_1(t) + \xi_1 z_2(t))/(\xi_1 - \xi_2) \leq 0, \forall t \geq t_0$$

or

$$\frac{dS}{dt} - \xi_i S \leq 0, t = t_0 \Leftrightarrow \frac{dS}{dt} - \xi_i S \leq 0, \forall t \geq t_0, \quad i = 1, 2 \quad (4.10)$$

We have thus proved that (4.10) holds if the inequality (4.3) is satisfied for all  $t \geq t_0$ . But if

$$S \leq 0, \quad t = t_0$$

and

$$\frac{dS}{dt} - \xi_i S \leq 0, \forall t \geq t_0$$

is satisfied for  $i = 1$  or  $2$ , then  $S \leq 0, \forall t \geq t_0$ , that is,  $S$  belongs to regions similar to  $\Omega$  for the first order constraint. It may then finally be concluded that

$$\frac{d^2 S}{dt^2}(x, u; t) + a_1 \frac{dS}{dt}(x; t) + a_2 S(x; t) \leq 0, \forall t \geq t_0$$

implies that

$$S \leq 0, \quad \forall t \geq t_0$$

if

$$S \leq 0, \quad t = t_0$$

and if

$$\frac{dS}{dt} - \xi_i S \leq 0, \quad t = t_0 \quad (4.11)$$

holds for either  $i = 1$  or  $i = 2$ . The solution thus always satisfies  $S(x;t) \leq 0$  if the constraining hyperplane  $\pi_2$  is never violated, and if the initial state of the system satisfies the inequalities (4.11). Notice that (4.11) is a very weak restriction, since it can be satisfied by choosing the parameters  $a_1$  and  $a_2$  such that  $\min(\xi_i)$  is negative enough. ( $\xi_i$  are in the sequel frequently referred to as the eigenvalues of the hyperplane).

### 4.3. Higher order constraints

The constraining hyperplane technique is easily generalized to state variable constraints of arbitrary order. Let  $S(x;t)$  be of order  $q$ , that is,

$$\frac{dS}{dt} = \frac{dS}{dt}(x;t)$$

.

.

.

$$\frac{d^q S}{dt^q} = \frac{d^q S}{dt^q}(x, u; t)$$

The feasible region  $S \leq 0$  is then approximated with the half-space

$$\frac{d^q S}{dt^q} + a_1 \frac{d^{q-1} S}{dt^{q-1}} + \dots + a_q S \leq 0 \quad (4.12)$$

which is generated by the constraining hyperplane

$$\pi_q = \left\{ \left( S, \frac{dS}{dt}, \dots, \frac{d^q S}{dt^q} \right) \mid \frac{d^q S}{dt^q} + a_1 \frac{d^{q-1} S}{dt^{q-1}} + \dots + a_q S = 0 \right\} \quad (4.13)$$

The parameters  $a_i$  are assumed real and positive, and it is assumed that the zeroes  $\xi_1, \xi_2, \dots, \xi_q$  of

$$p^q + a_1 p^{q-1} + \dots + a_q = 0$$



are real, distinct and satisfy  $\xi_1 < \xi_2 < \dots < \xi_q < 0$ .

We will now prove that (4.12) implies  $S \leq 0$ ,  $\forall t \geq t_0$ , provided that the initial state satisfies some relations similar to (4.11). The differential inequality (4.12) is then rewritten as the inhomogeneous differential equation

$$\frac{d^q S}{dt^q} + a_1 \frac{d^{q-1} S}{dt^{q-1}} + \dots + a_q S + \epsilon^2(t) = 0$$

and the state variables

$$z_i = \frac{d^{q-i} S}{dt^{q-i}}$$

are introduced. Then

$$\begin{aligned} \frac{dz}{dt} &= \begin{bmatrix} -a_1 & -a_2 & \dots & -a_q \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \end{bmatrix} z + \begin{bmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} \epsilon^2(t) = \\ &= Az + B\epsilon^2(t) \end{aligned} \tag{4.14}$$

The linear transformation  $w = Tz$ , where  $T^{-1}$  is given by the Vandermonde matrix

$$T^{-1} = \begin{bmatrix} \xi_1^{q-1} & \xi_2^{q-1} & \dots & \xi_q^{q-1} \\ \vdots & \vdots & & \vdots \\ \xi_1 & \xi_2 & & \xi_q \\ 1 & 1 & & 1 \end{bmatrix}$$

then transforms (4.14) into

$$\frac{dw}{dt} = \begin{bmatrix} \xi_1 & 0 \\ \cdot & \cdot \\ 0 & \xi_q \end{bmatrix} w + TB\epsilon^2(t) \quad (4.15)$$

Before proceeding with the analysis of the system (4.15), an explicit expression of  $T$  will be derived.

Introduce the polynomials

$$C^i(p) = \prod_{\substack{j=1 \\ j \neq i}}^q (p - \xi_j) = p^{q-1} + c_1^i p^{q-2} + \dots + c_{q-1}^i$$

with  $c_j^i > 0$ . With the assumptions made about  $\xi_i$ ,  $C^i(p)$  has the following properties.

$$\begin{aligned} C^i(\xi_j) &= 0 & i \neq j \\ C^i(\xi_i) &\neq 0 & (\xi_i \neq \xi_j) \\ C^q(\xi_q) &> 0 & (\xi_1 < \xi_2 < \dots < \xi_q) \end{aligned}$$

Further, if

$$C^i(\xi_i) > 0$$

then

$$\begin{aligned} C^{i+1}(\xi_{i+1}) &< 0 \\ C^{i-1}(\xi_{i-1}) &< 0 \end{aligned}$$

It is now easily verified that

$$T = \begin{bmatrix} 1/C^1(\xi_1) & & 0 \\ & \ddots & \\ 0 & & 1/C^q(\xi_q) \end{bmatrix} \quad \begin{bmatrix} 1 & c_1^1 & \dots & c_{q-1}^1 \\ \vdots & \vdots & & \vdots \\ 1 & c_1^q & \dots & c_{q-1}^q \end{bmatrix}$$

and

$$TB = \begin{bmatrix} -1/C^1(\xi_1) \\ \vdots \\ -1/C^q(\xi_q) \end{bmatrix}$$

The solution of (4.15) then is

$$w_i(t) = e^{\xi_i(t-t_0)} w_i(t_0) + \int_{t_0}^t e^{\xi_i(t-s)} [-1/C^i(\xi_i)] \epsilon^2(s) ds$$

and if we arbitrarily assume that  $C^i(\xi_i) > 0$ , it follows that  $w_i(t) \leq 0, \forall t \geq t_0$ , if and only if  $w_i(t_0) \leq 0$ . (If  $C^i(\xi_i) < 0$ , the inequalities are just reversed). Since  $w = Tz$ , the inequality

$$w_i(t) = [1/C^i(\xi_i)] (1, c_1^i, \dots, c_{q-1}^i) \begin{bmatrix} \frac{d^{q-1} S}{dt^{q-1}} \\ \vdots \\ S \end{bmatrix} \leq 0 \quad (4.16)$$

thus holds for all  $t \geq t_0$ , if and only if it holds for  $t = t_0$ . (4.16) may be further simplified to

$$\left\{ \prod_{\substack{j=1 \\ j \neq i}}^q \left( \frac{d}{dt} - \xi_j \right) \right\} S \leq 0 \quad \forall t \geq t_0, \quad i = 1, \dots, q \quad (4.17)$$

if and only if (4.17) holds for  $t = t_0$ , where  $(\frac{d}{dt} - \xi_j)S$  is interpreted as the operator  $(\frac{d}{dt} - \xi_j)$  acting on  $S$ . (If  $C^i(\xi_i) < 0$ , the inequalities

(4.16) are reversed, but (4.17) are unchanged).

The order of the differential inequalities (4.17) are  $q - 1$ , and to these the same procedure may be applied until the order of the inequalities is 1 or 2. The conditions derived for first and second order constraints are then applicable.

Sufficient conditions for the solution to satisfy the constraint  $S \leq 0, \forall t \geq t_0$ , can now be summarized as follows:

Assume that the inequality

$$\frac{d^q S}{dt^q}(x, u; t) + a_1 \frac{d^{q-1} S}{dt^{q-1}}(x; t) + \dots + a_q S(x; t) \leq 0$$

holds for all  $t \geq t_0$ , and let  $\xi_1 < \xi_2 < \dots < \xi_q < 0$  be the zeroes of the polynomial

$$p^q + a_1 p^{q-1} + \dots + a_q = 0$$

Then a sufficient condition for

$$S \leq 0 \quad \forall t \geq t_0$$

to hold, is the existence of a sequence of  $q - 1$  zeroes  $\xi_{i_1}, \dots, \xi_{i_{q-1}}$ , such that

$$S \leq 0$$

$$\left(\frac{d}{dt} - \xi_{i_1}\right) S \leq 0$$

$$\left(\frac{d}{dt} - \xi_{i_1}\right) \left(\frac{d}{dt} - \xi_{i_2}\right) S \leq 0$$

$$\vdots$$

$$\prod_{j=1}^{q-1} \left(\frac{d}{dt} - \xi_{i_j}\right) S \leq 0$$

(4.18)

hold for  $t = t_0$ .

The inequalities (4.18) generate half-spaces to which  $S, \dots, \frac{d^{q-1}S}{dt^{q-1}}$  must belong for  $t = t_0$ . If  $S(t_0) < 0$ , (4.18) are easily satisfied by choosing the parameters  $a_i$  such that the eigenvalues  $\xi_i$  are negative enough. All the half-spaces then tend to the half-space  $S \leq 0$ , while the corresponding hyperplanes become perpendicular to the  $S$ -axis.

## 5. A SECOND-ORDER DIFFERENTIAL DYNAMIC PROGRAMMING ALGORITHM FOR MIXED STATE-CONTROL VARIABLE CONSTRAINTS

$g(x, u; t) \leq 0$

As was shown in section 4, the constraining hyperplane technique transforms the original state variable inequality constrained problem into a problem with mixed constraints,  $g(x, u; t) \leq 0$ . The fact that  $g$  now depends explicitly on the state will have consequences for the numerical solution of the problem, since existing algorithms, see e.g. [2] [6] [10], deal only with pure control variable constraints  $g(u; t) \leq 0$ . Thus a suitable principle for the algorithm must be selected and then generalized.

Our choice is based on personal computational experiences. Methods based on Differential Dynamic Programming have in general proved to be superior to other existing methods, and we have thus chosen this approach. Further, the choice between first and second order methods is simple, since the additional work required for second order methods always pays back.

In this section we will thus develop a second order Differential Dynamic Programming algorithm for a particular class of continuous time control problems. This class consists of those problems for which the optimal control is a continuous function of time, i.e. bang-bang control problems are not covered. It will also be assumed that the problem has normality properties [2]. The algorithm is a generalization of the algorithm given by Jacobson and Mayne [10] for the case  $g(u; t) \leq 0$ , and with some exceptions, the same technique will be used to derive the algorithm for the general case  $g(x, u; t) \leq 0$ .

## 5.1. Differential Dynamic Programming

Let us briefly recapitulate the problem that results when the constraining hyperplane technique is applied. Given the dynamic system

$$\frac{dx}{dt} = f(x, u; t) \quad x(t_0) = x_0 \quad (5.1)$$

we want to determine the control strategy  $u(t)$ ,  $t_0 \leq t \leq t_f$ , that minimizes the cost functional

$$J = F(x(t_f); t_f) + \int_{t_0}^{t_f} L(x, u; t) dt \quad (5.2)$$

subject to the constraints

$$\begin{aligned} \psi(x(t_f); t_f) &= 0 \\ g(x, u; t) &\leq 0 \quad \forall t \in [t_0, t_f] \end{aligned} \quad (5.3)$$

where  $\psi$  and  $g$  are nonlinear vector functions of dimension  $s (\leq n)$  and  $p$ . Notice that the dimension of  $g$  is arbitrary. However, as will become clear later, it is necessary to put restrictions on the number  $\hat{p}$  of active constraints  $\hat{g}$ .

A possible way to handle the terminal constraints, is to adjoin  $\psi$  to the cost functional by means of Lagrange multipliers [2], and in the sequel we will thus consider the augmented cost functional

$$\bar{J} = F(x(t_f); t_f) + b^T \psi(x(t_f); t_f) + \int_{t_0}^{t_f} L(x, u; t) dt \quad (5.4)$$

It is assumed that the problem (5.1) - (5.3) has properties such that the minimal solution constitutes a minimal solution also of (5.4). For some problems this does not hold. However, by adding the quadratic form  $c\psi^T\psi$  to  $F$ , the extremal of (5.4) can in general be made a minimizing solution [16].

Now define  $V^\circ(x, b; t)$  as the minimal contribution to the cost over the time interval  $[t, t_f]$ , when the state of the system at time  $t$  is  $x(t)$ , i. e.

$$V^{\circ}(x, b; t) = \min_{u(\tau)} \left\{ F + b^T \psi + \int_t^{t_f} L ds \right\} \quad (5.5)$$

$$t \leq \tau \leq t_f$$

$$g(x, u; \tau) \leq 0$$

Assuming that  $V^{\circ}(x, b; t)$  exists and is twice continuously differentiable with respect to  $x$  and  $t$ ,  $\forall t \in [t_0, t_f]$ ,  $V^{\circ}(x, b; t)$  satisfies the well-known Hamilton-Jacobi-Bellman partial differential equation<sup>+</sup>

$$-\frac{\partial V^{\circ}}{\partial t}(x, b; t) = \min_u \left\{ L(x, u; t) + V_x^{\circ}(x, b; t) f(x, u; t) \right\} \quad (5.6)$$

$$g(x, u; t) \leq 0$$

If  $\bar{u}(t)$ ,  $\bar{x}(t)$  and  $\bar{b}$  is a nominal solution close to the optimal solution  $u^{*t}(t) = \bar{u}(t) + \delta u(t)$ ,  $x^{*t}(t) = \bar{x}(t) + \delta x(t)$  and  $b^{*t} = \bar{b} + \delta b$ , (5.6) can be written

$$-\frac{\partial V^{\circ}}{\partial t}(\bar{x} + \delta x, \bar{b} + \delta b; t) = \min_{\delta u} \left\{ L(\bar{x} + \delta x, \bar{u} + \delta u; t) + \right.$$

$$\left. g(\bar{x} + \delta x, \bar{u} + \delta u; t) \leq 0 \right.$$

$$\left. + V_x^{\circ}(\bar{x} + \delta x, \bar{b} + \delta b; t) f(\bar{x} + \delta x, \bar{u} + \delta u; t) \right\} \quad (5.7)$$

Equation (5.7) will in principle yield no further information about the unknown quantities  $\delta u$ ,  $\delta x$  and  $\delta b$ , since  $V^{\circ}(x, b; t)$  is not known. But assume that  $V^{\circ}(\bar{x}, \bar{b}; t)$  is known, and also the first and second order partial derivatives with respect to  $x$  and  $b$ , all evaluated at  $\bar{x}, \bar{b}; t$ . Also assume that  $V^{\circ}$  is sufficiently smooth to be expanded in a second order Taylor expansion. We can then approximate  $V^{\circ}(x, b; t)$  with

<sup>+</sup> Notice, that with the definition of  $V^{\circ}$  (5.5) and the assumption that  $V^{\circ}$  possesses continuous partial derivatives with respect to  $x$  and  $t$ ,  $V_x^{\circ}(t)$  can be identified with the adjoint variables  $\lambda(t)$  (sec. 3).

$$\begin{aligned}
 V^\circ(x, b; t) &= V^\circ(\bar{x} + \delta x, \bar{b} + \delta b; t) = V^\circ(\bar{x}, \bar{b}; t) + V_x^\circ \delta x + V_b^\circ \delta b + \\
 &+ \langle \delta x, V_{xb}^\circ \delta b \rangle + \frac{1}{2} \langle \delta x, V_{xx}^\circ \delta x \rangle + \frac{1}{2} \langle \delta b, V_{bb}^\circ \delta b \rangle \quad (5.8)
 \end{aligned}$$

for  $x(t)$  and  $b$  sufficiently close to  $\bar{x}(t)$  and  $\bar{b}$ . Similarly  $V_x^\circ(x, b; t)$  may be expanded to first order as<sup>+</sup>

$$V_x^\circ(x, b; t) = V_x^\circ(\bar{x} + \delta x, \bar{b} + \delta b; t) = V_x^\circ(\bar{x}, \bar{b}; t) + \delta x^T V_{xx}^\circ + \delta b^T V_{xb}^\circ \quad (5.9)$$

All quantities are evaluated at  $\bar{x}, \bar{b}; t$ . We also introduce

$$a^\circ(\bar{x}, \bar{b}; t) = V^\circ(\bar{x}, \bar{b}; t) - \bar{V}(\bar{x}, \bar{b}; t) \quad (5.10)$$

that is, the difference between the optimal cost with initial state  $\bar{x}$  at time  $t$ , and the nominal cost produced by the nominal control  $\bar{u}(t)$  from the same initial condition. To simplify the notations, the superscript indicating the optimal solution will from now on be dropped.

Substituting (5.8) - (5.10) into (5.7), we then have

$$\begin{aligned}
 -\frac{\partial \bar{V}}{\partial t} - \frac{\partial a}{\partial t} - \frac{\partial V}{\partial t} x \delta x - \frac{\partial V}{\partial t} b \delta b - \langle \delta x, \frac{\partial V}{\partial t} xb \delta b \rangle - \frac{1}{2} \langle \delta x, \frac{\partial V}{\partial t} xx \delta x \rangle - \\
 - \frac{1}{2} \langle \delta b, \frac{\partial V}{\partial t} bb \delta b \rangle = \min_{\delta u} \left\{ L(\bar{x} + \delta x, \bar{u} + \delta u; t) + \right. \\
 \left. g(\bar{x} + \delta x, \bar{u} + \delta u; t) \leq 0 \right\} \\
 + \langle V_x^T + V_{xx} \delta x + V_{xb} \delta b, f(\bar{x} + \delta x, \bar{u} + \delta u; t) \rangle \quad (5.11)
 \end{aligned}$$

<sup>+</sup> That higher order terms may be neglected in these expansions is justified in [10] where also a detailed error analysis is given. Our notations will differ from those in [10]. We will thus consider  $V_x$  as an element of the dual space of  $R^n$  and not as the gradient. This means that transposes will appear throughout the derivation when compared with [10].  $\langle \cdot, \cdot \rangle$  is used as the scalar product between two elements of  $R^n$ .



Equation (5.11) is the fundamental relation in Differential Dynamic Programming, and from this the algorithm will be derived. Notice that  $V$  and the partial derivatives are evaluated at  $\bar{x}, \bar{b}; t$ . Since  $V$  is approximated with a second order expansion, we then have the following relations between the total and the partial time derivatives

$$\frac{d}{dt} (\bar{V} + a) = \frac{\partial}{\partial t} (\bar{V} + a) + V_x f(\bar{x}, \bar{u}; t)$$

$$\frac{dV_x}{dt} = \frac{\partial V_x}{\partial t} + f^T(\bar{x}, \bar{u}; t) V_{xx}$$

$$\frac{dV_{xx}}{dt} = \frac{\partial V_{xx}}{\partial t}$$

(5.12)

$$\frac{dV_b}{dt} = \frac{\partial V_b}{\partial t} + f^T(\bar{x}, \bar{u}; t) V_{bx}^T$$

$$\frac{dV_{bb}}{dt} = \frac{\partial V_{bb}}{\partial t}$$

$$\frac{dV_{xb}}{dt} = \frac{\partial V_{xb}}{\partial t}$$

## 5.2. Derivation and outline of the algorithm

The derivation of the algorithm is based on the fundamental equations (5.11) and (5.12), where the former is used twice. First the optimal solution is characterized in terms of neighbouring optimal solutions under the assumption that these are known. This step constitutes a major part of the derivation. When the optimal variation  $\delta u$  is determined (in terms of  $\delta x$  and  $\delta b$ ), the left hand side of (5.11) will then be identified with the right hand side of (5.11) where the optimal variation  $\delta u$  is substituted. This will result in a set of partial differential equations for the optimal return  $V(\bar{x}, \bar{b}; t)$  and its derivatives. The relations (5.12) will then yield the total time derivatives, and boundary conditions for

these can be determined from (5.5). These differential equations will thus characterize  $V(x, b; t)$  to second order about  $\bar{x}, \bar{b}; t$ , if either the nominal control  $\bar{u}(t)$  (resulting in the nominal state  $\bar{x}(t)$ ) is sufficiently close to the optimal solution  $u^*(t)$ , or if the problem is Linear-Quadratic and no control variable constraints are active.

These differential equations and the computed expression for the optimal variation  $\delta u$ , constitute the fundamental relations on which the algorithm will be based.

In principle, it will then be apparent how the unknown quantity  $\delta x$  should be determined. We could just apply the new control  $\bar{u} + \delta u$  to the dynamic system (5.1) since  $\delta u$  has been expressed in terms of  $\delta x$  and  $\delta b$ . However, it should be emphasized that we are in general dealing with highly non-Linear-Quadratic problems, and the second order approximations of the optimal cost may be very bad. Then there is no guarantee that the new nominal solution  $\bar{x} + \delta x$ ,  $\bar{u} + \delta u$ ,  $\bar{b} + \delta b$  is a better approximation of the optimal solution than the previous one.

A very elegant and powerful way to keep this problem under control, the "step-size adjustment technique", was proposed in [10], and is included in the algorithm.

For similar reasons, the computation of the corrections  $\delta b$  may be critical for the efficiency of the algorithm. In the outline of the algorithm, this problem is treated. The computational technique proposed is a generalization of the method proposed by Gershwin and Jacobson [7], and has proved in practice to be a substantial improvement.

We then introduce the Hamiltonian

$$H(x, u, V_x; t) = L(x, u; t) + V_x f(x, u; t)$$

and consider the right hand side of (5.11) with  $\delta x$  and  $\delta b$  equal to zero.

$$\min_{\delta u} \left\{ H(\bar{x}, \bar{u} + \delta u, V_x; t) \right\} \quad (5.13)$$

$$g(\bar{x}, \bar{u} + \delta u; t) \leq 0$$

Determine the variation  $\delta u^*$  that minimizes  $H$  subject to the constraint  $g(\bar{x}, \bar{u} + \delta u; t) \leq 0$ , and assume that the minimum is

$$H(\bar{x}, u^*, V_x; t)$$

where  $u^* = \bar{u} + \delta u^*$ . Thus  $u^*$  would be the optimal solution if the corresponding trajectory was  $\bar{x}$  and the corresponding multipliers  $\bar{b}$ .

Consequently,  $u^*$  must be corrected by an amount  $\delta u$  that takes into account that  $\bar{x}$  and  $\bar{b}$  differ from the optimal solution with  $\delta x$  and  $\delta b$  respectively. We then reintroduce these variations, and also the  $\delta u$  required to maintain optimality.

$$\min_{\delta u} \left\{ H(\bar{x} + \delta x, u^* + \delta u, V_x(\bar{x} + \delta x, \bar{b} + \delta b; t); t) \right\} \quad (5.14)$$

$$g(\bar{x} + \delta x, u^* + \delta u; t) \leq 0$$

Depending on whether the constraints  $g(\bar{x}, u^*; t)$  are active or not, we must now separate between two cases, A and B.

Case A. Let us first consider the case where some of the constraints  $g$  are active at  $\bar{x}, u^*$ . (The unconstrained case B will then follow as a simple special case). Denote by  $\hat{u}(=u^*)$  the minimizing control, and assume that  $\hat{p}$  of the constraints  $g$  are active ( $\hat{p} \leq m$ ). Let  $\hat{g}$  stand for these constraints, that is,  $\hat{g}(\bar{x}, \hat{u}; t) = 0$ . We assume that  $\hat{g}$  satisfies the constraint qualification (compare (2.4))

$$\text{rank } \hat{g}_u(\bar{x}, \hat{u}; t) = \hat{p} \quad (5.15)$$

We could now handle the active constraints  $\hat{g}$  in the same way as the terminal constraints  $\psi$ , i. e. by introducing Lagrange multipliers  $\lambda$  and a corresponding Lagrangian

$$\mathcal{L}(\bar{x}, u, \lambda, V_x; t) = H(\bar{x}, u, V_x; t) + \langle \lambda, \hat{g}(\bar{x}, u; t) \rangle$$

Necessary conditions for a local minimum in  $u = \hat{u}$  then are

$$\mathcal{L}_u(\bar{x}, \hat{u}, \lambda, V_x; t) = H_u(\bar{x}, \hat{u}, V_x; t) + \lambda^T \hat{g}_u(\bar{x}, \hat{u}; t) = 0$$

$$\hat{g}(\bar{x}, \hat{u}; t) = 0$$

where the multipliers  $\lambda \geq 0$  exist and are unique if the constraint qualification (5.15) is satisfied. This approach is used in [10] for the pure control variable constraints  $g(u; t) \leq 0$ .

An alternative possibility is to use the multiplier function technique [16]. Since this approach is very well suited for neighbouring or disturbed solutions, the minimization problems (5.13) and (5.14) will be approached with this technique. Restricting ourselves to nonsingular problems, the

$\hat{p}$ -dimensional vector function  $\mu = \mu(x, u, b; t)$  will be called a multiplier function if and only if

- i)  $\mu$  exists and is twice differentiable with respect to  $u$  in a neighbourhood of  $\bar{x}, \hat{u}, \bar{b}; t$ .
- ii)  $\mu(\bar{x}, \hat{u}, \bar{b}; t) = \lambda$

Necessary conditions for optimality then are

$$H_u(\bar{x}, \hat{u}, V_x; t) + \mu^T(\bar{x}, \hat{u}, \bar{b}; t) \hat{g}_u(\bar{x}, \hat{u}; t) = 0$$

$$\hat{g}(\bar{x}, \hat{u}; t) = 0$$
(5.16)

The choice of  $\mu$  is now free within the frame of the definition. In particular

$$\mu(x, u, b; t) = - \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u H_u^T$$

is easily shown to be an admissible multiplier function, and this will be used in the following.

Now reintroduce the variations  $\delta x$ ,  $\delta b$  and  $\delta u$  according to (5.14), and assume that the constraints  $\hat{g}$  are still active (since the variations are small). Analogous to (5.16) the following conditions then are necessary for minimum, and from these conditions the optimal correction  $\delta u$  will be determined.

$$H_u(\bar{x} + \delta x, \hat{u} + \delta u, V_x(\bar{x} + \delta x, \bar{b} + \delta b; t); t) +$$

$$+ \mu^T(\bar{x} + \delta x, \hat{u} + \delta u, \bar{b} + \delta b; t) \hat{g}_u(\bar{x} + \delta x, \hat{u} + \delta u; t) = 0$$

$$\hat{g}(\bar{x} + \delta x, \hat{u} + \delta u; t) = 0$$
(5.17)

To determine  $\delta u$  as a function of  $\delta x$  and  $\delta b$ , we expand (5.17) to first order about  $\bar{x}, \hat{u}, \bar{b}; t$ . Then

$$\begin{aligned}
& H_u^T + H_{uu} \delta u + H_{ux} \delta x + f_u^T (V_{xx} \delta x + V_{xb} \delta b) + \\
& + \hat{g}_u^T (\mu_u + \mu_x \delta x + \mu_u \delta u + \mu_b \delta b) + \mu \hat{g}_{uu} \delta u + \mu \hat{g}_{ux} \delta x = 0
\end{aligned} \tag{5.18}$$

$$\hat{g}_u + \hat{g}_u \delta u + \hat{g}_x \delta x = 0$$

All quantities are evaluated at  $\bar{x}, \hat{\alpha}, \bar{b}; t$ . The notation  $\mu \hat{g}_{uu}$  symbolizes

$$\Sigma \mu_i \frac{\partial^2 \hat{g}_i}{\partial u^2}$$

and similarly for  $\mu \hat{g}_{ux}$ .  $\mu_u$ ,  $\mu_x$  and  $\mu_b$  are determined through a straightforward differentiation, i. e. ,

$$\begin{aligned}
\mu_u &= - \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u (H_{uu} + \mu \hat{g}_{uu}) \\
\mu_x &= - \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u (H_{ux} + f_u^T V_{xx} + \mu \hat{g}_{ux}) \\
\mu_b &= - \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u f_u^T V_{xb}
\end{aligned} \tag{5.19}$$

Substituting (5.19) into (5.18), noticing that  $H_u + \mu \hat{g}_u^T = 0$  and  $\hat{g} = 0$  according to (5.16), the necessary conditions (5.18) reduce to

$$\begin{aligned}
& Y \left[ (H_{uu} + \mu \hat{g}_{uu}) \delta u + (H_{ux} + f_u^T V_{xx} + \mu \hat{g}_{ux}) \delta x + f_u^T V_{xb} \delta b \right] = 0 \\
& \hat{g}_u \delta u + \hat{g}_x \delta x = 0
\end{aligned} \tag{5.20}$$

where

$$Y = I_m - \hat{g}_u^T \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u$$

is an orthogonal projection.  $I_m$  stands for the unit matrix of dimension

$m$  (= the number of control variables). Thus

$$(H_{uu} + \mu \hat{g}_{uu}) \delta u + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b$$

belongs to the range space of the linear transformation  $\hat{g}_u^T$ , and then

$$(H_{uu} + \mu \hat{g}_{uu}) \delta u + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b = \hat{g}_u^T \alpha$$

with the  $\hat{p}$ -dimensional vector  $\alpha$  determined by the condition  $\hat{g}_u \delta u + \hat{g}_x \delta x = 0$ . To solve for  $\alpha$ , we assume that  $H_{uu} + \mu \hat{g}_{uu}$  is non-singular,<sup>+</sup> in which case

$$\delta u = - (H_{uu} + \mu \hat{g}_{uu})^{-1} \left[ (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b - \hat{g}_u^T \alpha \right] \quad (5.21)$$

Substitute (5.21) into  $\hat{g}_u \delta u + \hat{g}_x \delta x = 0$ . Then

$$\begin{aligned} & - \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x - \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb} \delta b + \\ & + \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \alpha + \hat{g}_x \delta x = 0 \end{aligned}$$

and

$$\begin{aligned} \alpha = & \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + \right. \\ & \left. + \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb} \delta b - \hat{g}_x \delta x \right] \quad (5.22) \end{aligned}$$

The inverse  $\left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1}$  exists since  $\hat{g}_u$  is assumed to have full rank  $\hat{p} \leq m$ . With  $\alpha$  determined by (5.22) the optimal variation  $\delta u$  (5.21) finally becomes

<sup>+</sup> Although this assumption is very natural for nonsingular optimal control problems, it may probably be relaxed. See e.g. [16].

$$\begin{aligned}
\delta u = & - (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x - (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb} \delta b + \\
& + (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + \\
& + (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb} \delta b - \\
& - (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_x \delta x
\end{aligned}$$

or

$$\begin{aligned}
\delta u = & - (H_{uu} + \mu \hat{g}_{uu})^{-1} \left\{ I_m - \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \right\} \cdot \\
& \cdot \left\{ (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b \right\} - \\
& - (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_x \delta x \tag{5.23}
\end{aligned}$$

To simplify the notations, we introduce the  $m \times \hat{p}$ -matrix

$$Q = (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1}$$

and the  $m \times m$ -matrix

$$Z = I_m - Q \hat{g}_u = I_m - (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u$$

where  $I_m$  stands for the unit matrix of dimension  $m \times m$ . Since  $(H_{uu} + \mu \hat{g}_{uu})^{-1} Z^T$  is symmetric, the variation  $\delta u$  necessary to main-

tain optimality (5.23), can then be summarized as the linear feedback

$$\delta u = \beta_1 \delta x + \beta_2 \delta b \quad (5.24)$$

where

$$\beta_1 = - Z(H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) - Q \hat{g}_x \quad (5.25)$$

$$\beta_2 = - Z(H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb}$$

Except for the  $Q \hat{g}_x$ -term, these equations are identical to the equations derived in [10] for pure control variable constraints  $g(u;t) \leq 0$  (and of course yield the equations in [10] as a special case).

To proceed with the derivation, the optimal admissible variation (5.24) shall be inserted into the basic relation (5.11). However, at this stage it is convenient to first establish some properties of  $Z$ ,  $Q$ ,  $\beta_1$  and  $\beta_2$ , since this will simplify the subsequent derivation.

1.  $Z$  and  $Z^T$  are projections ( $Z^2 = Z$ ), and  $Z^T$  projects the range of  $\hat{g}_u^T$  on zero. However, the projections are not orthogonal (in the metric used).

$$2. \quad Z^T H_u^T = o_m \quad (5.26)$$

where  $o_m$  is the null element in  $R^m$ . (5.26) follows from the fact that  $H_u^T = -\hat{g}_u^T \mu$  belongs to the range of  $\hat{g}_u^T$ , and from property 1.

$$3. \quad Z(H_{uu} + \mu \hat{g}_{uu})^{-1} Z^T = (H_{uu} + \mu \hat{g}_{uu})^{-1} Z^T = Z(H_{uu} + \mu \hat{g}_{uu})^{-1} \quad (5.27)$$

The symmetry of  $Z(H_{uu} + \mu \hat{g}_{uu})^{-1}$ , that is, the second equality in (5.27), is easily verified. Then  $Z(H_{uu} + \mu \hat{g}_{uu})^{-1} Z^T = Z^2 (H_{uu} + \mu \hat{g}_{uu})^{-1} = Z(H_{uu} + \mu \hat{g}_{uu})^{-1}$ , where the last equality follows from the projection property of  $Z$ .



$$4. \quad ZQ = O_{m \times \hat{p}}$$

A straightforward multiplication yields

$$ZQ = Q - (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} = Q - Q = O_{m \times \hat{p}}$$

where  $O_{m \times \hat{p}}$  is the  $m \times \hat{p}$ -dimensional null matrix.

$$5. \quad \hat{g}_u^T Q = I_{\hat{p}}$$

This follows from the definition of  $Q$ .

$$6. \quad \hat{g}_u^T \beta_1 = - \hat{g}_x^T \quad (5.28)$$

From the definition of  $\beta_1$  (5.25), we get

$$\begin{aligned} \hat{g}_u^T \beta_1 &= - \hat{g}_u^T Z (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) - \hat{g}_u^T Q \hat{g}_x^T = \\ &= - \hat{g}_u^T Q \hat{g}_x^T = - \hat{g}_x^T \end{aligned}$$

since  $Z^T \hat{g}_u^T = O_{m \times \hat{p}}$  according to property 1, and  $\hat{g}_u^T Q = I_{\hat{p}}$  according to 5.

$$7. \quad \hat{g}_u^T \beta_2 = O_{\hat{p} \times s} \quad (5.29)$$

From the definition of  $\beta_2$  (5.25) follows

$$\hat{g}_u^T \beta_2 = - \hat{g}_u^T Z (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb}$$

and since  $\hat{g}_u^T Z = O_{\hat{p} \times m}$  according to 4, (5.29) is proved.

$$8. \quad \beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_1 = - \beta_2^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) = - V_{xb}^T f_u \beta_1 \quad (5.30)$$

The second equality is trivial. To prove the first, we notice that

$$\beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_1 = -\beta_2^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) - \beta_2^T (H_{uu} + \mu \hat{g}_{uu}) Q \hat{g}_x$$

and

$$\beta_2^T (H_{uu} + \mu \hat{g}_{uu}) Q = \beta_2^T \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} = O_{m \times p}$$

since  $\hat{g}_u \beta_2 = O_{\hat{p} \times s}$  according to (5.29)

(We will now drop the subscripts on the null elements, since the dimensions in the following will be obvious).

For future availability we will also derive an expression for the variation

$$\delta\mu = \mu(\bar{x} + \delta x, \hat{u} + \delta u, \bar{b} + \delta b; t) - \mu(\bar{x}, \hat{u}, \bar{b}; t)$$

in terms of  $\delta x$  and  $\delta b$ . Approximating  $\delta\mu$  with a first order series expansion about  $\bar{x}, \hat{u}, \bar{b}; t$ , and substituting  $\mu_u, \mu_x$  and  $\mu_b$  according to (5.19), we get

$$\begin{aligned} \delta\mu = & - \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u \left[ (H_{uu} + \mu \hat{g}_{uu}) \delta u + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + \right. \\ & \left. + f_u^T V_{xb} \delta b \right] \end{aligned}$$

or

$$\begin{aligned} \delta\mu = & \left[ \hat{g}_u \hat{g}_u^T \right]^{-1} \hat{g}_u \left[ (Z^T - I_m) (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + \right. \\ & \left. + (H_{uu} + \mu \hat{g}_{uu}) Q \hat{g}_x \delta x + (Z^T - I_m) f_u^T V_{xb} \delta b \right] \end{aligned}$$

where  $\delta u$  has been replaced by the local linear feedback (5.24).

Since, by definition,  $Z^T - I_m = -\hat{g}_u^T Q^T$ ,  $\delta\mu$  finally reduces to

$$\delta\mu = -Q^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) \delta x + \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_x \delta x - Q^T f_u^T V_{xb} \delta b \quad (5.31)$$

Now return to the fundamental equation (5.11), and expand the right hand side to second order under the assumption that the constraints are active, that is, expand

$$\min_{\delta u} \left\{ H(\bar{x} + \delta x, \hat{u} + \delta u, V_x(\bar{x} + \delta x, \bar{b} + \delta b; t); t) + \mu^T (\bar{x} + \delta x, \hat{u} + \delta u, \bar{b} + \delta b; t) g(\bar{x} + \delta x, \hat{u} + \delta u; t) \right\}$$

to second order about  $\bar{x}, \hat{u}, \bar{b}$ . Then we have

$$\begin{aligned} \min_{\delta u} \left\{ H + H_x \delta x + H_u \delta u + \langle \delta u, H_{ux} \delta x \rangle + \frac{1}{2} \langle \delta x, H_{xx} \delta x \rangle + \right. \\ + \frac{1}{2} \langle \delta u, H_{uu} \delta u \rangle + \langle \delta x, V_{xx} f \rangle + \langle \delta b, V_{xb}^T f \rangle + \\ + \langle \delta x, V_{xx} f \delta x \rangle + \langle \delta b, V_{xb}^T f \delta x \rangle + \langle \delta x, V_{xx} f \delta u \rangle + \\ + \langle \delta b, V_{xb}^T f \delta u \rangle + \langle \mu, \hat{g} \rangle + \langle \mu, \hat{g}_x \delta x \rangle + \\ + \langle \mu, \hat{g}_u \delta u \rangle + \langle \delta \mu, \hat{g} \rangle + \langle \delta \mu, \hat{g}_x \delta x \rangle + \langle \delta \mu, \hat{g}_u \delta u \rangle + \\ \left. + \frac{1}{2} \langle \delta u, \mu \hat{g}_{uu} \delta u \rangle + \frac{1}{2} \langle \delta x, \mu \hat{g}_{xx} \delta x \rangle + \langle \delta u, \mu \hat{g}_{ux} \delta x \rangle + \right. \\ \left. + \text{higher order terms} \right\} \end{aligned}$$

We now insert the minimizing variation  $\delta u = \beta_1 \delta x + \beta_2 \delta b$  and the variation  $\delta \mu$  given by (5.31). After some simplifying calculations, where we also make use of the relations  $\hat{g} = 0$  and  $H_u + \mu^T \hat{g}_u = 0$ , the following expression is obtained.

$$\begin{aligned}
& H + (H_x + \mu \hat{g}_x^T + f_{xx}^T V) \delta x + f_{xb}^T V \delta b + \\
& + \langle \delta x, \left\{ (H_{ux} + \mu \hat{g}_{ux}^T + f_{ux}^T V_{xx})^T \beta_2 + \beta_1^T (H_{uu} + \mu \hat{g}_{uu}^T) \beta_2 + \right. \\
& + (f_x + f_u \beta_1)^T V_{xb} - (\hat{g}_x + \hat{g}_u \beta_1)^T Q^T f_u^T V_{xb} - \\
& - (H_{ux} + \mu \hat{g}_{ux}^T + f_{ux}^T V_{xx})^T Q \hat{g}_u \beta_2 + \hat{g}_x^T \left[ \hat{g}_u^T (H_{uu} + \mu \hat{g}_{uu}^T)^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u \beta_2 \left. \right\} \delta b \rangle + \\
& + \langle \delta b, \left\{ \frac{1}{2} \beta_2^T (H_{uu} + \mu \hat{g}_{uu}^T) \beta_2 + V_{xb}^T f_u \beta_2 - V_{xb}^T f_u Q \hat{g}_u \beta_2 \right\} \delta b \rangle + \\
& + \langle \delta x, \left\{ \frac{1}{2} (H_{xx} + \mu \hat{g}_{xx}^T) + V_{xx} f_x + \frac{1}{2} \beta_1^T (H_{uu} + \mu \hat{g}_{uu}^T) \beta_1 + \right. \\
& + \beta_1^T (H_{ux} + \mu \hat{g}_{ux}^T + f_{ux}^T V_{xx}) - (H_{ux} + \mu \hat{g}_{ux}^T + f_{ux}^T V_{xx})^T Q (\hat{g}_x + \hat{g}_u \beta_1) + \\
& + \hat{g}_x^T \left[ \hat{g}_u^T (H_{uu} + \mu \hat{g}_{uu}^T)^{-1} \hat{g}_u^T \right]^{-1} (\hat{g}_x + \hat{g}_u \beta_1) \left. \right\} \delta x \rangle \quad (5.32)
\end{aligned}$$

All quantities are evaluated at  $\bar{x}, \bar{u}, \bar{b}; t$ . The series expansion (5.32) is now identified with the left hand side of equation (5.11). Identifying terms of equal power in  $\delta x$  and  $\delta b$ , and making use of the equations (5.12), a set of ordinary differential equations for the optimal cost and its partial derivatives are obtained. It should be kept in mind that these are evaluated at  $\bar{x}, \bar{b}; t$ .

1. Identifying terms of power zero in  $\delta x$  and  $\delta b$ , we get

$$-\frac{\partial \bar{V}}{\partial t} - \frac{\partial a}{\partial t} = H$$

Equation (5.12) then implies

$$-\frac{da}{dt} = H + \frac{d\bar{V}}{dt} - V_x f(\bar{x}, \bar{u}; t) = H - L(\bar{x}, \bar{u}; t) - V_x f(\bar{x}, \bar{u}; t)$$

where the last equality follows from the relation

$$\frac{d\bar{V}}{dt} = -L(\bar{x}, \bar{u}; t)$$

Then

$$- \frac{da}{dt} = H(\bar{x}, \hat{u}, V_x; t) - H(\bar{x}, \bar{u}, V_x; t) \quad (5.33)$$

2. Terms of power one in  $\delta x$  and zero in  $\delta b$ .

$$- \frac{\partial V}{\partial t} = H_x + \mu^T \hat{g}_x + f^T V_{xx}$$

(5.12) then implies

$$- \frac{dV}{dt} = H_x + \mu^T \hat{g}_x + (f - f(\bar{x}, \bar{u}; t))^T V_{xx} \quad (5.34)$$

where all quantities are evaluated at  $\bar{x}, \hat{u}, \bar{b}; t$  unless otherwise indicated.

3. Terms of power zero in  $\delta x$  and one in  $\delta b$ .

$$- \frac{\partial V}{\partial t} b = f^T V_{xb}$$

Then

$$- \frac{dV}{dt} b = (f - f(\bar{x}, \bar{u}; t))^T V_{xb} \quad (5.35)$$

4.1 Mixed terms.

$$\begin{aligned} - \frac{\partial V}{\partial t} x b &= (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T \beta_2 + \beta_1^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2 + \\ &+ (f_x + f_u \beta_1)^T V_{xb} - (\hat{g}_x + \hat{g}_u \beta_1)^T Q^T f_u^T V_{xb} - \\ &- (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T Q \hat{g}_u \beta_2 + \hat{g}_x^T [\hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T]^{-1} \hat{g}_u \beta_2 \end{aligned}$$

According to (5.29),  $\hat{g}_u \beta_2 = 0$ , and then the last two terms equal zero. Further,

$$(\hat{g}_x + \hat{g}_u \beta_1)^T Q^T f_u^T V_{xb} = 0$$

since  $\hat{g}_x + \hat{g}_u \beta_1 = 0$  (5.28). Finally, from (5.30), follows

$$(H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T \beta_2 = -\beta_1^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2$$

and then

$$-\frac{\partial V_{xb}}{\partial t} = (f_x + f_u^T \beta_1)^T V_{xb}$$

which, together with (5.12), implies

$$-\frac{dV_{xb}}{dt} = (f_x + f_u^T \beta_1)^T V_{xb} \quad (5.36)$$

5. Terms quadratic in  $\delta b$ .

$$\begin{aligned} -\frac{\partial V_{bb}}{\partial t} &= \beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2 + V_{xb}^T f_u \beta_2 + \beta_2^T f_u^T V_{xb} - \\ &\quad - V_{xb}^T f_u Q \hat{g}_u \beta_2 - \beta_2^T \hat{g}_u^T Q^T f_u^T V_{xb} \end{aligned}$$

Again, the last two terms will vanish since  $\hat{g}_u \beta_2 = 0$  (5.29). It is also easily verified by straightforward calculations, plus (5.27) and the projection property of  $Z$ , that

$$\beta_2^T f_u^T V_{xb} = V_{xb}^T f_u \beta_2 = -\beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2$$

Then (5.12) implies

$$-\frac{dV_{bb}}{dt} = -\beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2 \quad (5.37)$$

6. Terms quadratic in  $\delta x$ .

$$\begin{aligned}
 -\frac{\partial V_{xx}}{\partial t} &= H_{xx} + \mu \hat{g}_{xx} + f_x^T V_{xx} + V_{xx} f_x + \beta_1^T (H_{uu} + \mu \hat{g}_{uu}) \beta_1 + \\
 &+ \beta_1^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T \beta_1 - \\
 &- (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T Q (\hat{g}_x + \hat{g}_u \beta_1) - (\hat{g}_x + \hat{g}_u \beta_1)^T Q^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) + \\
 &+ \hat{g}_x^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} (\hat{g}_x + \hat{g}_u \beta_1) + \\
 &+ (\hat{g}_x + \hat{g}_u \beta_1)^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_x
 \end{aligned}$$

The last four terms will vanish since  $(\hat{g}_x + \hat{g}_u \beta_1) = 0$  (5.28). Then

$$\begin{aligned}
 -\frac{dV_{xx}}{dt} &= H_{xx} + \mu \hat{g}_{xx} + f_x^T V_{xx} + V_{xx} f_x + \beta_1^T (H_{uu} + \mu \hat{g}_{uu}) \beta_1 + \\
 &+ \beta_1^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T \beta_1
 \end{aligned} \tag{5.38}$$

Summarizing the equations (5.33) - (5.38), we thus get the following differential equations for the optimal cost and its partial derivatives:

$$\begin{aligned}
 -\frac{da}{dt} &= H - H(\bar{x}, \bar{u}, V_x; t) \\
 -\frac{dV_x}{dt} &= H_x + \mu \hat{g}_x^T + (f - f(\bar{x}, \bar{u}; t))^T V_{xx} \\
 -\frac{dV_b}{dt} &= (f - f(\bar{x}, \bar{u}; t))^T V_{xb} \\
 -\frac{dV_{xb}}{dt} &= (f_x + f_u \beta_1)^T V_{xb}
 \end{aligned} \tag{5.39}$$

$$-\frac{dV_{bb}}{dt} = -\beta_2^T (H_{uu} + \mu \hat{g}_{uu}) \beta_2$$

$$-\frac{dV_{xx}}{dt} = H_{xx} + \mu \hat{g}_{xx} + f_x^T V_{xx} + V_{xx} f_x + \beta_1^T (H_{uu} + \mu \hat{g}_{uu}) \beta_1 + \\ + \beta_1^T (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) + (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx})^T \beta_1$$

where

$$\beta_1 = -Z (H_{uu} + \mu \hat{g}_{uu})^{-1} (H_{ux} + \mu \hat{g}_{ux} + f_u^T V_{xx}) - Q \hat{g}_x \quad (5.40)$$

$$\beta_2 = -Z (H_{uu} + \mu \hat{g}_{uu})^{-1} f_u^T V_{xb}$$

and

$$Z = I_m - (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1} \hat{g}_u \quad (5.41)$$

$$Q = (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \left[ \hat{g}_u (H_{uu} + \mu \hat{g}_{uu})^{-1} \hat{g}_u^T \right]^{-1}$$

All quantities in (5.39) - (5.41) are evaluated at  $\bar{x}, \hat{u}, \bar{b}; t$ , unless otherwise indicated.

The boundary conditions of (5.39) are identical to the boundary conditions given in [10] for the case where no control variable constraints are present, that is,

$$a(\bar{x}(t_f), \bar{b}; t_f) = 0$$

$$V_x(\bar{x}(t_f), \bar{b}; t_f) = F_x(\bar{x}(t_f); t_f) + \bar{b}^T \psi_x(\bar{x}(t_f); t_f)$$

$$V_b(\bar{x}(t_f), \bar{b}; t_f) = \psi^T(\bar{x}(t_f); t_f) \quad (5.42)$$

$$V_{xb}(\bar{x}(t_f), \bar{b}; t_f) = \psi_x^T(\bar{x}(t_f); t_f)$$

$$V_{bb}(\bar{x}(t_f), \bar{b}; t_f) = 0$$

$$V_{xx}(\bar{x}(t_f), \bar{b}; t_f) = F_{xx}(\bar{x}(t_f); t_f) + \bar{b} \psi_{xx}(\bar{x}(t_f); t_f)$$



Case B. When no constraints are active at  $\hat{u}$ , the differential equations corresponding to (5.39) are easily obtained from these equations if we put

$$\mu = 0$$

$$Z = I_m$$

$$Q = 0$$

The boundary conditions are the same.

The influence of  $Z$  and  $Q$  can be given a simple interpretation. Let us first assume that  $\hat{g}_x = 0$ . The correction  $\delta u = \beta_1 \delta x + \beta_2 \delta b$  then becomes

$$\delta u = -Z(H_{uu} + \mu g_{uu})^{-1} \left[ (H_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b \right]$$

and thus

$$g_u \delta u = 0$$

since  $g_u Z = 0$ . The new control  $\hat{u} + \delta u$  then satisfies  $g(\hat{u} + \delta u) = 0$  to first order as was required (5.20). An alternative way to express this, is to say that  $Z$  projects the correction  $\delta u$  which should be made if no constraints were present, into the tangent plane of  $g$ . However,  $Z$  is not an orthogonal projection, since  $\delta u$  is the minimizing correction in the tangent plane, and thus the curvature of  $H$  must be taken into account.

Then consider the general case  $\hat{g}_x \neq 0$ . From (5.24) and (5.25) we have

$$\delta u = -Z(H_{uu} + \mu g_{uu})^{-1} \left[ (H_{ux} + \mu g_{ux} + f_u^T V_{xx}) \delta x + f_u^T V_{xb} \delta b \right] - Q g_x \delta x$$

$\delta u$  now consists of a projection of the unconstrained correction on the tangent plane, plus a term where the change of the tangent plane due to  $\delta x$  is taken into account.

Having characterized the optimal cost in terms of a neighbouring optimal solution, it is now obvious how an algorithm for numerical solution should be constructed. With some smaller, but from a numerical point of view essential modifications, the algorithm we outline here is a generalization of the algorithm given by Jacobson and Mayne in [10].

### Algorithm:

1. Guess a nominal control  $\bar{u}(t)$ ,  $t_0 \leq t \leq t_f$ , and compute the corresponding nominal trajectory  $\bar{x}(t)$ . Store  $\bar{u}(t)$  and  $\bar{x}(t)$ . Guess a nominal set of multipliers  $\bar{b}$ , and compute the corresponding cost  $\bar{V}(x_0, \bar{b}; t_0)$ .
2. Compute boundary conditions for  $a$ ,  $V_x$  and  $V_{xx}$  from (5.42).
3. Integrate  $\frac{da}{dt}$ ,  $\frac{dV_x}{dt}$  and  $\frac{dV_{xx}}{dt}$  backwards from  $t_f$  to  $t_0$  while minimizing  $H(\bar{x}, u, V_x; t)$  subject to  $g(\bar{x}, u; t) \leq 0$ . If no constraints are active, then  $Z = I_m$  and  $Q = 0$ . If  $\hat{p}$  constraints  $\hat{g}$  are active, then  $Z$  and  $Q$  are given by (5.41). Compute  $\beta_1$  according to (5.40), and store the minimizing control  $\hat{u}(t)$ ,  $\beta_1$  and  $a(\bar{x}, \bar{b}; t)$  for all  $t$ ,  $t_0 \leq t \leq t_f$ . If possible, store also  $Z$  and  $(H_{uu} + \mu \hat{g}_{uu})^{-1}$ , since this may later save a lot of computations.
4. If  $|a(x_0, \bar{b}; t_0)| < \eta_1$  where  $\eta_1$  is a small quantity, the predicted possible change in cost is small, and thus  $\bar{u}(t), \bar{x}(t)$  is considered as the optimal solution of the cost functional

$$\bar{J} = F + \bar{b}^T \psi + \int_{t_0}^{t_f} L ds \quad (5.43)$$

Otherwise proceed to 5.

If in addition  $\|\psi(\bar{x}(t_f); t_f)\| < \eta_2$ , where  $\eta_2$  is a small quantity, then  $\bar{u}, \bar{x}, \bar{b}$  is considered as the optimal solution of the original problem.

If  $|a| < \eta_1$  but  $\|\psi\| > \eta_2$ , go to 6.

5. Apply the new control

$$u = \hat{u} + \beta_1 \delta x = \hat{u} + \beta_1 (x - \bar{x}) \quad (5.44)$$

to the system, initially over the whole time interval  $[t_0, t_f]$ . When integrating the system equations, the constraints  $g(x, u; t)$  are checked

so that they are not violated. Even for rather small variations this may happen in the neighbourhood of entry and exit points. The new control is then, if possible, determined by the condition  $\underline{g} = 0$ .

If the reduction in cost  $\bar{V} - V$  is greater than zero, and with sufficient accuracy agrees with the predicted reduction in cost  $|a(x_0, \bar{b}; t_0)|$ , e.g.

$$\frac{\bar{V} - V}{|a|} > c \quad (5.45)$$

then choose  $u(t)$  and  $x(t)$  as the new improved nominal solution and go to 2. A suitable choice of  $c$  has proved to be 0.5, but for many problems the convergence may be greatly improved by other choices of  $c$ . If (5.45) does not hold, we apply the "step-size adjustment technique", that is, the new control (5.44) is applied only over a smaller part  $[t_s, t_f]$  of the whole interval  $[t_0, t_f]$ . A detailed description of this is found in [10].

6. When the cost functional (5.43) is minimized, that is  $|a(x_0, \bar{b}; t_0)| < \eta_1$ , the multipliers must be adjusted by an amount  $\delta b$ , so that the minimum of

$$F + (\bar{b} + \delta b)^T \psi + \int_{t_0}^{t_f} L ds$$

corresponds to  $\psi = 0$ . Then compute boundary values for  $V_b$ ,  $V_{xb}$  and  $V_{bb}$  according to (5.42).

7. With  $\hat{u}(t)$ ,  $\bar{x}(t)$ ,  $\beta_1$ ,  $Z$ , and  $(H_{uu} + \mu \underline{g}_{uu})^{-1}$  available in a storage

area,  $\frac{dV_b}{dt}$ ,  $\frac{dV_{xb}}{dt}$  and  $\frac{dV_{bb}}{dt}$  are integrated from  $t_f$  to  $t_0$ . For all

times  $t$ , we then compute and store  $\beta_2$ . Although  $\frac{dV_b}{dt}$  should equal

zero (5.39) when  $\bar{u}(t) (= \hat{u}(t))$ ,  $\bar{x}(t)$  minimizes the cost (5.43), we have found it necessary and highly improving upon accuracy to integrate this equation also. The reason for this is that although  $f(\bar{x}, \bar{u}; t) - f(\bar{x}, \hat{u}; t)$

should be rather small, the product  $(f(\bar{x}, \bar{u}; t) - f(\bar{x}, \hat{u}; t))^T V_{xb}$  may be

large, since the magnitude of  $V_{xb}$  is in no way restricted.

8. To compute the corrections  $\delta b$ , we notice, that for the optimal solution

$$V_b(x_0, b^*; t_0) = \psi^T(x_f^*(t_f); t_f) = 0$$

Then expand  $V_b$  to first order about  $\bar{b}$ .

$$V_b(x_0, b; t_0) = V_b(x_0, \bar{b}; t_0) + \delta b^T V_{bb}(x_0, \bar{b}; t_0) \quad (5.46)$$

To be able to solve for  $\delta b$ , we assume that  $V_{bb}(x_0, \bar{b}; t_0)$  is nonsingular, which is equivalent to claim normality of the problem [2]. The correction  $\delta b$  is thus given by

$$\delta b = -V_{bb}^{-1}(x_0, \bar{b}; t_0) V_b^T(x_0, \bar{b}; t_0)$$

Since (5.46) is an exact representation of  $V_b(x_0, b; t_0)$  only for Linear-Quadratic problems with linear terminal constraints  $\psi$ , it is necessary to be able to modify  $\delta b$  when it turns out that (5.46) is a bad approximation. A straightforward way is to choose

$$\delta b = -\epsilon V_{bb}^{-1}(x_0, \bar{b}; t_0) V_b^T(x_0, \bar{b}; t_0) \quad 0 < \epsilon \leq 1 \quad (5.47)$$

with  $\epsilon$  initially equal to one. To get a rule for the modification of  $\epsilon$ , we also predict the change in the optimal cost  $V(x_0, b; t_0)$  when the multipliers  $\bar{b}$  are changed the amount  $\delta b$ .

To predict the change in  $V$ , we use the second order expansion of  $V$  about  $\bar{b}$ , that is,

$$V(x_0, \bar{b} + \delta b; t_0) - V(x_0, \bar{b}; t_0) = V_b \delta b + \frac{1}{2} \langle \delta b, V_{bb} \delta b \rangle$$

Substituting  $\delta b$  (5.47), the predicted change in the optimal cost then is

$$V(x_0, \bar{b} + \delta b; t_0) - V(x_0, \bar{b}; t_0) = -(\epsilon - \frac{1}{2}\epsilon^2) V_b V_{bb}^{-1} V_b^T \quad (5.48)$$

Notice that (5.48) always is positive, since  $V_{bb}$  is negative definite and  $0 < \epsilon \leq 1$ . However, (5.48) may be a bad prediction from a numerical point of view, especially in the neighbourhood of the optimal multipliers  $b^*$ . The reason is that we do not compute the optimal cost  $V(x_0, \bar{b}; t_0)$  of the functional (5.43), but the approximation  $\bar{V}(x_0, \bar{b}; t_0)$ , where

$$\bar{V}(x_0, \bar{b}; t_0) = V(x_0, \bar{b}; t_0) - a(x_0, \bar{b}; t_0) \quad (5.49)$$

Since the cost functional (5.43) is considered to be minimized when the predicted improvement  $|a(x_0, \bar{b}; t_0)|$  is less than  $\eta_1$ , the difference  $\bar{V} - V$  may be as large as  $\eta_1$  when we try to change the multipliers  $\bar{b}$ . Close to  $b^*$ , when the corrections  $\delta b$  are small, this error may then be dominating over the predicted change (5.48). The relevant change in cost to predict is thus  $V(x_0, \bar{b} + \delta b; t) - \bar{V}(x_0, \bar{b}; t_0)$ . Substituting (5.49) into (5.48), we then have

$$V(x_0, \bar{b} + \delta b; t_0) - \bar{V}(x_0, \bar{b}; t_0) = a(x_0, \bar{b}; t_0) - \left(\epsilon - \frac{1}{2}\epsilon^2\right) V_b \bar{V}_{bb}^{-1} V_b^T \quad (5.50)$$

Also notice, that except from improving the accuracy, (5.50) makes the algorithm more flexible since an accurate minimization of (5.43) is not critical. It may then be possible to allow for a greater value of the parameter  $\eta_1$ , and this can sometimes improve the convergence rate of the algorithm. (For Linear-Quadratic problems with linear terminal constraints and no active control variable constraints, the choice of  $\eta_1$  is now completely arbitrary. The convergence will be one-step, and the predicted change (5.50) is identical to the actual change for any value of  $\eta_1$ ).

## 9. Apply the new control

$$u = \alpha + \beta_1 \delta x + \beta_2 \delta b$$

to the system over the whole time interval  $[t_0, t_f]$ . Compute the corresponding state  $x(t)$  and loss  $\tilde{V}(x_0, \bar{b} + \delta b; t_0)$ . Store  $x(t)$  and  $u(t)$ . We now apply two separate tests to judge if  $\delta b$  should be accepted as an improvement of the multipliers  $\bar{b}$  or not. The first is to examine if the magnitude of  $\psi$  is decreased, that is, if the condition

$$\| \psi(\bar{x}(t_f); t_f) \| - \| \psi(x(t_f); t_f) \| > 0$$

holds. If not, we proceed directly to 10. The second test is a comparison of the actual change in cost

$$\tilde{V}(x_0, \bar{b} + \delta b; t_0) - \bar{V}(x_0, \bar{b}; t_0)$$

with the predicted change (5.50). If the condition

$$\gamma_1 < \frac{\bar{V}(x_0, \bar{b} + \delta b; t_0) - \bar{V}(x_0, \bar{b}; t_0)}{V(x_0, \bar{b} + \delta b; t_0) - \bar{V}(x_0, \bar{b}; t_0)} < \gamma_2 \quad (5.51)$$

where  $\gamma_1$  and  $\gamma_2$  are suitably chosen (e.g.  $0 < \gamma_1 < 1$ ,  $\gamma_2 > 1$ ), is satisfied,  $\delta b$  is accepted as an improvement, and  $u(t), x(t)$  is a new improved nominal solution. The new nominal multipliers are  $\bar{b} + \delta b$  and the nominal cost is  $\bar{V}$ . Then return to 2. If (5.51) is not satisfied, proceed to 10.

10. If any of the tests for acceptance of  $\delta b$  is violated, set  $\epsilon = \epsilon/2$  and return to 8. If no correction has been approved after a certain number of reductions of  $\epsilon$  (e.g. 10), set  $\epsilon = 1$  and return to 8. Condition (5.51) is then released, and the only demand on  $\delta b$  is that  $\|\psi\|$  should be reduced. In practice, this will often get the algorithm out of the difficult situation.

This completes the outline of the algorithm. However, to get an efficient computer program, problems such as storage exploitation, integration routines, interpolation and extrapolation methods should also be considered. These problems are treated in [18].

## 6. EXAMPLES

In this section different properties of the constraining hyperplane technique will be illustrated. To get an idea about the accuracy and efficiency of the method, three problems with explicit solutions available are solved. It will be shown that the solutions of the transformed problems tend to satisfy the necessary conditions given by Speyer et al. (cf. sec. 3) as the slopes of the hyperplanes increase. The hyperplane technique will also be compared with Kelley's penalty function method [11] in respect to sensitivity and computational accuracy.

In a fourth problem, the efficiency of the combination of constraining hyperplanes and a second order Differential Dynamic Programming algorithm is compared with the slack variable technique proposed by Jacobson and Lele [9].

It should be pointed out, that all the computations have been executed in single precision on a UNIVAC 1108.

### Example 6.1

We will consider the system

$$\begin{aligned} \frac{dx_1}{dt} &= x_2 & x_1(0) &= 0 \\ \frac{dx_2}{dt} &= u & x_2(0) &= 1 \end{aligned} \tag{6.1}$$

with the cost functional

$$J = \frac{1}{2} \int_0^1 u^2 dt \tag{6.2}$$

the terminal constraints

$$\begin{aligned} \psi_1 &= x_1(1) = 0 \\ \psi_2 &= x_2(1) + 1 = 0 \end{aligned} \tag{6.3}$$

and the second order state variable inequality constraint

$$S(\mathbf{x}; t) = x_1(t) - \ell \leq 0 \quad \ell = 1/9 \tag{6.4}$$

In [2] it is shown that if  $\ell \geq 1/4$ , the constraint (6.4) is not active, if  $1/6 \leq \ell \leq 1/4$ , there is a tangency point, and if  $\ell < 1/6$ , the constraint is active over a time interval of length greater than zero.

For  $\ell = 1/9$  the analytic solution of the problem is [2]:

$$\begin{aligned}
 x_1(t) &= \begin{cases} [1 - (1 - 3t)^3]/9 & 0 \leq t \leq 1/3 \\ 1/9 & 1/3 \leq t \leq 2/3 \\ [1 - (3t - 2)^3]/9 & 2/3 \leq t \leq 1 \end{cases} \\
 x_2(t) &= \begin{cases} (1 - 3t)^2 & 0 \leq t \leq 1/3 \\ 0 & 1/3 \leq t \leq 2/3 \\ -(3t - 2)^2 & 2/3 \leq t \leq 1 \end{cases} \quad (6.5) \\
 u(t) &= \begin{cases} 18t - 6 & 0 \leq t \leq 1/3 \\ 0 & 1/3 \leq t \leq 2/3 \\ 12 - 18t & 2/3 \leq t \leq 1 \end{cases}
 \end{aligned}$$

Thus the entry time is  $t_1 = 1/3$ , and the exit time is  $t_2 = 2/3$ . The corresponding cost is  $J = 4.0$ .

Since  $S$  is of second order, the constraining hyperplane is

$$\frac{d^2 S}{dt^2} + a_1 \frac{dS}{dt} + a_2 S = 0$$

Substituting  $S = x_1 - \ell$ ,  $\frac{dS}{dt} = \dot{x}_1$ ,  $\frac{d^2 S}{dt^2} = \ddot{x}_1 = u$ , we then have

$$u + a_1 \dot{x}_1 + a_2 (x_1 - \ell) = 0$$

The transformed problem thus is to minimize the cost functional (6.2) subject to the terminal constraints (6.3) and the mixed state-control variable constraint



$$g(x, u; t) = u + a_1 x_2 + a_2 (x_1 - l) \leq 0$$

The problem was solved for three different hyperplanes A, B and C. The parameters  $a_i$  of these hyperplanes and the corresponding eigenvalues  $\xi_i$  are given in Table I.

Table I. Constraining hyperplanes

	$a_1$	$a_2$	$\xi_1$	$\xi_2$
A	45	500	- 20	- 25
B	105	2750	- 50	- 55
C	165	6800	- 80	- 85

In Figs 2 and 3 the optimal solution (6.5) is compared with the computed solutions for the different hyperplanes. Notice, that the solutions corresponding to B and C are almost identical to the optimal solution, and the maximum deviations in  $x_1$  are 0.00016 and 0.00003 respectively. This accuracy was obtained with  $\eta_1 = \eta_2 = 0.002$ , that is, the latest nominal solution was considered optimal when both the predicted improvement in cost  $a(x_0, b; t_0)$  and  $\|\psi\|$  (Euklidian norm) were less than 0.002.

Some interesting qualities of the constraining hyperplane technique can be observed in Figs 2 and 3. Firstly, the entry or contact time  $t_1$  is reached too early, and around  $t_1$  the solution is sensitive to the slope of the hyperplane. The first phenomenon is naturally explained by the construction of the hyperplane, which must intervene before the constraint  $S$  is reached. An investigation of the sensitivity is presented below. Secondly, the exit time  $t_2$  is found with very good accuracy, and the computed solutions agree very well with the optimal solution in the interval  $[t_2, t_f]$ . This is due to the separability of the problem [2] and to the construction of the hyperplane, which never prevents  $S(t)$  to move from  $S(t) = 0$  into the halfspace  $S(t) \leq 0$ .

The total number of iterations (including iterations on the multipliers b) and the computed costs for different hyperplanes are shown in Table II.

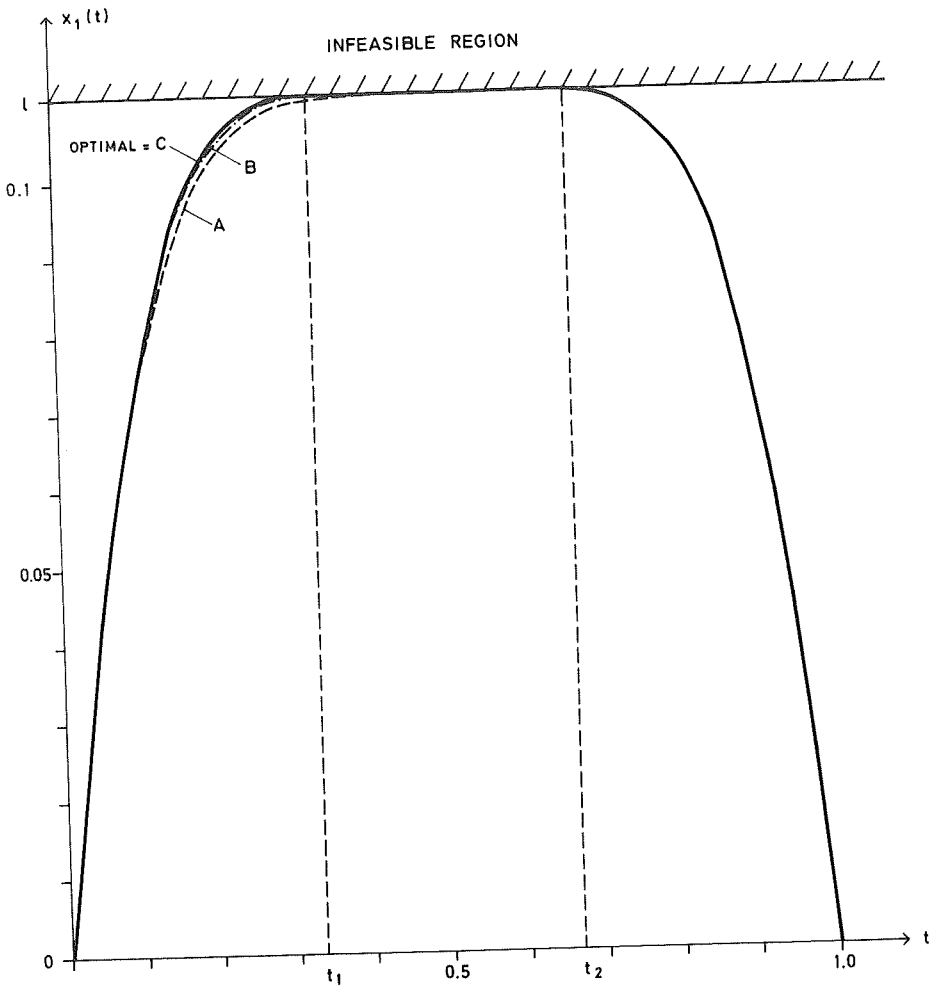


Fig 2. - Example 6.1: Computed solutions  $x_1(t)$  for the hyperplanes A(---), B(- . -) and C compared with the optimal solution (—). (Notice that the difference between the computed solution for hyperplane B and the optimal solution is exaggerated.)

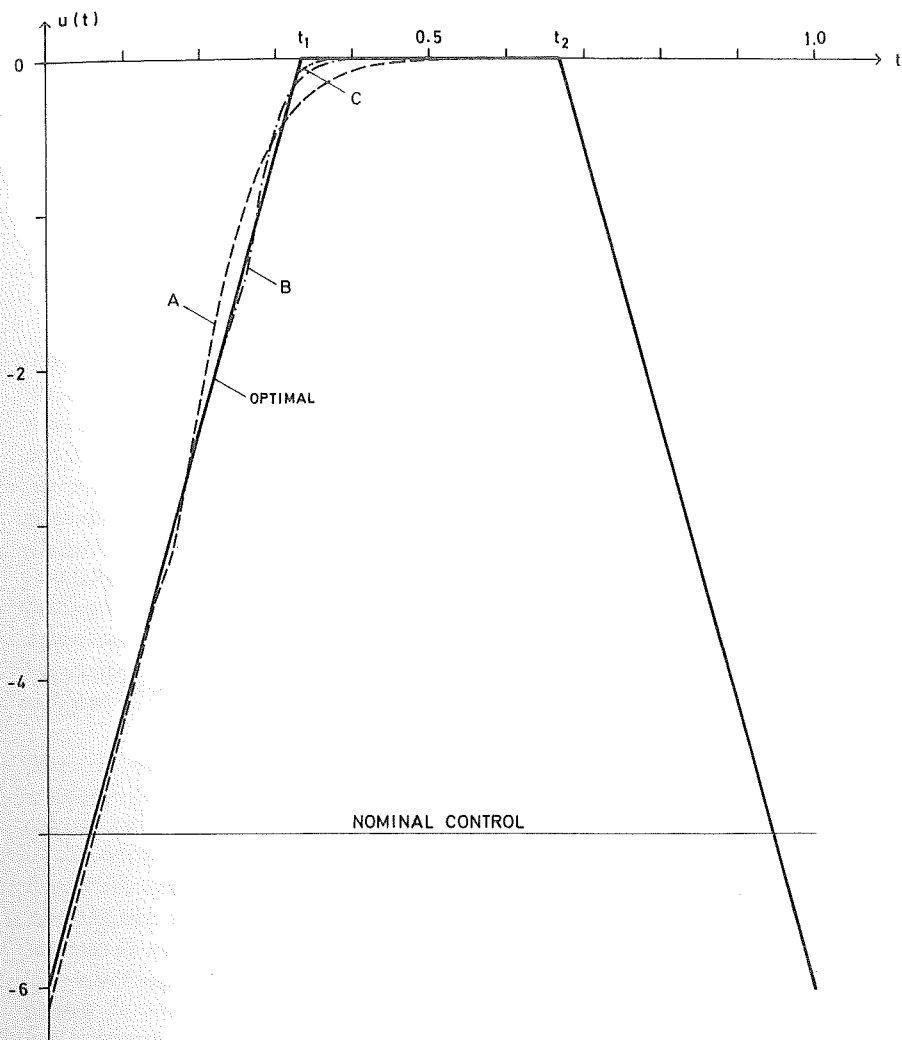


Fig 3. - Example 6.1: Computed solutions  $u(t)$  for the hyperplanes A(---), B(-.-) and C(-.-.-) compared with the optimal solution (—).

The initial guess of the terminal multipliers was  $\bar{b}^T = (-5, 5)$ , and the problem was "balanced" [19] by adding the term  $20\psi^T\psi$  to the cost functional. This slightly decreased the number of iterations, but was not critical for convergence. Nor was the computed solutions sensitive to the initial guesses  $\bar{b}$  and  $\bar{u}(t)$ .

Table II. Example 6.1: Computed costs and number of iterations for different hyperplanes ( $\eta_1 = 0.002$ )

	Computed cost	No. of iterations
A	4.0213	5
B	4.0015	12
C	4.0004	18

The computed adjoint variables  $\lambda(t)$  illustrate further interesting properties of the constraining hyperplane technique. We have:

I. Bryson's necessary conditions:

$$\lambda_1(t) = \begin{cases} 18 & 0 \leq t < 1/3 \\ -18 & 1/3 \leq t \leq 1 \end{cases}$$

$$\lambda_2(t) = \begin{cases} 16 - 18t & 0 \leq t < 1/3 \\ 18t - 12 & 1/3 \leq t \leq 1 \end{cases}$$

II. Speyer's necessary conditions:

$$\lambda_1(t) = \begin{cases} 18 & 0 \leq t < 1/3 \\ 0 & 1/3 \leq t < 2/3 \\ -18 & 2/3 \leq t \leq 1 \end{cases}$$

$$\lambda_2(t) = \begin{cases} 6 - 18t & 0 \leq t \leq 1/3 \\ 0 & 1/3 \leq t \leq 2/3 \\ 18t - 12 & 2/3 \leq t \leq 1 \end{cases}$$

$$\eta(t) = 0 \quad 0 \leq t \leq 1$$

Thus  $\lambda(t)$  suffer from discontinuities in both Bryson's and Speyer's necessary conditions. In Figs 4 and 5 these conditions are compared with the computed adjoint variables ( $V_X(t)$ ). It can be seen that the adjoint variables converge toward Speyer's necessary conditions as the slope of the hyperplane increases, and not toward Bryson's. A plausible explanation is the similarity between the constraining hyperplane technique and the derivation of the generalized Kuhn-Tucker theorem in function space [14], from which Speyer's necessary conditions are derived. It is thus suggested that the necessary conditions given by Jacobson, Lele and Speyer [8], can be derived from the constraining hyperplane technique through a limiting approach. This problem is subject to present research.

The overshoot in  $\lambda_1$  (Fig 4) is a typical property of hyperplane constrained problems. It should be emphasized that this is not due to numerical errors in the integration of  $\lambda$ , but to inherent properties of the problem. In Fig 6, computed solutions  $\lambda_1(t)$  for the hyperplane C and for different values of the acceptance parameter  $\eta_1$  ( $a(x_0, \bar{b}; t_0) < \eta_1$ ) are shown. It can be seen that the reduction of the overshoot is a matter of reducing the cost as far as the numerical accuracy permits. The corresponding values of the cost and the number of iterations are shown in Table III. This clearly illustrates the fast convergence to an approximate solution, and the slower convergence thereafter to an accurate solution. The corresponding control variables are shown in Fig 6. By inspection of Tables II and III, it can be seen, that if a modest accuracy is required, the fastest convergence is obtained with hyperplane A and the acceptance level  $\eta_1 = 0.002$ . Thus it is of no sense to use a steep hyperplane (C) unless the computation is carried out to a high degree of accuracy, that is, the acceptance parameter  $\eta_1$ , should be as small as the numerical accuracy permits. This similarity between state variable constrained problems and singular problems has been noted previously [8] [9].

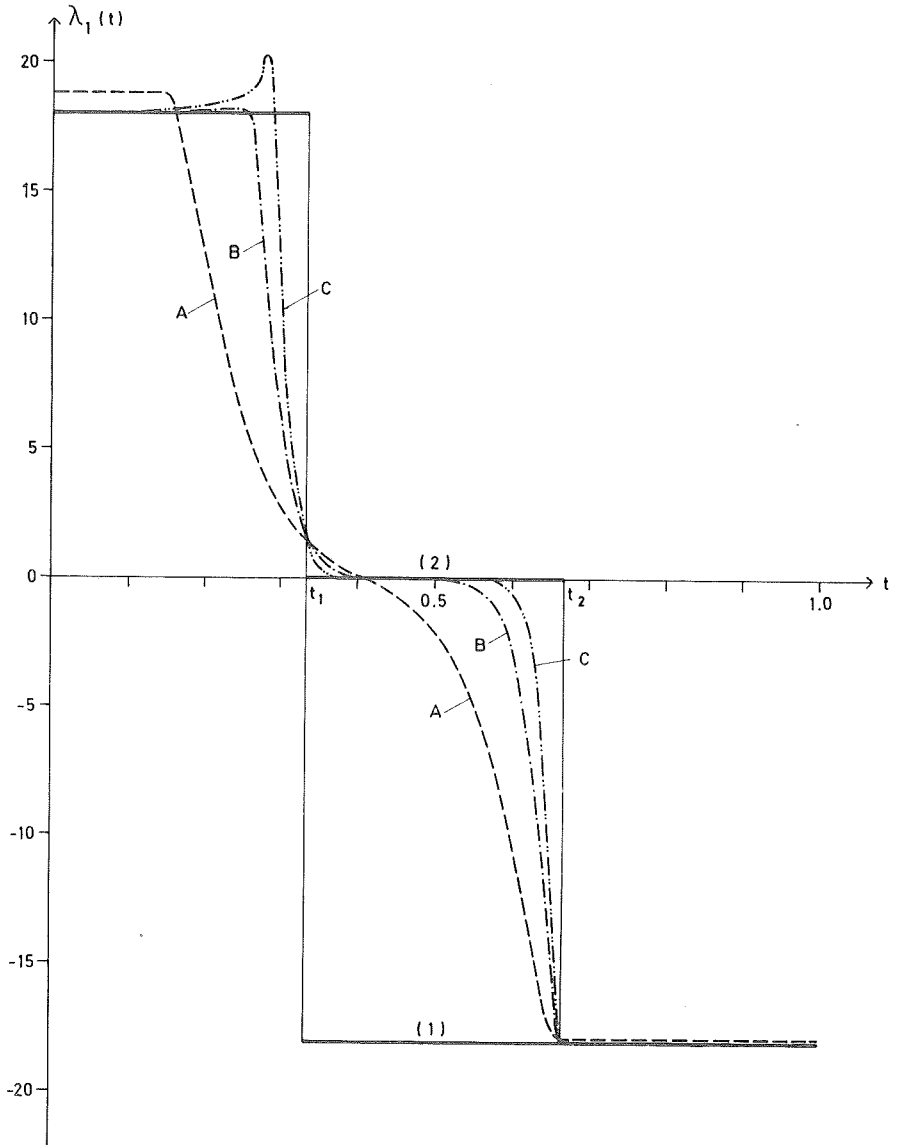


Fig 4. - Example 6.1: Computed adjoint variables  $\lambda_1(t)$  compared with Bryson's  $(\frac{1}{-})$  and Speyer's  $(\frac{2}{-})$  necessary conditions.

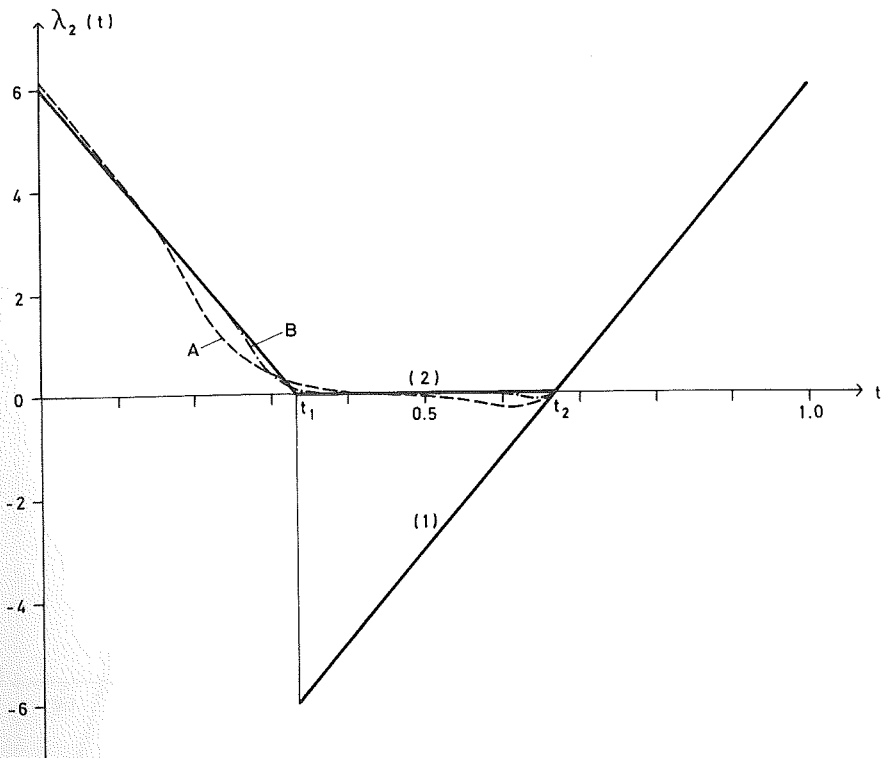


Fig 5. - Example 6.1: Computed adjoint variables  $\lambda_2(t)$  compared with Bryson's <sup>(1)</sup> and Speyer's <sup>(2)</sup> necessary conditions.

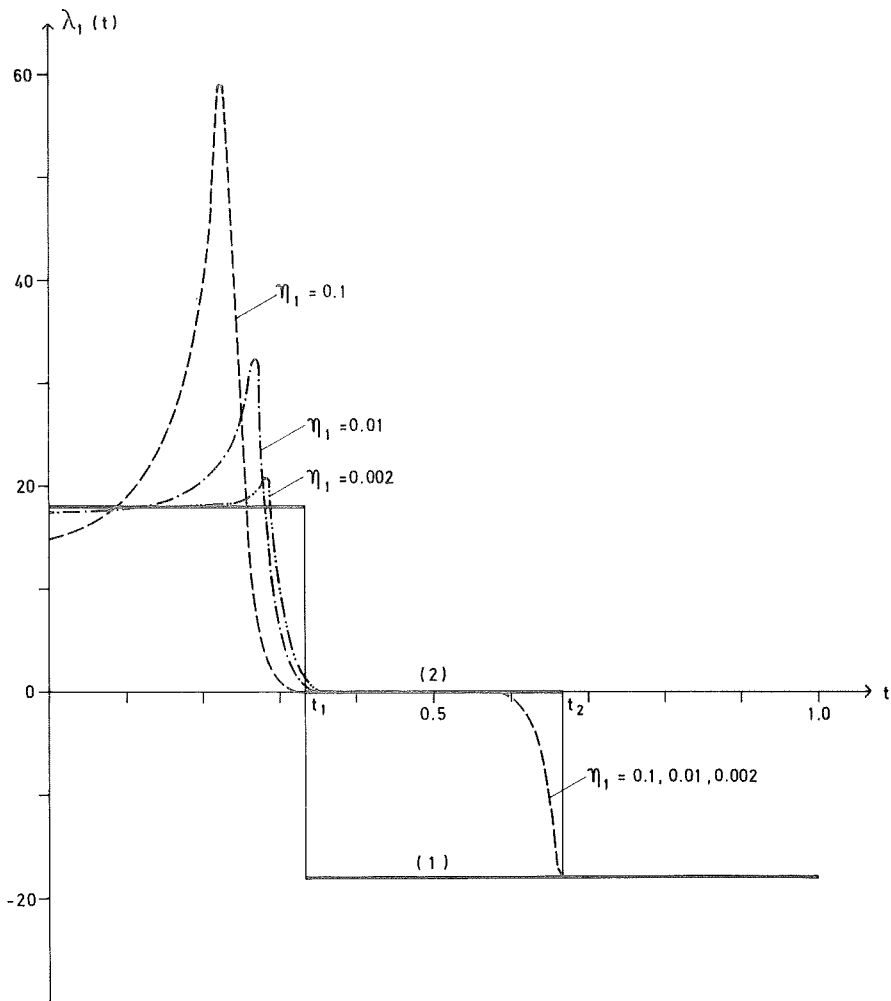


Fig 6. - Example 6.1: Computed adjoint variables  $\lambda_1(t)$  for  $\eta_1 = 0.1$  (---),  $\eta_1 = 0.01$  (-.-) and  $\eta_1 = 0.002$  (-.-.-) (hyperplane C).



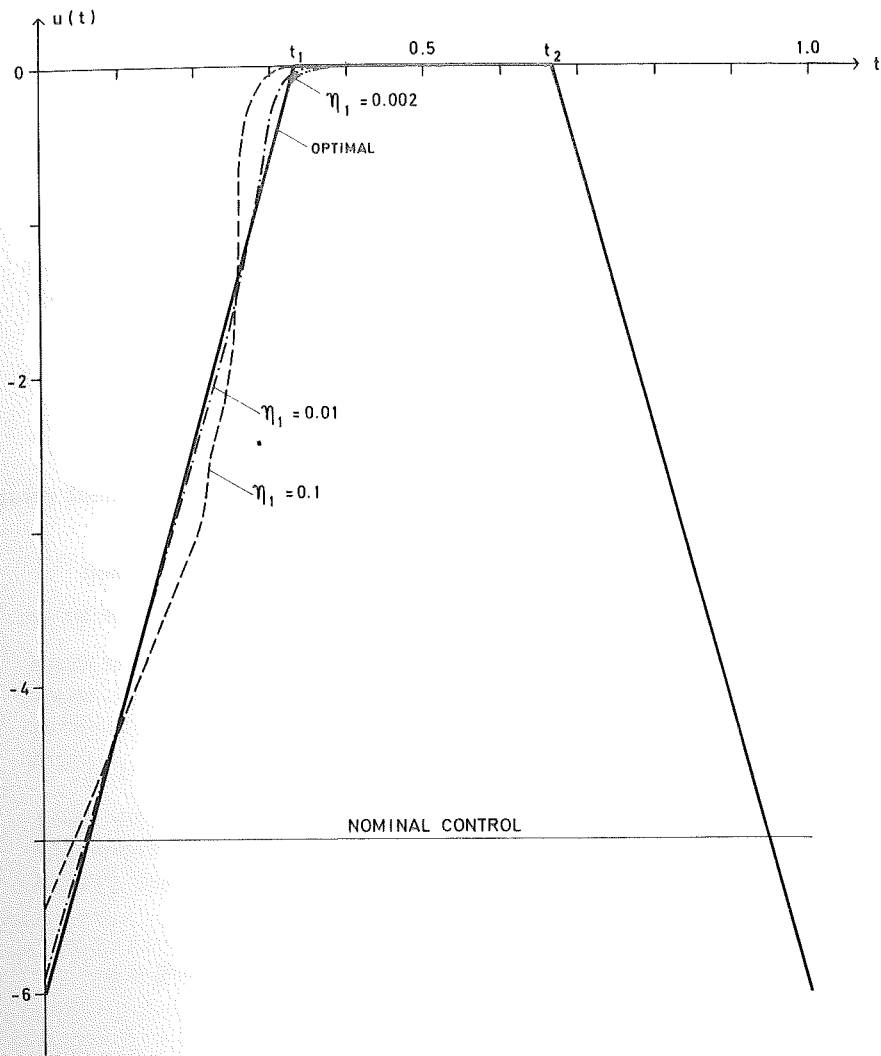


Fig 7. - Example 6.1: Computed control variables  $u(t)$  for  $\eta_1 = 0.1$  (---),  $\eta_1 = 0.01$  (-·-) and  $\eta_1 = 0.002$  (-··-) (hyperplane C).

Table III. Example 6.1: Computed costs and number of iterations for different values of  $\eta_1$  (hyperplane C).

$\eta_1$	Computed cost	No. of iterations
0.1	4.0401	11
0.01	4.0025	15
0.002	4.0004	18

In Fig 8, the second order partial derivative  $V_{x_1 x_1}$  of the optimal return is outlined. When the constraint is active,  $Z = 0$ ,  $Q = 1$ , and then

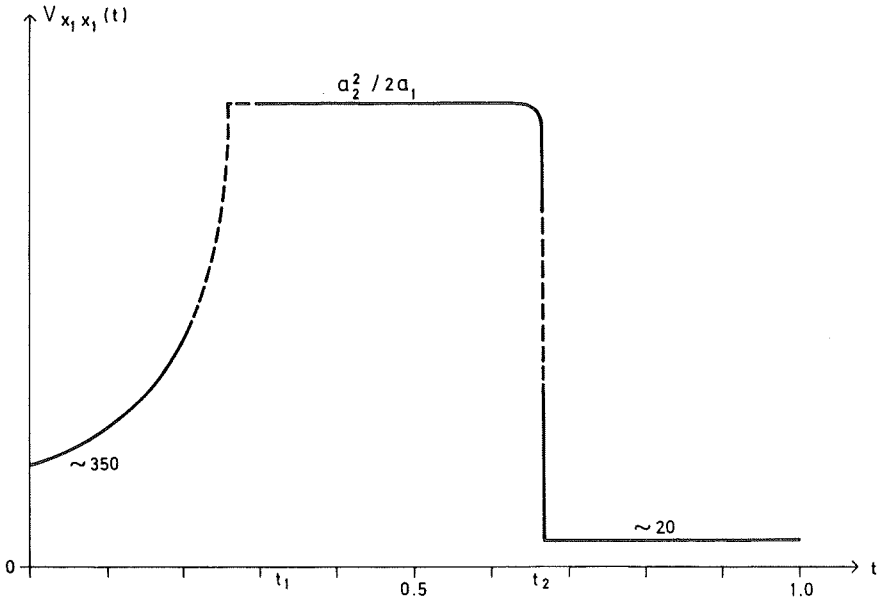


Fig 8. - Example 6.1: Outline of computed  $V_{x_1 x_1}(t)$ .

(5.39) reduces to

$$-\frac{dV_{xx}}{dt} = (f_x - f_u \hat{g}_x)^T V_{xx} + V_{xx} (f_x - f_u \hat{g}_x) + \hat{g}_x^T H_{uu} \hat{g}_x \quad (6.6)$$

But  $f_x - f_u \hat{g}_x$  is stable with eigenvalues equal to the eigenvalues of the hyperplane (Table I). When the constraint is active,  $V_{xx}$  then tends to the stationary solution

$$P_{ch} = \begin{bmatrix} \frac{a_2}{2a_1} & \frac{a_2}{2} \\ \frac{a_2}{2} & \frac{a_2}{2a_1} + \frac{a_1}{2} \end{bmatrix}$$

of (6.6). Thus  $V_{xx}(t) \rightarrow \infty$ ,  $t \in [t_1, t_2)$ , as the slope of the hyperplane tends to infinity, that is, when the admissible region tends to the half-space  $S(t) \leq 0$ .  $V$  then gets the interpretation as the optimal return of the adjoined functional  $\bar{J}$  subject to the constraint  $S(x; t) = 0$ ,  $t \in [t_1, t_2]$ .

From this point of view, it is interesting to compare the constraining hyperplane technique and the Kelley penalty function technique [11]. In [11] the following modified cost functional is formed

$$J = F(x(t_f); t_f) + \int_{t_0}^{t_f} \{L(x, u; t) + r_k h(S) S^2(x; t)\} dt \quad (6.7)$$

where

$$h(S) = \begin{cases} 1 & S \geq 0 \\ 0 & S < 0 \end{cases}$$

The cost (6.7) is then successively minimized for a monotone increasing sequence  $\{r_k\}$  with  $\lim_{k \rightarrow \infty} r_k = \infty$ . Disregarding the possi-

bility of local isolated minima of the original problem, it is obvious that the optimal solution of each  $r_k$ -problem does not satisfy the con-

straint  $S(t) \leq 0$ , and in general  $S(t) > 0$  over some interval  $(t_{1k}, t_{2k})$  with  $\lim_{k \rightarrow \infty} (t_{1k}, t_{2k}) = 0$ . Necessary conditions for the  $r_k$ -problem in  $(t_{1k}, t_{2k})$  then are

$$-\frac{d\lambda}{dt} = H_x^T + 2r_k S S_x^T \quad \lambda(t_{2k}^-) = \lambda_{2k}$$

$$H_u = 0$$

where

$$H(x, u, \lambda; t) = L(x, u; t) + \lambda^T f(x, u; t)$$

By restricting the analysis to the open interval  $(t_{1k}, t_{2k})$ , we may thus disregard the possible discontinuities in  $\lambda$  at entry or exit times.

Then consider the second order derivatives  $V_{xx}$  of the optimal return function of (6.7) in the interval  $(t_{1k}, t_{2k})$ . From (5.39) - (5.41) follows that

$$-\frac{dV_{xx}}{dt} = H_{xx} + f_x^T V_{xx} + V_{xx} f_x - V_{xx} f_u H_{uu}^{-1} f_u^T V_{xx} \quad (6.8)$$

since there are no control variable constraints present. It is easily verified that

$$P_k = \begin{bmatrix} \sqrt{4r_k} \sqrt{2r_k} & \sqrt{2r_k} \\ \sqrt{2r_k} & \sqrt{2\sqrt{2r_k}} \end{bmatrix}$$

is a stationary solution of (6.8) for any  $r_k$ -problem, and that  $V_{xx}$  in  $(t_{1k}, t_{2k})$  converges to  $P_k$  independent of the boundary condition  $V_{xx}(t_{2k}^-)$  [15]. Since  $P_k \rightarrow \infty$  as  $r_k \rightarrow \infty$ , the previous interpretation of  $V$  as the optimal return subject to  $S(x; t) = 0$ ,  $t \in [t_1, t_2]$ , is thus preserved. However, the rates with which the stationary solutions  $P_{ch}$  and  $P_k$  tend to infinity differ significantly. To see

this, consider the hyperplane C. Then

$$P_{ch} \approx \begin{bmatrix} 1.4 \times 10^5 & 3.4 \times 10^3 \\ 3.4 \times 10^3 & 1.0 \times 10^2 \end{bmatrix} \quad (6.9)$$

while  $r_k = a_2$  ( $= 6800$ ) yields

$$P_k \approx \begin{bmatrix} 1.8 \times 10^3 & 1.2 \times 10^2 \\ 1.2 \times 10^2 & 1.5 \times 10^1 \end{bmatrix}$$

It turns out that it is necessary to increase  $r_k$  to  $10^7 - 10^8$  to get a stationary solution  $P_k$  of the same magnitude as (6.9). This is well confirmed by computational results with penalty function methods. In [9] and [13] it has proved necessary to successively increase  $r_k$  to  $10^6 - 10^8$  to reach a sufficiently accurate solution. It may then be concluded, that the constraining hyperplane technique transforms the original state-variable constrained problem in a much "harder" and sensitive way than penalty function methods.

The difference in sensitivity will also affect the computation of the adjoint variables. Recalling the necessary conditions of Speyer et al., it is clear that  $\eta(t), t \in (t_1, t_2)$ , should be identified with  $\lim_{k \rightarrow \infty} (2r_k S)$  in the penalty function method, and with  $\lim (\mu(t)a_2)$  in the constraining hyperplane technique (for second order constraints). From the preceding sensitivity analysis, it can then be expected that  $\mu(t)$  tends to zero much slower than  $S$ , and consequently that the computation of  $\mu(t)a_2$  is less sensitive to numerical errors than the limit value of  $2r_k S$ .

The computed Lagrange multipliers  $\mu(t)$  for different hyperplanes are shown in Fig 9. It is easily verified that  $\mu(t)a_2 \rightarrow 0$  in the interior of  $[t_1, t_2]$ , and that  $\mu(t)a_2 \rightarrow \infty$  in decreasing neighbourhoods of  $t_1$  and  $t_2$ , thus providing the discontinuities in  $\lambda$  at entry and exit times.

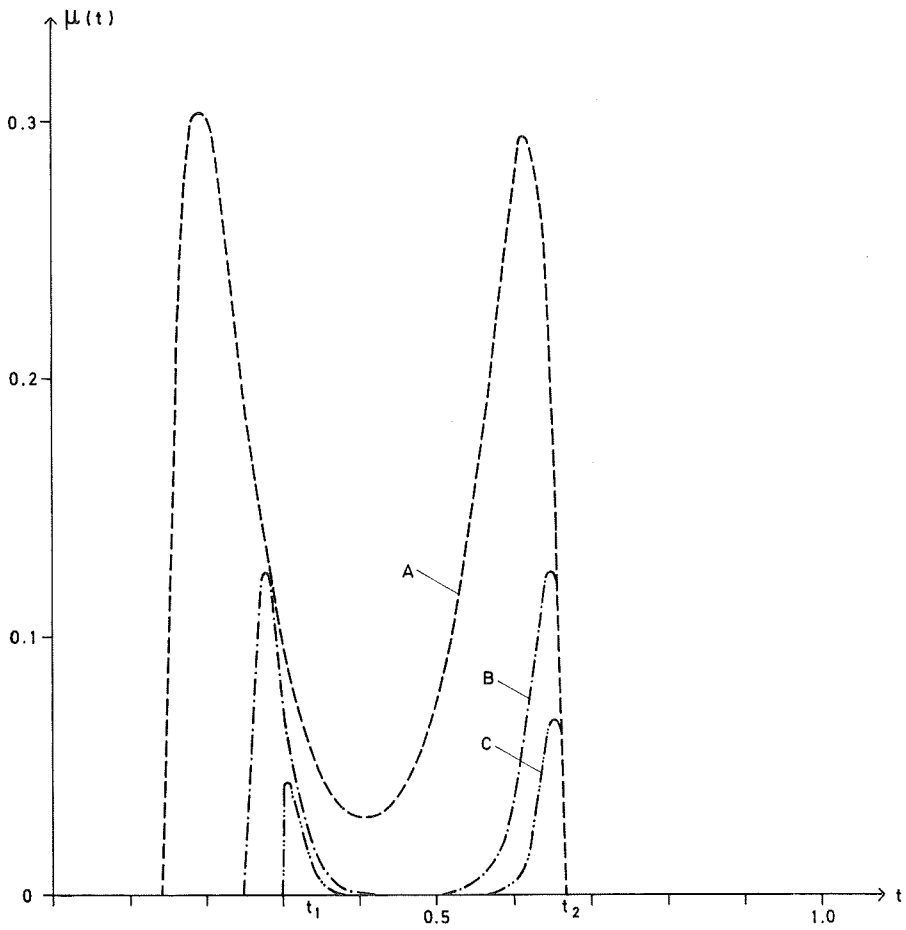


Fig 9. - Example 6.1: Computed Lagrange multipliers  $\mu(t)$  for the hyperplanes A (---), B (-·-) and C (-·-·-).

### Example 6.2

We will consider the system

$$\frac{dx_1}{dt} = x_2 \quad x_1(0) = 0 \quad (6.10)$$

$$\frac{dx_2}{dt} = u \quad x_2(0) = 0$$

with the cost functional

$$J = \frac{1}{2} \int_0^1 u^2 dt \quad (6.11)$$

the terminal constraints

$$\psi_1 = x_1(1) - 1 = 0 \quad (6.12)$$

$$\psi_2 = x_2(1) = 0$$

and the second order state variable inequality constraint

$$S(x; t) = x_1(t) - 8(t - 0.5)^2 + 0.5 \leq 0 \quad (6.13)$$

The optimal solution is:

$$x_1(t) = \begin{cases} (2048t^3 - 1512t^2)/243 & 0 \leq t \leq 9/16 \\ (16t^2 - 16t + 3)/2 & 9/16 \leq t \leq 13/16 \\ (-2048t^3 + 5208t^2 - 4272t + 1139)/27 & 13/16 \leq t \leq 1 \end{cases}$$
$$x_2(t) = \begin{cases} (2048t^2 - 1008t)/81 & 0 \leq t \leq 9/16 \\ 16t - 8 & 9/16 \leq t \leq 13/16 \\ (-2048t^2 + 3472t - 1424)/9 & 13/16 \leq t \leq 1 \end{cases} \quad (6.14)$$

$$u(t) = \begin{cases} (4096t - 1008)/81 & 0 \leq t \leq 9/16 \\ 16 & 9/16 \leq t \leq 13/16 \\ (-4096t + 3472)/9 & 13/16 \leq t \leq 1 \end{cases}$$

Thus the entry time is  $t_1 = 9/16$  and the exit time is  $t_2 = 13/16$ .

The optimal cost is  $J = 4736/27 \approx 175.407$ .

Since the constraint (6.13) is of second order, the constraining hyperplane is

$$\frac{d^2 S}{dt^2} + a_1 \frac{dS}{dt} + a_2 S = 0$$

and the transformed problem then is to minimize the cost functional (6.11) subject to the terminal constraints (6.12) and the mixed state-control variable constraint

$$g(x, u; t) = u - 16 + a_1(x_2 - 16t + 8) + a_2(x_1 - 8t^2 + 8t - \frac{3}{2}) \leq 0$$

The same hyperplanes as in example 6.1 were chosen (Table I). In Figs 10 and 11 the optimal solution (6.14) is compared with the computed solutions for the different hyperplanes, and it is clear that very little is gained by increasing the slope of the hyperplanes. The maximum deviations in  $x_1$  are in all cases about 0.006, and the acceptance parameters  $\eta_1 = 0.1$  and  $\eta_2 = 0.002$  were used. This relative insensitivity of the solution to the steepness of the hyperplanes is also illustrated by Table IV, where the resulting optimal costs and the number of iterations are shown. It can be seen that very small decreases in the optimal cost are obtained at the rate of a substantial increase in the number of iterations. (Notice that the computed costs in case B and C are smaller than the optimal cost. This is due to the fact that the terminal constraints are not exactly satisfied ( $\eta_2 = 0.002$ )).



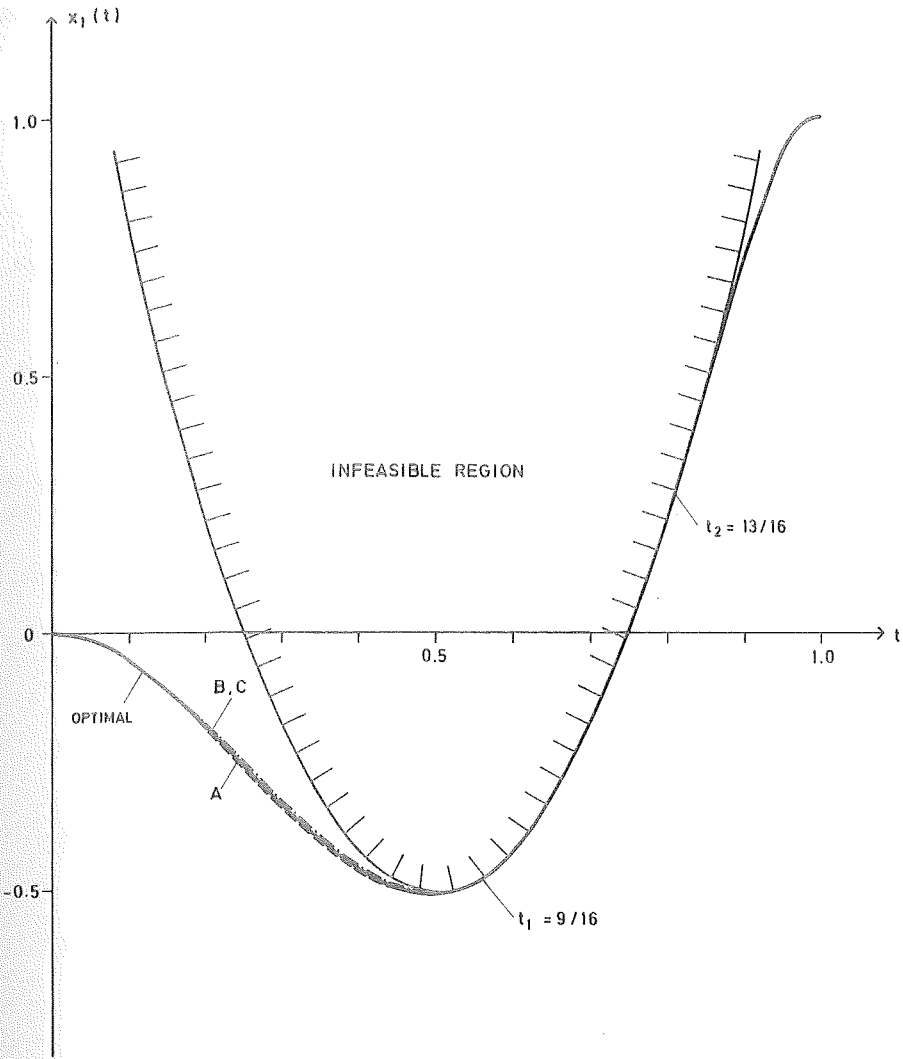


Fig 10. - Example 6.2: Computed solutions  $x_1(t)$  for the hyperplanes A (---), B (-·-) and C ( $\approx$  B) compared with the optimal solution (—).

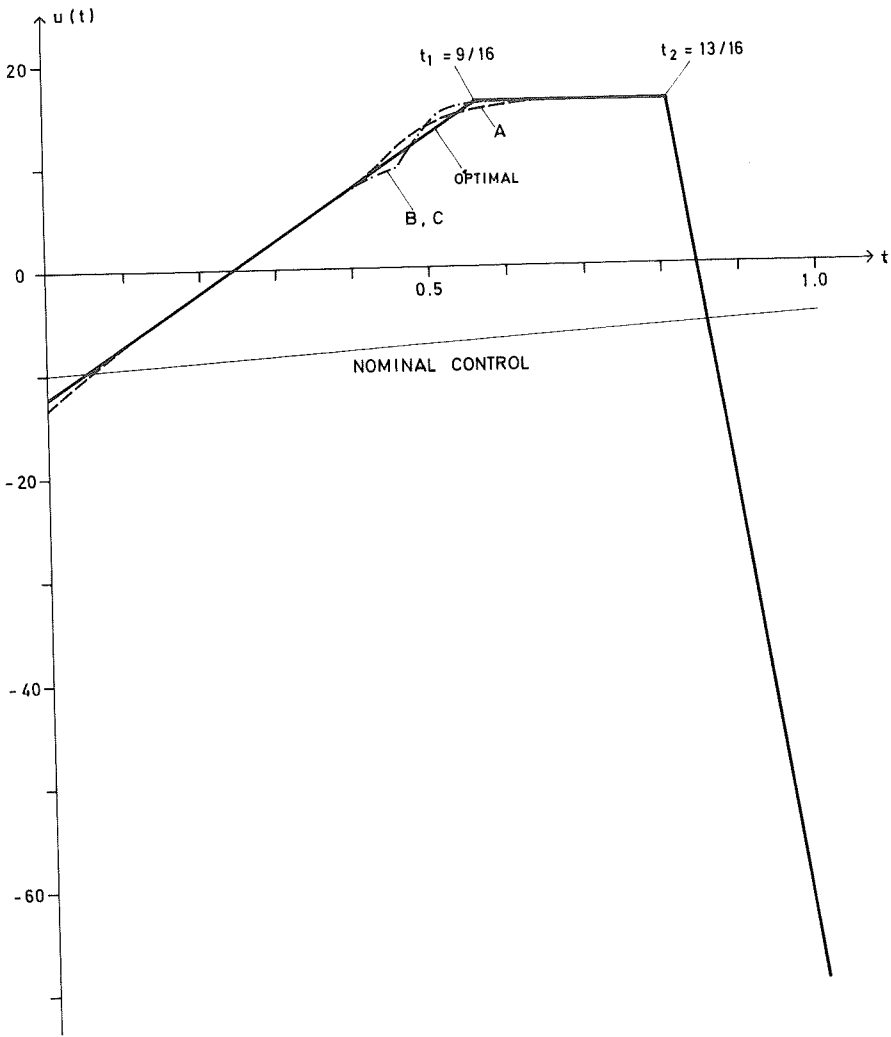


Fig 11. - Example 6.2: Computed solutions  $u(t)$  for the hyperplanes A (---), B (-.-) and C ( $\approx$  B) compared with the optimal solution (—).

Table IV. Example 6.2: Computed costs and number of iterations for different hyperplanes ( $\eta_1 = 0.1$ ).

	Computed cost	No. of iterations
A	175.47	6
B	175.35	17
C	175.22	22

The adjoint variables exhibit the same properties as in the preceding example. We have:

I. Bryson's necessary conditions:

$$\lambda_1(t) = \begin{cases} 4096/81 & 0 \leq t < 9/16 \\ -4096/9 & 9/16 \leq t \leq 1 \end{cases}$$

$$\lambda_2(t) = \begin{cases} (-4096t + 1008)/81 & 0 \leq t < 9/16 \\ (4096t - 3472)/9 & 9/16 \leq t \leq 1 \end{cases}$$

II. Speyer's necessary conditions:

$$\lambda_1(t) = \begin{cases} 4096/81 & 0 \leq t < 9/16 \\ 0 & 9/16 \leq t < 13/16 \\ -4096/9 & 13/16 \leq t \leq 1 \end{cases}$$

$$\lambda_2(t) = \begin{cases} (-4096t + 1008)/81 & 0 \leq t \leq 9/16 \\ -16 & 9/16 \leq t \leq 13/16 \\ (4096t - 3472)/9 & 13/16 \leq t \leq 1 \end{cases}$$

$$\eta(t) = 0$$

$$0 \leq t \leq 1$$

In Fig 12 the computed adjoint variables are compared with Bryson's and Speyer's necessary conditions, and it can be seen that they tend to satisfy Speyer's conditions as the slope of the hyperplane increases.

### Example 6.3

We will consider the system

$$\frac{dx_1}{dt} = x_2$$

$$x_1(0) = 0$$

$$\frac{dx_2}{dt} = x_3$$

$$x_2(0) = 1$$

(6.15)

$$\frac{dx_3}{dt} = u$$

$$x_3(0) = 2$$

with the cost functional

$$J = \frac{1}{2} \int_0^1 u^2 dt$$

(6.16)

the terminal constraints

$$\psi_1 = x_1(1) = 0$$

$$\psi_2 = x_2(1) + 1 = 0$$

$$\psi_3 = x_3(1) - 2 = 0$$

(6.17)

and the third order state variable inequality constraint

$$S(x; t) = x_1(t) - \ell \leq 0$$

(6.18)

The explicit solution for different values of the parameter  $\ell$  was given in [8]. It was shown that if  $\ell > 3/8$ , the constraint is not

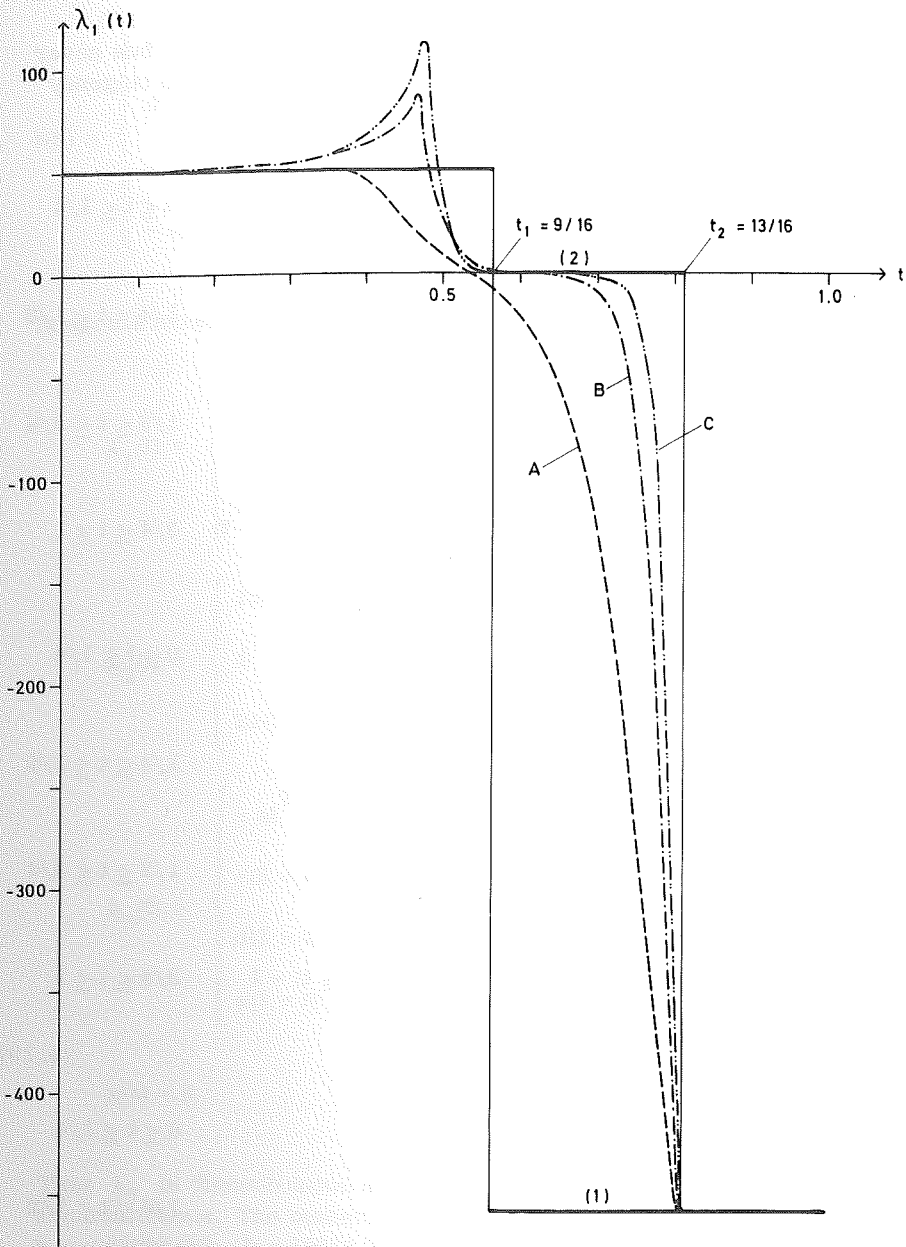


Fig 12. - Example 6.2: Computed adjoint variables  $\lambda_1(t)$  compared with Bryson's (1) and Speyer's (2) necessary conditions.

active, if  $9/40 \leq \ell \leq 3/8$  there is only a tangency point at  $t = 1/2$ , and if  $\ell < 9/40$  there are two tangency points. The solution thus never remains on the constraint. For  $\ell = 293/2080 \approx 0.14087$  the tangency points are  $t_1 = 1/4$ ,  $t_2 = 3/4$ , and the optimal solution is:

$$x_1(t) = \begin{cases} \frac{at^5}{60} + \frac{bt^4}{24} + \frac{ct^3}{6} + t^2 + t & 0 \leq t \leq 1/4 \\ \frac{dt^4}{24} + \frac{et^3}{6} + \frac{ft^2}{2} + gt + h & 1/4 \leq t \leq 1/2 \\ \frac{d(1-t)^4}{24} + \frac{e(1-t)^3}{6} + \frac{f(1-t)^2}{2} + g(1-t) + h & 1/2 \leq t \leq 3/4 \\ \frac{a(1-t)^5}{60} + \frac{b(1-t)^4}{24} + \frac{c(1-t)^3}{6} + (1-t)^2 + (1-t) & 3/4 \leq t \leq 1 \end{cases}$$

$$x_2(t) = \begin{cases} \frac{at^4}{12} + \frac{bt^3}{6} + \frac{ct^2}{2} + 2t + 1 & 0 \leq t \leq 1/4 \\ \frac{dt^3}{6} + \frac{et^2}{2} + ft + g & 1/4 \leq t \leq 1/2 \\ -\frac{d(1-t)^3}{6} - \frac{e(1-t)^2}{2} - f(1-t) - g & 1/2 \leq t \leq 3/4 \\ -\frac{a(1-t)^4}{12} - \frac{b(1-t)^3}{6} - \frac{c(1-t)^2}{2} - 2(1-t) - 1 & 3/4 \leq t \leq 1 \end{cases}$$

(6.19)

$$x_3(t) = \begin{cases} \frac{at^3}{3} + \frac{bt^2}{2} + ct + 2 & 0 \leq t \leq 1/4 \\ dt^2 + et + f & 1/4 \leq t \leq 1/2 \\ \frac{d(1-t)^2}{2} + e(1-t) + f & 1/2 \leq t \leq 3/4 \\ \frac{a(1-t)^3}{3} + \frac{b(1-t)^2}{3} + c(1-t) + 2 & 3/4 \leq t \leq 1 \end{cases}$$

$$u(t) = \begin{cases} at^2 + bt + c & 0 \leq t \leq 1/4 \\ dt + e & 1/4 \leq t \leq 1/2 \\ -d(1-t) - e & 1/2 \leq t \leq 3/4 \\ -a(1-t)^2 - b(1-t) - c & 3/4 \leq t \leq 1 \end{cases}$$

Speyer's necessary conditions are:

$$\lambda_1(t) = \begin{cases} -2a & 0 \leq t < 1/4 \\ 0 & 1/4 \leq t < 3/4 \\ 2a & 3/4 \leq t \leq 1 \end{cases}$$

$$\lambda_2(t) = \begin{cases} 2at + b & 0 \leq t \leq 1/4 \\ d & 1/4 \leq t \leq 3/4 \\ 2a(1-t) + b & 3/4 \leq t \leq 1 \end{cases}$$

$$\lambda_3(t) = \begin{cases} -at^2 - bt - c & 0 \leq t \leq 1/4 \\ -dt - e & 1/4 \leq t \leq 1/2 \\ d(1-t) + e & 1/2 \leq t \leq 3/4 \\ a(1-t)^2 + b(1-t) + c & 3/4 \leq t \leq 1 \end{cases}$$

$$\eta(t) = 0 \quad 0 \leq t \leq 1$$

Thus  $\lambda_1$  is discontinuous at the entry and exit times, while  $\lambda_2$  and  $\lambda_3$  are continuous. The numerical values of the parameters are:

$a = -43008/13$ ,  $b = 19776/13$ ,  $c = -1824/13$ ,  $d = -1728/13$ ,  
 $e = 864/13$ ,  $f = -198/13$ ,  $g = 27/13$ ,  $h = -7/130$ , and the optimal  
cost is  $J = 734976/845 \approx 869.794$ .

Since the constraint is of order three, the constraining hyperplane is

$$\frac{d^3 S}{dt^3} + a_1 \frac{d^2 S}{dt^2} + a_2 \frac{dS}{dt} + a_3 S = 0$$

and the mixed state-control variable constraint becomes

$$g(x, u; t) = u + a_1 x_3 + a_2 x_2 + a_1 (x_1 - \ell) \leq 0$$

The problem was solved with the following parameters of the hyperplane:  $a_1 = 153$ ,  $a_2 = 7802$ ,  $a_3 = 132600$  (corresponding to:  $\xi_1 = -50$ ,  $\xi_2 = -51$ ,  $\xi_3 = -52$ ).

In Figs 13, 14 and 15 the computed solutions  $x_1(t)$ ,  $u(t)$  and  $\lambda_1(t)$  are shown together with the corresponding optimal solutions. Notice that the computed solution  $x_1(t)$  never reaches the constraint at the tangency points. This is a natural but undesirable consequence of the construction of the hyperplane. A simple way to overcome this problem is to slightly increase the parameter  $\ell$ , and this will also have a considerable influence on the computed cost. For the  $\ell$  chosen ( $\ell = 0.14087$ ), the optimal cost is  $J = 869.8$  while the computed cost is  $J = 903.6$ . Increasing  $\ell$  to 0.143, the original constraint is just slightly violated ( $\max x_1 = 0.1419$ ), but the computed cost is now reduced to  $J = 862.3$ . This sensitivity of the cost seems to be typical of optimal control problems with high order state variable inequality constraints.

The success of the algorithm proved to be dependent of good initial guesses of the terminal constraint multipliers  $b$  and of the initial nominal control. In this case  $B^T = (-6000, 1500, -100)$  was chosen, and the algorithm then converged to the optimal solution in 13 iterations ( $\eta_1 = 1.0$ ). A main problem was to get a new nominal control accepted over the whole interval  $[t_0, t_f]$  (cf. section 5). Once this was achieved, the algorithm proceeded smoothly. A possible way to overcome these difficulties is to start with a less steep hyperplane, and then use the computed terminal multipliers and optimal control as initial guesses for the actual hyperplane.



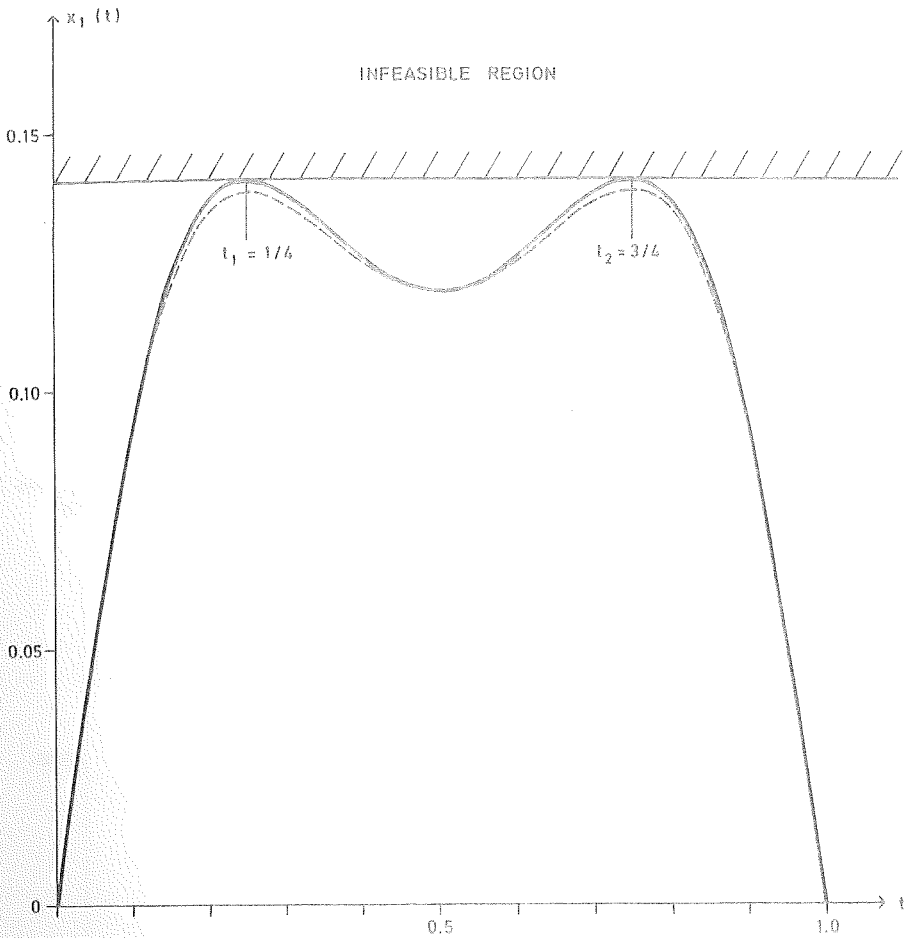


Fig 13. - Example 6.3: Computed solution  $x_1(t)$  (---) compared with the optimal solution (—).

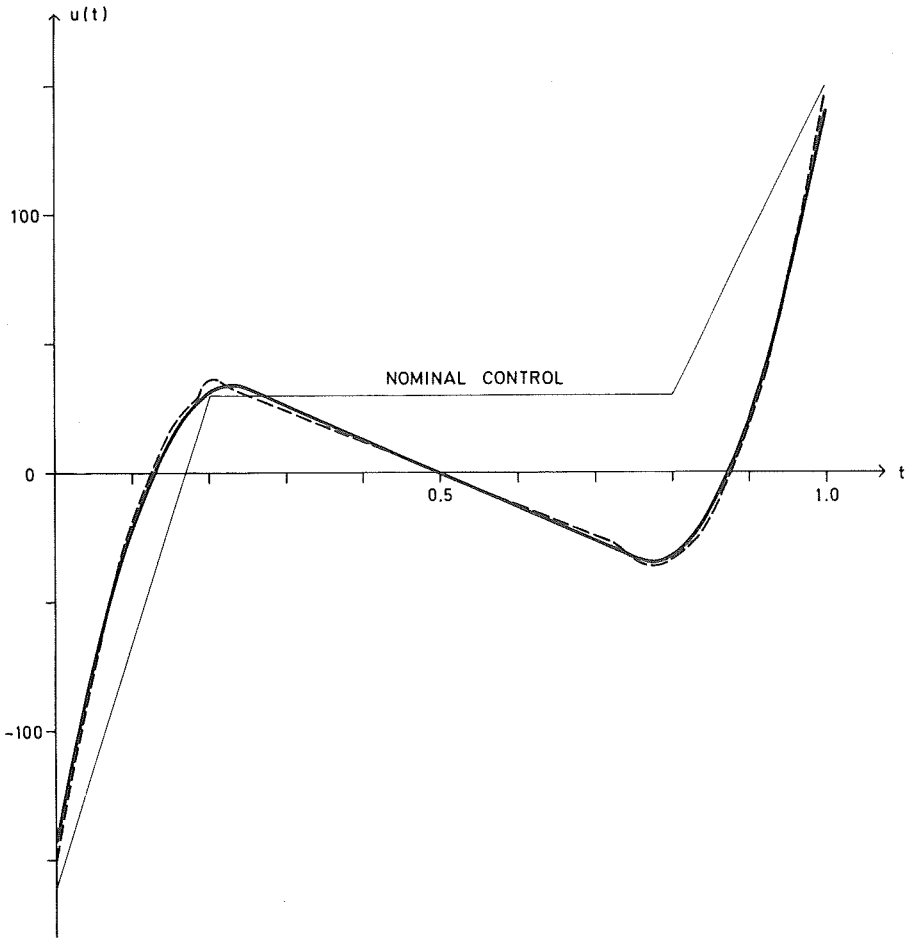


Fig 14. - Example 6.3: Computed solution  $u(t)$  (---) compared with the optimal solution (—).

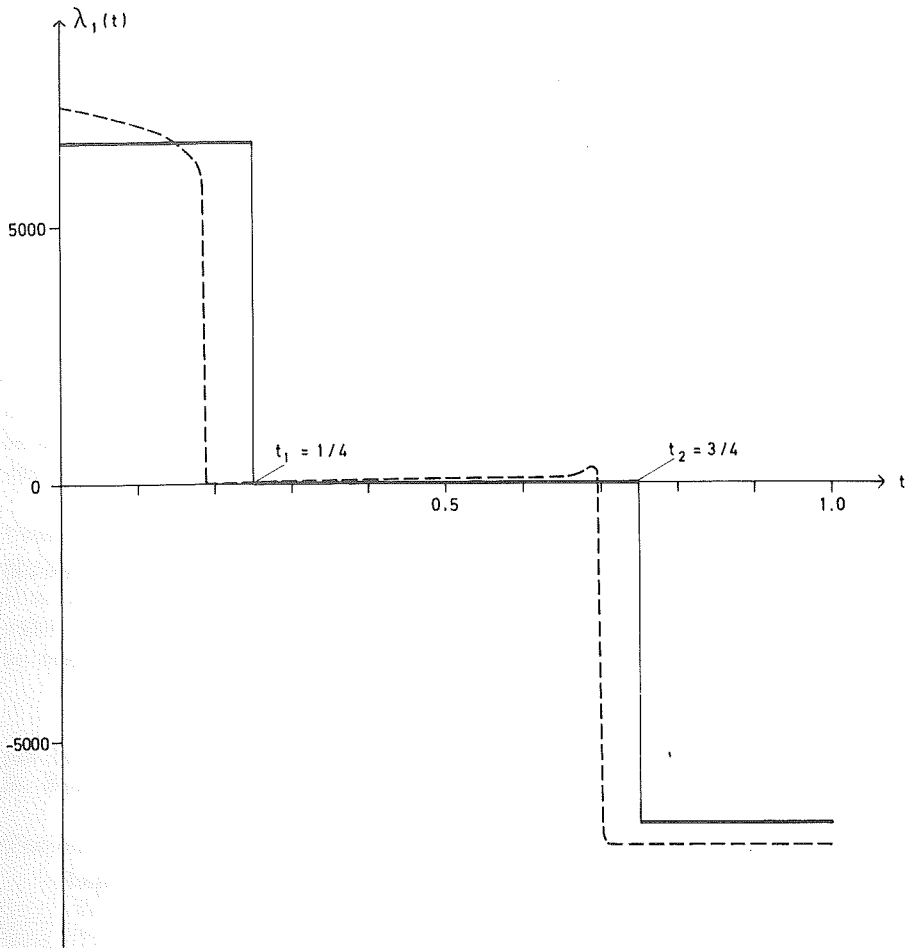


Fig 15. - Example 6.3: Computed solution  $\lambda_1(t)$  (---) compared with Speyer's necessary conditions ( $\longrightarrow$ ).

#### Example 6.4

In this example we will compare the efficiency of the constraining hyperplane technique with the slack variable technique proposed by Jacobson and Lele [9]. We will then consider the system

$$\frac{dx_1}{dt} = x_2 \quad x_1(0) = 0 \quad (6.20)$$

$$\frac{dx_2}{dt} = -x_2 + u \quad x_2(0) = -1$$

with the cost functional

$$J = \int_0^1 (x_1^2 + x_2^2 + 0.005u^2) dt \quad (6.21)$$

and two different state variable inequality constraints

$$S_1(x;t) = x_2 - 8(t - 0.5)^2 + 0.5 \leq 0 \quad (6.22)$$

$$S_2(x;t) = x_1 - 8(t - 0.5)^2 + 0.5 \leq 0$$

of first and second order respectively. The comparisons will be made with the results reported in [9].

#### I. First order constraint.

The constraining hyperplane is

$$\frac{dS_1}{dt} + a_1 S_1 = 0$$

and the transformed problem then is to minimize the cost functional (6.21) subject to the mixed state-control variable constraint

$$g(x, u; t) = u - x_2 - 16(t - 0.5) + a_1 [x_2 - 8(t - 0.5)^2 + 0.5] \leq 0$$

The problem was solved for  $a_1 = 20$  and  $a_1 = 50$  with the acceptance parameter  $\eta_1 = 0.001$ . In Figs 16 and 17 the computed solutions  $x_2(t)$  and  $u(t)$  are compared with the slack variable technique solutions. It can be seen that these agree fairly well with the solutions for  $a_1 = 20$ .

However, the efficiency of the methods differ significantly, and the obvious reason for this is that the slack variable technique transforms the original problem into a singular problem (where second order methods cannot be used). Thus the slack variable solution was reached after 16 iterations with a conjugate gradient algorithm [12], and the improvements in the last iterations were negligible. This should be compared with the constraining hyperplane technique, where 1 iteration was required for  $a_1 = 20$  ( $J = 0.17199$ ) and 2 iterations for  $a_1 = 50$  ( $J = 0.16996$ ). Notice that the same initial nominal control as in [9] was used. It may thus be suspected that the accuracy obtained with  $a_1 = 50$  is impossible to reach with the slack variable technique.

Contrary to the previous examples, the adjoint variables are continuous at the entry and exit times, and  $\mu(t)$  is not identically zero. This is illustrated in Fig 18, where the product  $a_1\mu(t)$  is shown for increasing values of  $a_1$ . Identifying with Speyer's necessary conditions, it can thus be concluded that  $\lim_{a_1 \rightarrow \infty} a_1\mu(t)$  exists and is equal to  $\eta(t)$ .

$$a_1 \rightarrow \infty$$

## II. Second order constraint.

The constraining hyperplane is

$$\frac{d^2 S_2}{dt^2} + a_1 \frac{dS_2}{dt} + a_2 S_2 = 0$$

and thus the state-control variable constraint becomes

$$g(x, u; t) = u - x_2 - 16 + a_1 \left[ x_2 - 16(t - 0.5) \right] + \\ + a_2 \left[ x_1 - 8(t - 0.5)^2 + 0.5 \right] \leq 0$$

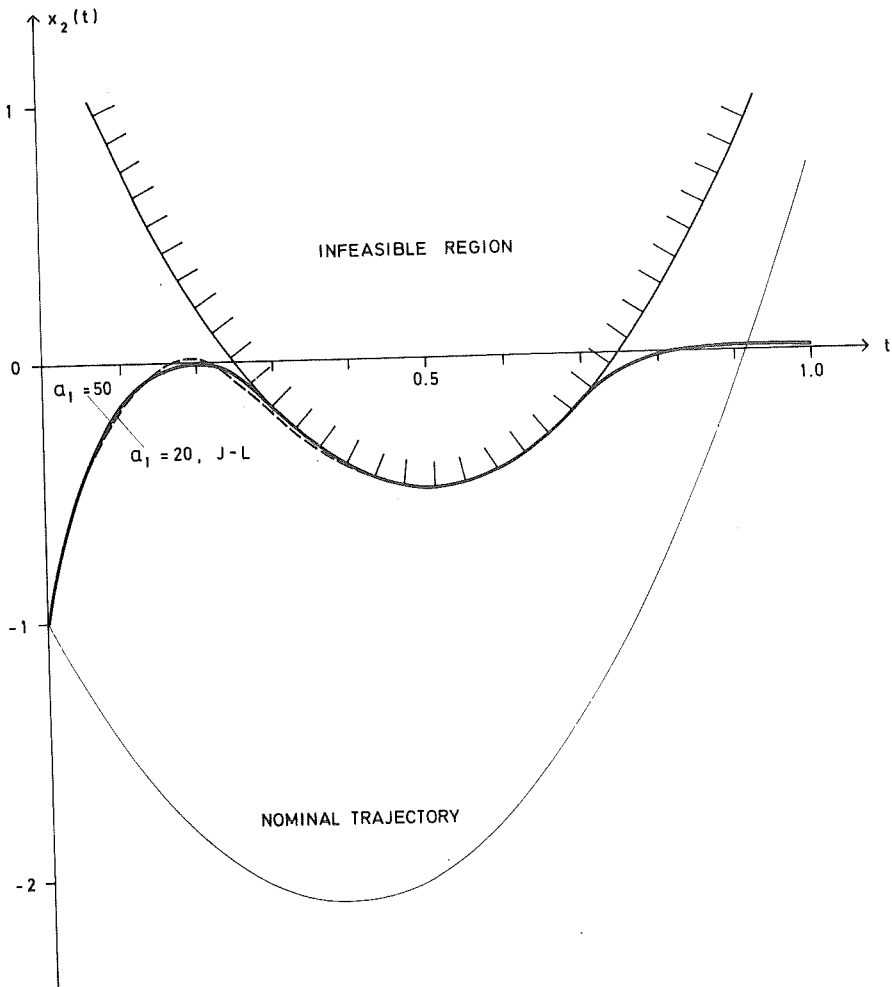


Fig 16. - Example 6.4: First order constraint  $x_2(t) - 8(t - 0.5)^2 + 0.5 \leq 0$ . Computed  $x_2(t)$  for  $a_1 = 20$  (---) and  $a_1 = 50$  (—), compared with slack variable technique solution (---) reported by Jacobson and Lele (J-L).

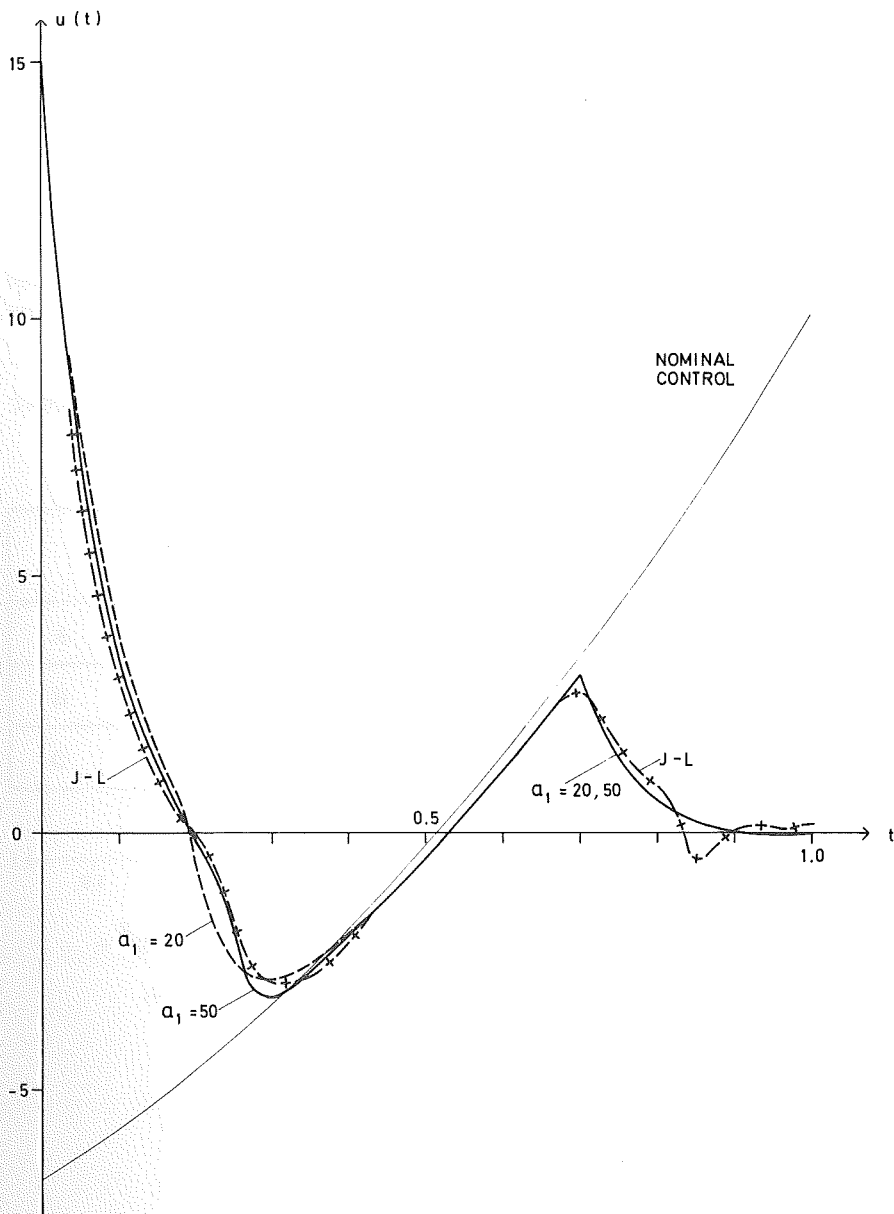


Fig 17. - Example 6.4: First order constraint  $x_2(t) - 8(t - 0.5)^2 + 0.5 \leq 0$ . Computed  $u(t)$  for  $a_1 = 20$  (---) and  $a_1 = 50$  (—), compared with slack variable technique solution (-+-).

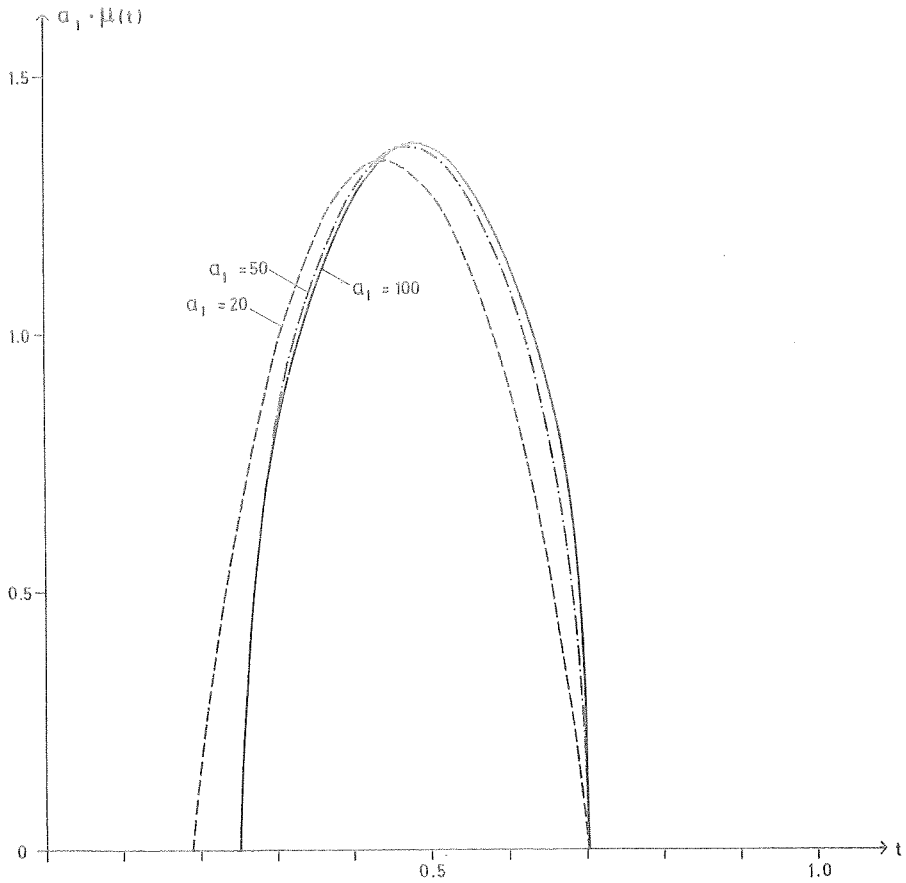


Fig 18. - Example 6.4: First order constraint  $x_2(t) - 8(t - 0.5)^2 + 0.5 \leq 0$ . Computed product  $a_1 \mu(t)$  for  $a_1 = 20$  (---),  $a_1 = 50$  (-·-) and  $a_1 = 100$  (—).



The hyperplanes given in Table I were used, and in Fig 19 the computed optimal control  $u(t)$  is compared with the slack variable technique solution. Notice that the same initial nominal control was used.

The slack variable solution was reached after 32 iterations with the conjugate gradient algorithm. In Table V the corresponding figures for the different hyperplanes are given together with the computed optimal costs. These clearly indicate that the combination of constraining hyperplanes and a second order method is superior as far as accuracy and efficiency are concerned.

Table V. Example 6.4: Second order constraint. Computed costs and number of iterations for different hyperplanes ( $\eta_1 = 0.001$ ).

	Computed cost	No. of iterations
A	0.77754	2
B	0.74345	4
C	0.74101	5

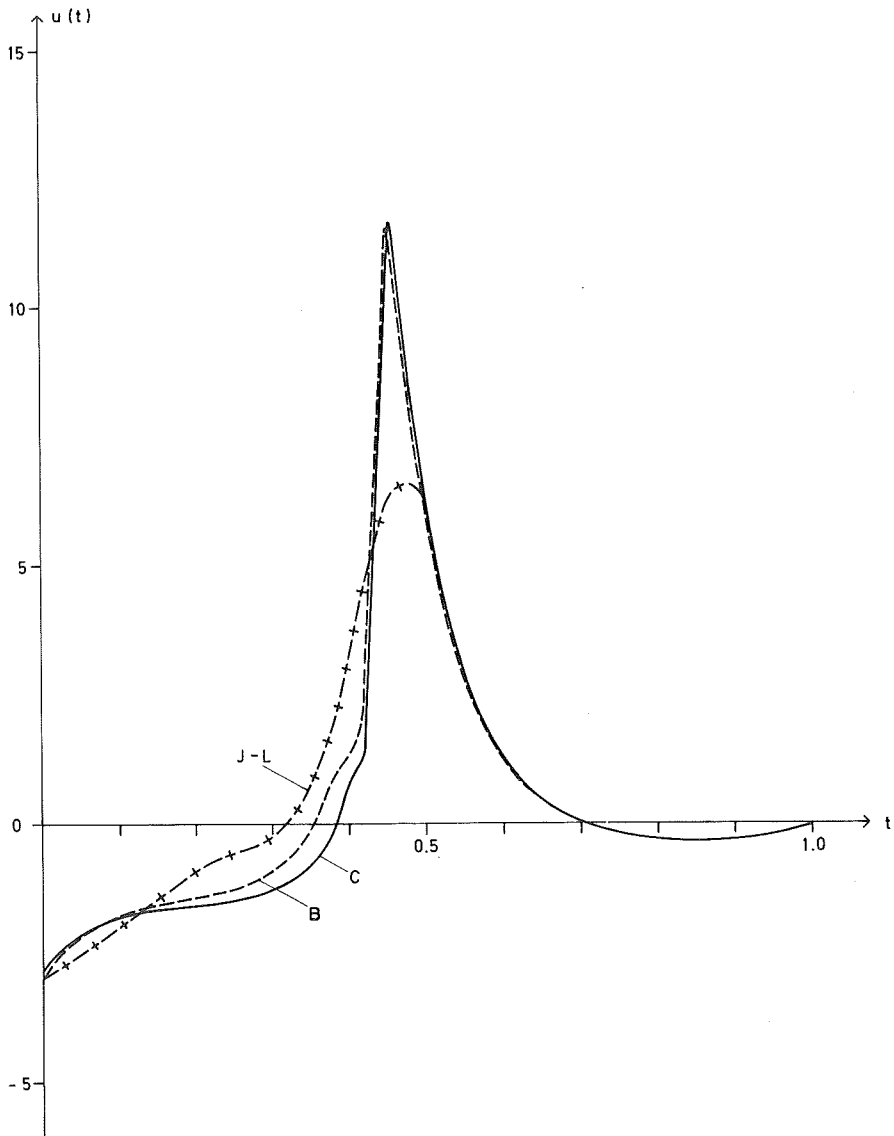


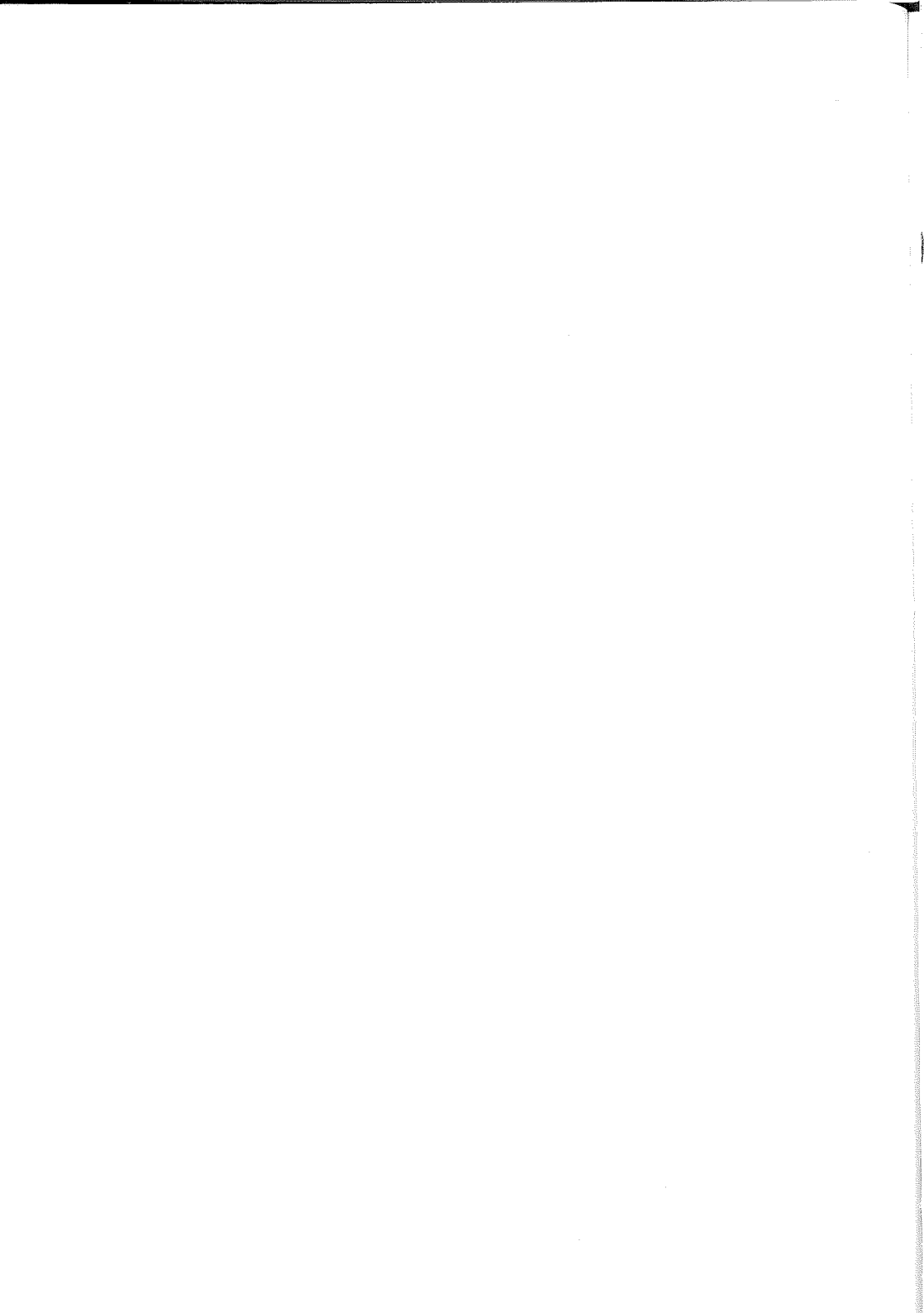
Fig 19. - Example 6.4: Second order constraint  $x_1(t) - 8(t - 0.5)^2 + 0.5 \leq 0$ . Computed  $u(t)$  for the hyperplanes B (---) and C (—), compared with slack variable technique solution (-+-).

## 7. REFERENCES

- [1] A. E. Bryson, W. F. Denham and S. E. Dreyfus, "Optimal Programming Problems with Inequality Constraints I: Necessary Conditions for Extremal Solutions", AIAA J., vol. 1, 1963, 2544-2550.
- [2] A. E. Bryson and Y. C. Ho, "Applied Optimal Control", Blaisdell, Waltham, 1969.
- [3] S. S. L. Chang, "Optimal Control in Bounded Phase Space", Automatica, vol. 1, 1962, 55-67.
- [4] W. F. Denham and A. E. Bryson, "Optimal Programming Problems with Inequality Constraints II: Solution by Steepest-Ascent", AIAA J., vol. 2, 1964, 25-34.
- [5] S. E. Dreyfus, "Dynamic Programming and the Calculus of Variations", Academic Press, New York and London, 1965.
- [6] P. Dyer and S. R. McReynolds, "The Computation and Theory of Optimal Control", Academic Press, New York and London, 1970.
- [7] S. D. Gershwin and D. H. Jacobson, "A Discrete-Time Differential Dynamic Programming Algorithm with Application to Optimal Orbit Transfer", Harvard Univ. Technical Report No. 566, Aug. 1968.
- [8] D. H. Jacobson, M. M. Lele and J. L. Speyer, "New Necessary Conditions of Optimality for Control Problems with State-Variable Inequality Constraints", J. Math. Anal. Appl., vol. 35, 1971, 255-284.
- [9] D. H. Jacobson and M. M. Lele, "A Transformation Technique for Optimal Control Problems with a State Variable Inequality Constraint", IEEE Trans. Automatic Control, vol. 14, 1969, 457-464.
- [10] D. H. Jacobson and D. Q. Mayne, "Differential Dynamic Programming", American Elsevier Publishing Comp., New York, 1970.

- [11] H. J. Kelley, "Methods of Gradients", in G. Leitman, Ed., "Optimization Techniques", Academic Press, New York and London, 1962.
- [12] L. S. Lasdon, S. K. Mitter and A. D. Waren, "The Conjugate Gradient Method for Optimal Control Problems", IEEE Trans. Automatic Control, vol. 12, 1967, 132-138.
- [13] L. S. Lasdon, A. D. Waren and R. K. Rice, "An Interior Penalty Method for Inequality Constrained Optimal Control Problems", IEEE Trans. Automatic Control, vol. 12, 1967, 388-395.
- [14] D. G. Luenberger, "Optimization by Vector Space Methods", J. Wiley and Sons, New York, 1969.
- [15] K. Mårtensson, "On the Matrix Riccati Equation", Inf. Sciences, vol. 3, 1971, 17-49. (Part 1 of this thesis)
- [16] K. Mårtensson, "A New Approach to Constrained Function Optimization", Research Report 7112, Division of Automatic Control, Lund Institute of Technology, March 1971. (Part 2 of this thesis)
- [17] K. Mårtensson, "Optimal Control of a Travelling Overhead Crane - a Feasibility Study", Research Report, Division of Automatic Control, Lund Institute of Technology, to appear. (Part 4 of this thesis)
- [18] K. Mårtensson, "Computational Methods for Optimal Control Problems", Research Report, Division of Automatic Control, Lund Institute of Technology, to appear.
- [19] J. E. Nagra, "Balance Function for the Optimal Control Problem", J. Optimization Theory and Applications, vol. 8, 1971, 35-48.
- [20] J. L. Speyer and A. E. Bryson, "Optimal Programming Problems with a Bounded State Space", AIAA J., vol. 6, 1968, 1488-1491.
- [21] F. A. Valentine, "The Problem of Lagrange with Differential Inequalities as Added Side Conditions", in Contributions to the Calculus of Variations, Chicago University Press, Chicago, 1937, 407-448.

- [22] W. E. Williamson and B. D. Tapley, "A Modified Perturbation Method for Solving Optimal Control Problems with State Variable Inequality Constraints", AIAA J., vol. 9, 1971, 2222-2228.



OPTIMAL CONTROL OF A TRAVELLING OVERHEAD  
CRANE - A FEASIBILITY STUDY

ABSTRACT

The possibility to use optimal control theory to design efficient control strategies for a travelling overhead crane is studied. Two different mathematical models corresponding to torque control and acceleration control are derived, and the computed minimum-time control strategies are presented. Optimal control in the presence of disturbances and the possibility to realize the optimal control strategies is considered. The study illustrates the current status of numerical methods for the solution of optimal control problems, and serves as a vehicle to discuss advantages and disadvantages of existing algorithms.

## 1. INTRODUCTION

The theory for optimal control of dynamic systems has developed very rapidly during the last ten years. A large number of important contributions have been published, and the theory has been considerably extended since the interest in optimal control arose in the late fifties. In parallel with the theoretical work, considerable progress has also been made in the development of efficient methods for numerical solution of the problems. Although there still is a considerable distance between the theoretical results and the computational methods, many powerful methods now exist for different kinds of problems. In particular, methods for problems where the system, or process, can be modelled by a set of ordinary differential equations are well established. However, apart from a large number of space vehicle problems, very few industrial applications have been reported. There are certainly many explanations to this gap between theory and practice, but a main reason may be the preparatory work required to develop computer programs to solve the stated problems. Thus even a simple feasibility study will in general be both expensive and time-consuming. Consequently, it is still rather difficult to form an opinion about the applicability of optimal control theory, and about the efficiency of existing computational methods.

In this part we will consider these aspects in a feasibility study of optimal control of a travelling overhead crane. The purpose of the paper is thus not only to present computed optimal strategies for a particular process, but also to discuss problems such as computer and computer program requirements, necessary preparatory work, and possibilities to realize the optimal control strategies. It is also believed that the paper illustrates the range of different problems that can be analysed with a pre-designed computer program package.

The problem considered originates from the container terminal illustrated in Fig 1. When the ship is unloaded, the containers are first transferred by a quayside crane to a waiting lorry. The lorry then drives to a storage house or an open storage area, where another crane stacks the containers in predetermined positions. The empty lorry then closes the loading cycle by returning to the quayside crane. The bottle-necks of this cycle are the times required for the cranes to transfer and exactly position the containers. Minimizing these transfer times would reduce the residence time of the ship, and would result in large economic profits.

Since the two cranes have rather identical properties, we will concentrate on the stack crane, and for this we will consider a typical transfer. This problem is stated in Section 2, where also the different kinds of restric-



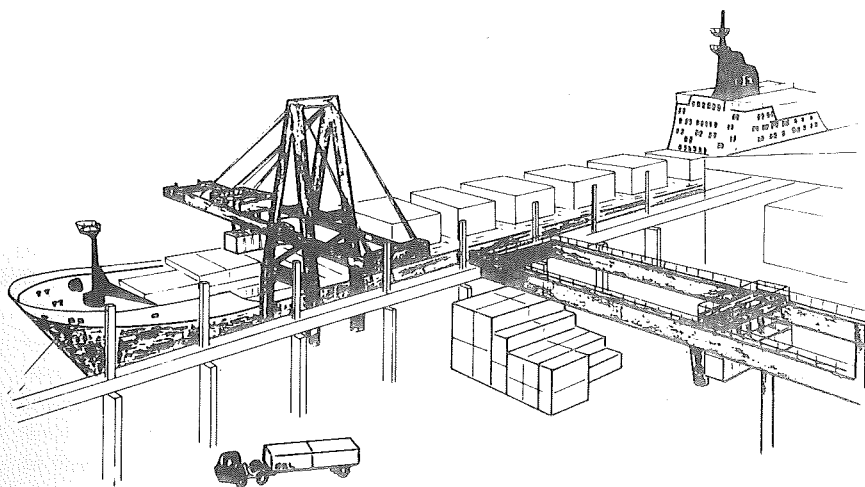


Fig 1. - Outline of the container terminal.

tions and constraints that may appear are defined.

In Section 3 we will derive two different sets of ordinary differential equations describing the operation of the crane. In Model 1 we assume that the available control variables are the accelerations of the trolley and the winch. This corresponds to a crane designed for manual operation, in which case the crane is equipped with different regulators to make the operation independent of the mass of the load. The second model (Model 2) is derived under the assumption that the control variables are the torques of the trolley driving motors and the winching driving motors. This model will depend on the masses of the load and the trolley, and thus is impossible to operate manually in an effective way. However, this model is believed to be of interest, because if a computer is used to realize the optimal strategies, it is possible to save a lot of the conventional control equipment.

In Section 4 we consider optimal control of Model 1. The problem stated in Section 2 is formulated in mathematical terms, the control and state variable constraints are specified, and the possibility to choose different cost functionals to simplify the numerical computations are discussed. We also motivate the selection of method for numerical solution. Approximate minimum time control strategies are presented for different stack profiles, and some experiences from the numerical computation of these are accounted for. The implementation of optimal strategies in the pre-

sence of disturbances is also treated, and a possible method is indicated.

Optimal control of Model 2 is considered in Section 5. The analysis is restricted to minimum time control strategies and it is shown that the structure of the solutions are similar to those of Model 1.

It should be emphasized that all the numerical solutions have been computed with a general-purpose programming package. This possibility to reduce the preparatory programming work to a minimum is considered as an important progress to make optimal control theory available as a standard tool for analysis and synthesis.

## 2. STATEMENT OF THE PROBLEM

Since the properties of the quayside crane and the stack crane are similar, the feasibility study is restricted to the stack crane. The particular transfer considered is illustrated in Fig 2. When the lorry arrives, the gantry is is

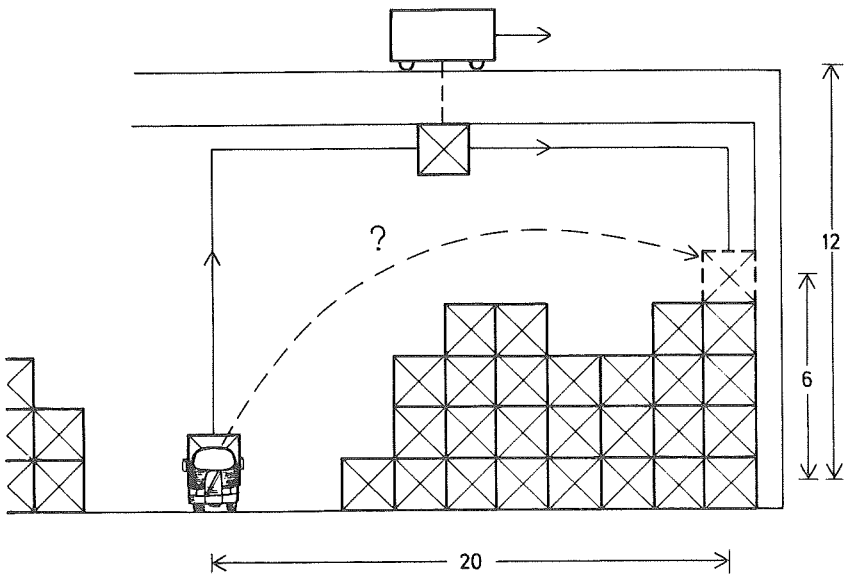


Fig 2. - The particular transfer considered. All distances are in meters.

assumed to be positioned so that the final transfer to the stack can be performed by the trolley and the winch. Thus the problem is reduced to two dimensions. It is assumed that the final position is 6 meters above and 20 meters away from the lorry, and that the distance between the load and the trolley is initially 12 meters. Further, it is assumed that the containers can be considered as point masses, and that the centre of mass is equal to the geometrical centre. All distances then refer to this point.

When the transfer begins, the trolley is positioned straight over the lorry, and has no velocity in either direction. Besides, the vertical velocity of the load is zero. At the end of the transfer, the system is similarly required to be in complete rest. In general this requires a skilful operator when the crane is controlled manually, since the oscillations of the load must be completely damped out. One possibility is then to operate the crane according to the full curve of Fig 2. However, it is easily seen that this strategy may be very time-consuming, and that it may be favourable to be able to operate the crane as indicated by the dashed line.

As stated in Section 1, the available control variables will be different for the two types of cranes studied. In Model 1 we assume that the control variables are the accelerations of the trolley and the winch, while in Model 2 it is assumed that the control variables are the torques of the driving motors. In both cases the trolley and the winch may be operated separately. The control variables as well as the trolley and winching velocities are bounded in both cases. This is due to the limited effect of the driving motors, and these constraints will in general depend on the characteristics of the electric motors. In this study we have for simplicity assumed that the constraints are simple magnitude limits. Finally, it may be necessary to take the actual profile of the stack into account. In this paper we have restricted the study to the particular situation indicated in Fig 2. However, the solution technique used is applicable to rather arbitrary stack profiles.

The storage area is in general an out-door area, and the transfer may be subject to heavy wind disturbances. The control strategy should thus also be able to compensate for these.

We may then summarize the problem as follows: Determine a control strategy for the trolley and the winch, so that the load is transferred to the predetermined final position as fast as possible. The control strategy must take all existing constraints into account, and the possibility to compensate for disturbances should be considered.

### 3. MATHEMATICAL MODELS OF THE CRANE

Different simplified models of travelling overhead cranes have been published. In [4] and [8], a crane with constant load cable length was modelled as a fourth order dynamic system with a single control variable. The control variable was assumed to be the acceleration of the trolley [4], respectively the torque of the trolley driving motors [8]. In [8] it was also assumed that the angular deviations of the load are small enough to allow for linearization, and thus the crane could be described with a simple linear model. A slightly more general model was considered in [1], where the control variables were chosen as the trolley acceleration and the winching velocity. However, the generalization was restricted to constant winching velocities. Similar to [8], the model was simplified by the assumption that the angular deviations are small. The crane could then be described as a fifth order dynamic system with two control variables, one of which is constant over the time interval considered.

The assumption that the angular deviations are small was necessary for the solution technique used in [4] and [8] to be applicable. Also, for the goods handling problems considered in [4] and [8], e.g. coal and ore loading, the accuracy is not critical, and thus the approximations may be well justified. However, for the container transfer problem considered in this paper, the accuracy of the terminal state is essential, and simulations have clearly indicated that approximations based on the assumption of small angular deviations are not accurate enough. It can further be noticed that the winching strategies of [4] and [8] are very special, and the full capacity of the crane is not considered.

In this section we will thus derive more general models, and it will be assumed that the trolley and the winch may be operated arbitrarily and separately. (Restrictions on the operation due to different constraints will be specified in Section 4 in connection with the mathematical formulation of the optimization problem.) In the first model, referred to as Model 1, we assume that the available control variables are the accelerations of the trolley and the winch. This corresponds to a crane designed for manual operation. In the second model (Model 2), the control variables are related to the torques of the electric driving motors of the trolley and the winch. The dynamic properties of the crane will then depend on the mass of the load, and the oscillations of the load will influence the motion of the trolley. In both models all frictional forces are neglected. Although these may be very important to include in practice for accuracy reasons, they have no essential influence on the basic dynamic proper-

ties of the models.<sup>+</sup>

To describe the system, a coordinate system  $(\xi, \eta)$  with the  $\xi$ -axis along the gantry according to Fig 3 is introduced.

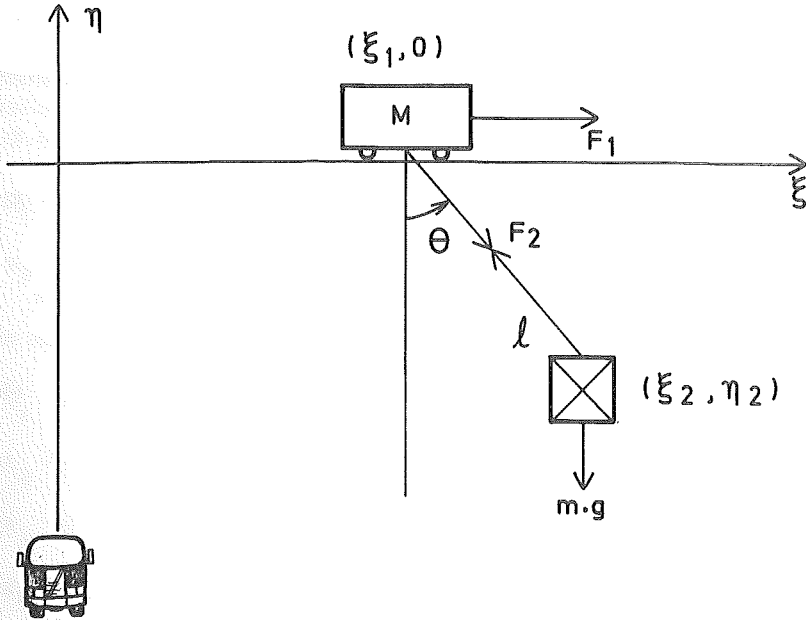


Fig 3. - Simplified model of the crane.

The following notations are used:

$M$  - mass of the trolley.

$m$  - mass of the load.

$(\xi_1, 0)$  - position of the trolley.

$(\xi_2, \eta_2)$  - position of the load.

<sup>+</sup> Notice that with the assumption that the angular deviations are not small enough to allow for linearization, the complexity of the problem is already raised above the level where simple solution methods might apply. The inclusion of the frictional forces would thus not further contribute to this complexity.

- $F_1$  - the force acting on the trolley due to the trolley driving motors.  
 $F_2$  - tension in the cable. This is directly proportional to the torque of the driving motors of the winch.  
 $\theta$  - angular deviation.  
 $l$  - length of the cable.  
 $g$  - gravity acceleration ( $9.81 \text{ m/s}^2$ ).

It is assumed that the load can be considered as an idealized pendulum with variable pendulum length. From Fig 3 then follows.

$$\xi_2 = \xi_1 + l \cdot \sin \theta \quad (3.1)$$

$$\eta_2 = - l \cdot \cos \theta$$

and a straightforward application of Newton's law to the load yields

$$m\ddot{\xi}_2 = - F_2 \cdot \sin \theta \quad (3.2)$$

$$m\ddot{\eta}_2 = F_2 \cdot \cos \theta - mg$$

Elimination of  $F_2$  and  $m$  in (3.2) then results in

$$\ddot{\xi}_2 \cdot \cos \theta + \ddot{\eta}_2 \cdot \sin \theta = - g \cdot \sin \theta \quad (3.3)$$

$\ddot{\xi}_2$  and  $\ddot{\eta}_2$  are obtained through differentiation of (3.1) with respect to time, that is,

$$\ddot{\xi}_2 = \ddot{\xi}_1 + \ddot{l} \cdot \sin \theta + 2 \dot{l} \dot{\theta} \cdot \cos \theta - l \dot{\theta}^2 \cdot \sin \theta + l \ddot{\theta} \cdot \cos \theta \quad (3.4)$$

$$\ddot{\eta}_2 = - \ddot{l} \cdot \cos \theta + 2 \dot{l} \dot{\theta} \cdot \sin \theta + l \dot{\theta}^2 \cdot \cos \theta + l \ddot{\theta} \cdot \sin \theta$$

Substituting (3.4) into (3.3), we get

$$\ddot{\xi}_1 \cdot \cos \theta + 2 \dot{l} \dot{\theta} + l \ddot{\theta} = - g \cdot \sin \theta$$

or

$$\ddot{\theta} = - \frac{g \cdot \sin \theta}{l} - \frac{2 \dot{l} \dot{\theta}}{l} - \frac{\ddot{\xi}_1 \cdot \cos \theta}{l} \quad (3.5)$$

Notice that (3.5) is independent of the choice of control variables, and (3.5) is the fundamental relation used in both models.

### 3.1. Model 1 - Acceleration control

Assuming that the accelerations of the trolley and the cable length are the available control variables, a model of the crane is easily derived from (3.5). We then introduce the control variables  $u_1$  and  $u_2$  as

$$\begin{aligned} u_1 &= \ddot{\xi}_1 \\ u_2 &= \ddot{l} \end{aligned} \tag{3.6}$$

and the state variables  $x_1, \dots, x_6$  as

$$\begin{aligned} x_1 &= \xi_1 \\ x_2 &= \dot{\xi}_1 \\ x_3 &= \theta \\ x_4 &= \dot{\theta} \\ x_5 &= l \\ x_6 &= \dot{l} \end{aligned} \tag{3.7}$$

Then by definition and from (3.5) we get

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u_1 \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= -\frac{g \cdot \sin x_3}{x_5} - \frac{2 x_4 x_6}{x_5} - \frac{u_1 \cdot \cos x_3}{x_5} \\ \dot{x}_5 &= x_6 \\ \dot{x}_6 &= u_2 \end{aligned} \tag{Model 1}$$

For transfers from the lorry to the stack, the initial state is

$$\begin{aligned} x_1(0) = x_2(0) = x_3(0) = x_4(0) = x_6(0) &= 0 \\ x_5(0) &= 12 \end{aligned} \quad (3.8)$$

From the model it can be seen that the position of the trolley and the load cable length are simple to control, since these parts of the model contain only pure integrations. However, the oscillations of the load are governed by a highly nonlinear equation, and a simple analytical solution of the problem stated in Section 2 is obviously excluded.

### 3.2 Model 2 - Torque control

In this case we will assume that the dynamics and the inertias of the cable drum and the driving motors can be neglected. As pointed out above, such factors may be necessary to consider in a practical application, but they will not change the basic properties of the model. Thus the forces  $F_1$  and  $F_2$  (Fig 3) can be considered as directly proportional to the torques of the driving motors.

Compared with Model 1, we have to express  $\ddot{\xi}_1$  and  $\ddot{\ell}$  in terms of the state variables and the forces  $F_1$  and  $F_2$ . We then apply Newton's law on the trolley too, that is,

$$M\ddot{\xi}_1 = F_1 + F_2 \cdot \sin \theta$$

or

$$\ddot{\xi}_1 = \frac{F_1}{M} + \frac{F_2}{M} \cdot \sin \theta \quad (3.9)$$

Substituting (3.9) into (3.5) yields

$$\ddot{\theta} = -\frac{g \cdot \sin \theta}{\ell} - \frac{2 \dot{\ell} \dot{\theta}}{\ell} - \frac{F_1 \cdot \cos \theta}{M\ell} - \frac{F_2 \cdot \cos \theta \cdot \sin \theta}{M\ell} \quad (3.10)$$

Now choose the same state variables (3.7) as for Model 1. From (3.9), (3.10), and by definition the time derivatives  $\dot{x}_1, \dots, \dot{x}_5$  are then determined. To determine  $\dot{x}_6 = \ddot{\ell}$ , consider the equations (3.2) and



(3.4), that is,

$$m\ddot{\eta}_2 = F_2 \cdot \cos \theta - mg$$

and

$$\ddot{\eta}_2 = -\ddot{l} \cdot \cos \theta + 2 \dot{l} \dot{\theta} \cdot \sin \theta + l \dot{\theta}^2 \cdot \cos \theta + l \ddot{\theta} \cdot \sin \theta$$

Substitute  $\ddot{\eta}_2$  into the first relation. Then

$$\frac{F_2 \cdot \cos \theta}{m} - g = -\ddot{l} \cdot \cos \theta + 2 \dot{l} \dot{\theta} \cdot \sin \theta + l \dot{\theta}^2 \cdot \cos \theta + l \ddot{\theta} \cdot \sin \theta \quad (3.11)$$

or

$$\ddot{l} = -\frac{F_2}{m} + \frac{g}{\cos \theta} + \frac{2 \dot{l} \dot{\theta} \cdot \sin \theta}{\cos \theta} + \frac{l \ddot{\theta} \cdot \sin \theta}{\cos \theta} + l \dot{\theta}^2 \quad (3.12)$$

Finally, we substitute  $\ddot{\theta}$  (3.10) into (3.12), and then

$$\ddot{l} = g \cdot \cos \theta - \frac{F_2}{m} - \frac{F_1}{m} \cdot \sin \theta - \frac{F_2}{M} \cdot \sin^2 \theta + l \dot{\theta}^2 \quad (3.13)$$

Thus the model obviously will depend on the masses of the trolley and the load. To simplify the notations, we then choose the control variables

$$u_1 = \frac{F_1}{M} \quad (3.14)$$

$$u_2 = \frac{F_2}{m}$$

and also introduce

$$\delta = \frac{m}{M} \quad (3.15)$$

From (3.9), (3.10), (3.13) - (3.15), and by definition, we then get the following model.

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

$$\dot{x}_2 = u_1 + u_2 \delta \cdot \sin x_3$$

$$\dot{x}_3 = x_4$$

(Model 2)

$$\dot{x}_4 = -\frac{g \cdot \sin x_3}{x_5} - \frac{2x_4 x_6}{x_5} - \frac{u_1 \cdot \cos x_3}{x_5} - \frac{u_2 \delta \cdot \cos x_3 \cdot \sin x_3}{x_5}$$

$$\dot{x}_5 = x_6$$

$$\dot{x}_6 = g \cdot \cos x_3 + x_5 x_4^2 - u_1 \cdot \sin x_3 - u_2 \delta \cdot \sin^2 x_3 - u_2$$

The initial condition is the same as for Model 1. Notice that the choice of control variables is rather arbitrary. However, with the particular choice made here (3.14), the control variables can be interpreted as accelerations, and it will be easier to compare the properties of the two models.

Also notice that in Model 2 the coupling between the load and the trolley is mutual, and the oscillations of the load will influence the velocity of the trolley. The strength in this coupling is determined by the magnitude of  $u_2$  and the quotient  $\delta$  between the mass of the load and the mass of the trolley.

#### 4. OPTIMAL CONTROL OF MODEL 1

In this section, the problem stated in Section 2 will be formulated in terms of an optimal control problem for Model 1. We will thus specify the control and state variable constraints, and different possibilities to choose the cost functional will be considered. The choice of numerical method to solve the problem is motivated, and a brief account for the necessary programming work is given. Further, we will present the computed minimum time control strategies both for a situation where the stack profile must be taken into account, and for a case where it can be neglected. Finally, the possibilities to handle disturbances and to implement the optimal strategies are discussed. For the different cases considered, we will give rough estimates of the computer requirements

and the execution times. All computations have been performed on a UNIVAC 1108.<sup>+</sup>

#### 4.1 Control and state variable constraints

In section 3 the following model (Model 1) was derived under the assumption that the available control variables are the accelerations of the trolley and the cable length:

$$\frac{dx_1}{dt} = x_2 \quad x_1(0) = 0$$

$$\frac{dx_2}{dt} = u_1 \quad x_2(0) = 0$$

$$\frac{dx_3}{dt} = x_4 \quad x_3(0) = 0$$

$$\frac{dx_4}{dt} = -\frac{g \cdot \sin x_3}{x_5} - \frac{2x_4x_6}{x_5} - \frac{u_1 \cdot \cos x_3}{x_5} \quad x_4(0) = 0$$

$$\frac{dx_5}{dt} = x_6 \quad x_5(0) = 12$$

$$\frac{dx_6}{dt} = u_2 \quad x_6(0) = 0$$

We also recall that the state variables are:  $x_1 = \xi_1$ ,  $x_2 = \dot{\xi}_1$ ,  $x_3 = \theta$ ,  $x_4 = \dot{\theta}$ ,  $x_5 = l$ ,  $x_6 = \dot{l}$ . At the end of the transfer, the container must

<sup>+</sup> For comparison some representative figures are:  
floating addition 2.6  $\mu$ s, multiplication 3.4  $\mu$ s and division 9.0  $\mu$ s.

be in the correct position in the stack, and the system is required to be in complete rest. These conditions at the terminal time  $t_f$ , are specified in the following terminal constraint vector  $\psi$ :

$$\psi(x(t_f); t_f) = \begin{pmatrix} x_1(t_f) - 20 \\ x_2(t_f) \\ x_3(t_f) \\ x_4(t_f) \\ x_5(t_f) - 6 \\ x_6(t_f) \end{pmatrix}$$

The requirement on the terminal state thus is

$$\psi(x(t_f); t_f) = 0$$

Since the torque of the electrical driving motors is limited, it is natural to assume that the magnitude of the accelerations and retardations are bounded. In this case it is assumed that

$$-0.46 \leq u_1 \leq 0.46$$

and

$$-0.61 \leq u_2 \leq 0.61$$

where the unit is  $\text{m/s}^2$ . The constraints may be summarized in vector notations as

$$h(u; t) = \begin{pmatrix} u_1 - 0.46 \\ -u_1 - 0.46 \\ u_2 - 0.61 \\ -u_2 - 0.61 \end{pmatrix} \leq 0 \quad (4.1)$$

Notice that the constraints (4.1) must be satisfied for all  $t$  such that  $0 \leq t \leq t_f$ .

Due to the limited working area of the regulators which make it possible

to choose the accelerations as control variables, the velocities of the trolley and the winch are bounded. It is thus assumed that

$$-1.4 \leq x_2 \leq 1.4$$

and

$$-0.5 \leq x_6 \leq 0.5$$

(The unit is m/s.) Similar to the control variable constraints, these state variable constraints are summarized in vector notations as

$$S(x;t) = \begin{pmatrix} x_2 - 1.4 \\ -x_2 - 1.4 \\ x_6 - 0.5 \\ -x_6 - 0.5 \end{pmatrix} \leq 0 \quad (4.2)$$

where the inequalities must be satisfied for all  $t$  such that  $0 \leq t \leq t_f$ .

These state variable constraints complicate the problem, since they are not possible to directly include in any existing computational method. To handle these constraints the constraining hyperplane technique was developed [12]. The basic idea in this technique is to convert (4.2) into mixed state-control variable constraints  $g(x, u; t) \leq 0$ , where the control variables appear explicitly. The transformed constraints can then be handled in the same way as the pure control variable constraints  $h(u; t) \leq 0$ . The transformation technique is applied to each component  $S_i$  of the constraints  $S \leq 0$ . If  $S_i$  is of order  $q$ , that is, the  $q$ -th total time derivate of  $S_i$  is the first that explicitly contains the control variables,  $S_i(x; t) \leq 0$  is approximated with the mixed state-control variable constraint

$$\frac{d^q S_i}{dt^q}(x, u; t) + a_1 \frac{d^{q-1} S_i}{dt^{q-1}}(x; t) + \dots + a_q S_i(x; t) \leq 0 \quad (4.3)$$

It is shown in [12] that the half-space generated by (4.3) in the

$\left( \frac{d^q S_i}{dt^q}, \dots, S_i \right)$ -space, tends to  $S_i \leq 0$  as the zeroes of the polynomial

$$p^q + a_1 p^{q-1} + \dots + a_q = 0 \quad (4.4)$$

tend to  $-\infty$ . In practice the choice of the parameters  $a_i$  depends on the time scale of the problem. However, this transformation technique has proved to yield very accurate solutions even for comparatively modest values of the zeroes of (4.4).

We thus directly convert all components of (4.2) with this technique. Since all  $S_i$ 's are of first order, the constraining hyperplanes are

$$\begin{aligned} u_1 + a_1 (x_2 - 1.4) &\leq 0 \\ -u_1 + b_1 (-x_2 - 1.4) &\leq 0 \\ u_2 + c_1 (x_6 - 0.5) &\leq 0 \\ -u_2 + d_1 (-x_6 - 0.5) &\leq 0 \end{aligned} \quad (4.5)$$

where a suitable choice of the parameters  $a_1, \dots, d_1$  in this case has been found to be 10.

Summarizing the control variable constraints (4.1) and the constraining hyperplanes (4.5) in a common constraint vector  $g(x, u; t)$ , we thus have

$$g(x, u; t) = \begin{pmatrix} u_1 - 0.46 \\ -u_1 - 0.46 \\ u_2 - 0.61 \\ -u_2 - 0.61 \\ u_1 + 10(x_2 - 1.4) \\ -u_1 + 10(-x_2 - 1.4) \\ u_2 + 10(x_6 - 0.5) \\ -u_2 + 10(-x_6 - 0.5) \end{pmatrix} \leq 0 \quad (4.6)$$

and these inequalities should then be satisfied for all  $t$  such that  $0 \leq t \leq t_f$ .

Notice that we have not yet included constraints due to a particular stack profile. This problem is treated separately in Section 4.5.

## 4.2 Choice of cost functional

Since we are mainly interested in minimum time transfers, one possibility to formulate the optimal control problem is as follows:

Determine a control strategy  $u(t)$  for Model 1, so that the cost functional

$$J = \int_0^{t_f} 1 \, dt$$

is minimized subject to the constraints

$$\psi(x(t_f); t_f) = 0$$

$$g(x, u; t) \leq 0$$

It is thus assumed that  $t_f$  is given implicitly. From the Maximum Principle [15] then follows that the optimal control strategy will either be a bang-bang solution in both  $u_1$  and  $u_2$ , or a combination of bang-bang arcs and singular subarcs.

However, there are several reasons why we have not chosen this approach. Firstly, there are no efficient computational methods developed for bang-bang or singular problems of this complexity. Secondly, the possibility to handle disturbances must be taken into account. For a large class of problems with continuous optimal control solutions this may be achieved by a local feedback around the optimal solution. However, for bang-bang problems similar possibilities do not exist. Different approaches to this problem have been presented, but the methods are either not applicable in real time since they do not produce causal feedbacks [5] [9], or will not satisfy the terminal constraints accurate enough [6] [7].

An alternative possibility then is to formulate the problem as a problem with explicitly given terminal time. We must then choose the terminal time  $t_f$  so that it is possible to reach the terminal state ( $\psi = 0$ ) with a

control strategy satisfying  $g(x, u; t) \leq 0$ . With  $t_f$  explicitly given, it is then natural to choose the cost functional with regard to the desired properties of the optimal solution. However, at this stage it is also important to keep in mind that a numerical solution shall be computed. Thus it is obviously favourable if the cost functional can be chosen also with regard to the properties of the numerical method. In this case we have as a first attempt chosen

$$J = \int_0^{t_f} \{u_1^2 + u_2^2\} dt \quad (4.7)$$

Physically this implies that we look for the minimum energy control strategy, and further this choice of  $J$  implies that bang-bang solutions and singular solutions are excluded.

A possible way to compute approximate minimum time strategies is now to successively decrease the terminal time until it becomes impossible to satisfy all the terminal constraints. This technique has been used for both models, and although it has proved to be difficult to determine the minimum time strategies with very high accuracy, this approximation technique will prove to have many advantages. For example, it will clearly illustrate the sensitivity of the control strategy structure with respect to the terminal time.

Although the cost functional (4.7) is chosen to exclude singular and bang-bang solutions, it will be shown in Section 4.4 that (4.7) must be further modified to simplify the numerical computations. However, (4.7) constitutes the basic cost functional, and the necessary modifications will be discussed in connection with the presentation of the computed optimal solutions in Section 4.4.

### 4.3 The computer program

For the regular problem formulated in the preceding section, there are several different numerical methods available, e.g. successive-sweep methods [3], conjugate gradient methods [11] and methods based on Differential Dynamic Programming [9]. The ultimate choice between these methods is not uniquely determined, but is merely a weighing between the preparatory programming work and the efficiency of the algorithm. For example, the conjugate gradient method requires rela-



tively little programming work, but the algorithm has poor convergence properties close to the optimal solution. We have chosen a second order method based on Differential Dynamic Programming, and although this method initially requires a great deal of preparatory programming work, we think that this is repaid by the efficiency and flexibility of this algorithm compared with the other methods mentioned above. In contrast with first order methods, the second order method chosen gives a local linear feedback around the optimal solution, and this feedback may sometimes be used to handle the disturbances on the process.

The derivation and a detailed description of the algorithm is given in [12] (part 3 of this thesis). To avoid reiterations, all necessary references will thus be made to this part. We will then use the notation  $(3.x.x)$  when referring to the relation  $(x.x)$  in part 3.

However, the basic structure of the computer program deserves a commentary. The program consists of a main program which is independent of the particular problem considered, and a subroutine where the particular optimal control problem is specified. The necessary preparatory programming work is thus restricted to the subroutine. In this the system equations, the cost functional, the terminal constraints, the control variable constraints, the Hamiltonian, the partial derivatives and the different parameters are specified. Besides, the minimization of the Hamiltonian is carried out in this subroutine. To our knowledge, this is the first attempt to design a program package for numerical solution of a wide class of optimal control problems. The package has proved applicable to a wide range of problems, and will be described in detail in [13]. However, it should be noted, that since the main program is designed for general problems, the computer memory requirements are rather large ( $\sim 30$  k). Besides, the execution time may be considerably larger than for a program where all the structure of the particular problem considered is taken into account. These facts should be considered when the possibility to compute optimal solutions with a small computer is estimated on the basis of computations on a big computer.

#### 4.4 Minimum time control strategies

In this section we will illustrate the possibility to compute approximate minimum time control strategies for the case where the stack profile can be neglected. We will then consider the reformulated fixed terminal

time problem, and successively decrease  $t_f$  until the terminal constraints can not be satisfied.

In principle, the cost functional (4.7) makes the problem well-defined, and for different values of  $t_f$  in the range of interest, it can be verified that an optimal solution satisfying all constraints exist. However, the best way to include the terminal constraints  $\psi$  in the problem is in general to adjoin  $\psi$  by means of a Lagrange multiplier vector  $b$  to the cost functional. This approach is used in the computer program described above, and thus the actual cost considered would be

$$\bar{J} = b^T \psi + \int_0^{t_f} \{u_1^2 + u_2^2\} dt \tag{4.8}$$

The multipliers  $b$  are then determined so that the extremal solution of (4.8) satisfies  $\psi = 0$ . In optimal control literature it is always tacitly assumed that this extremal constitutes a minimum of (4.8). However, comparing with finite-dimensional optimization, this is obviously a rather strong assumption, and it was in this case found that the extremal is not a minimum of (4.8). Consequently, no minimum solution of (4.8) could be determined, and in particular it was found that the Riccati equation (3.5.39)

$$\begin{aligned} -\frac{dV_{xx}}{dt} = & H_{xx} + \mu g_{xx} + f_x^T V_{xx} + V_{xx} f_x + \beta_1^T (H_{uu} + \mu g_{uu}) \beta_1 + \\ & + \beta_1^T (H_{ux} + \mu g_{ux} + f_u^T V_{xx}) + (H_{ux} + \mu g_{ux} + f_u^T V_{xx})^T \beta_1 \end{aligned} \tag{4.9}$$

got unbounded solutions. This problem was overcome with the technique described in [14] (part 2 of this thesis) for finite-dimensional problems. Through inclusion of the quadratic term  $c\psi^T\psi$  in the cost functional, it thus became possible to convert the extremal of (4.8) into a minimum of

$$\bar{J} = c\psi^T\psi + b^T\psi + \int_0^{t_f} \{u_1^2 + u_2^2\} dt \tag{4.10}$$

where  $c = 2$  was found to be sufficient.

With the cost functional (4.10), the problem was solved for a decreasing

sequence of terminal times  $t_f$ , with  $t_f$  initially equal to 25.0 secs., and the minimum time was determined to  $t_f = 18.12$  secs. (For  $t_f = 18.10$  the algorithm failed to satisfy the terminal constraints within the specified numerical accuracy,  $\|\psi\| \leq 0.01$ ). The computed control strategies and the movements of the trolley and the load are shown in Fig 4. It can be seen that the control strategies are very close to bang-zero-bang solutions, and that  $u_1$  exhibits a rather unexpected behaviour with three retardation-acceleration phases for  $t \approx 4, 10$  and 15.

Since there is no analytic solution available, the exact minimum time can only be roughly estimated. For example, the time required to bring the trolley to the final position (not considering the state of the load) is easily shown to be about 17.4 secs. Consequently the additional time required when the oscillations and the position of the load are taken into account is rather small. The computed minimum time should also be compared with the full curve strategy of Fig 2. This transfer takes about 55.0 secs. and thus the minimum time strategy implies a reduction with a factor three.

The corresponding state variables  $x_2$  (trolley velocity),  $x_3$  (angular deviation) and  $x_6$  (winching velocity) are shown in Fig 5. The behaviour of the trolley velocity  $x_2$  well illustrates the advantage of the constraining hyperplane technique compared with the interior point constraint method suggested by Bryson et al. [2]. Rather unexpectedly, the trolley reaches the velocity constraint twice. With the solution technique suggested in [2], it would then be necessary to a priori know this property of the optimal solution, and also to have sufficiently good estimates of the entry times. In the constraining hyperplane technique, the structure as well as the entry times fall out automatically.

There are several possibilities to speed up the iteration procedure on the unknown optimal terminal time  $t_f$ . Firstly, the structure of the control variables is rather informative. This is illustrated in Fig 6, where the computed optimal control variables are shown for  $t_f = 18.2, 18.6$  and 19.0. It can be seen that the bang-bang structure is very rapidly lost, and the control variable  $u_1(t)$  becomes rather smooth already for  $t_f = 18.2$ . Very little in time is thus gained at the price of a complex control signal. Secondly, the eigenvalues of  $V_{bb}$  (3.5.39) constitute a useful measure of how much  $t_f$  may be decreased. When  $t_f$  approaches the minimum time, one or more eigenvalues of  $V_{bb}$  tend to zero, and the corresponding components of the multiplier vector  $b$  tend to plus or minus infinity (3.5.47).

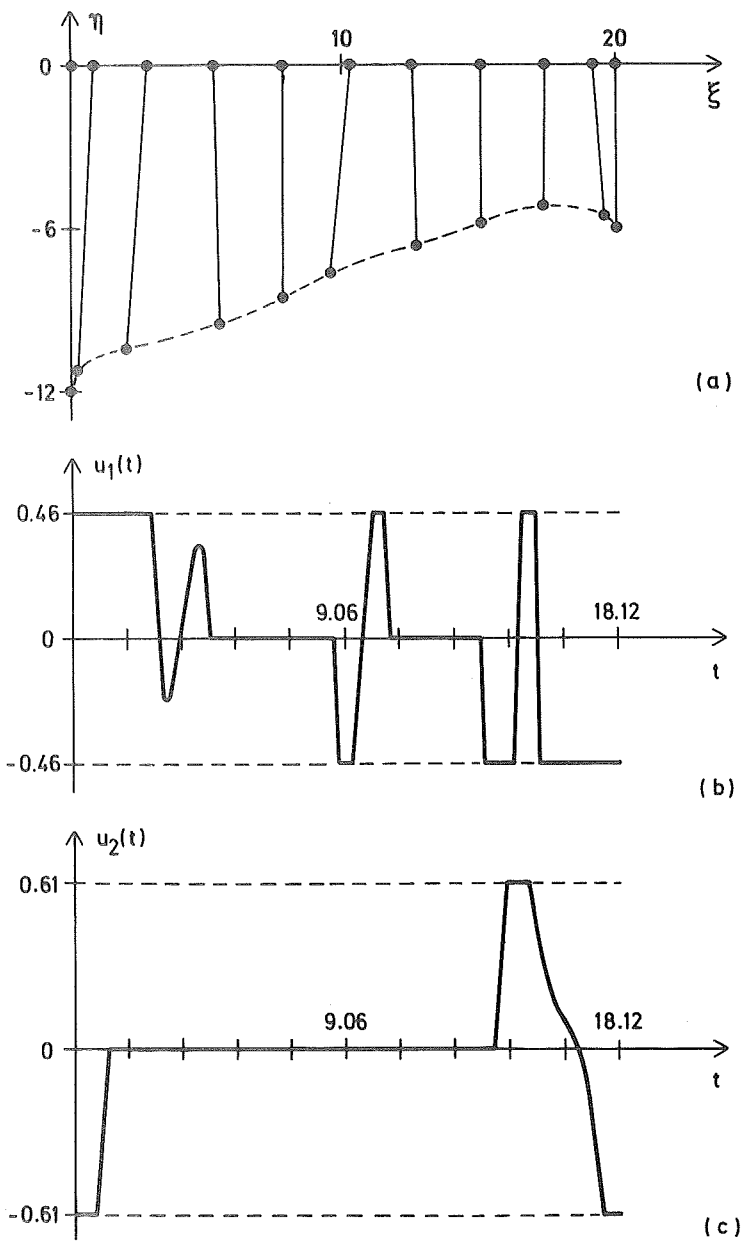


Fig 4. - Computed optimal solution for  $t_f = 18.12$  secs.  
 a) - movements of the trolley and the load.  
 The distance between two successive points is 1.812 secs.  
 b) - acceleration of the trolley ( $u_1(t)$  ).  
 c) - acceleration of the winch ( $u_2(t)$  ).

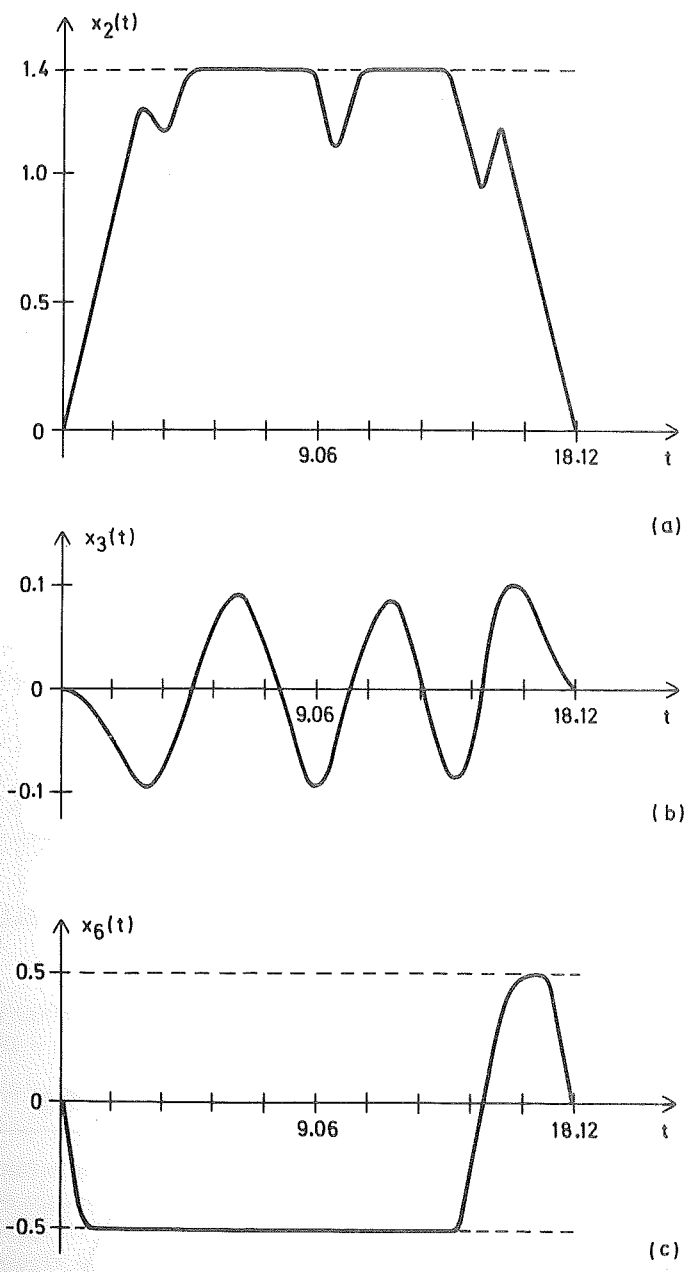


Fig 5. - Computed optimal solution for  $t_f = 18.12$  secs.  
 a) - trolley velocity ( $x_2(t)$  ).  
 b) - angular deviation ( $x_3(t)$  ).  
 c) - winching velocity ( $x_6(t)$  ).

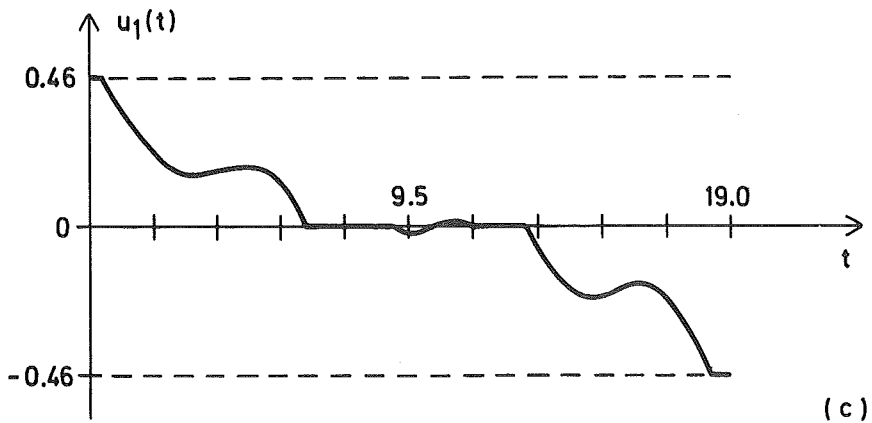
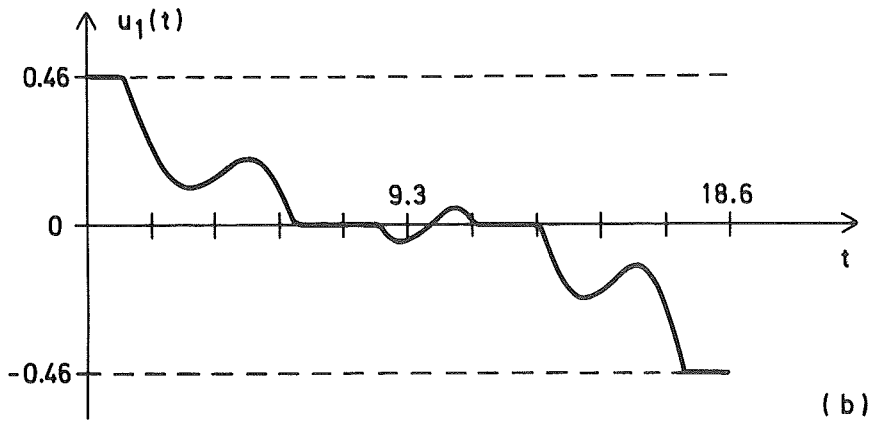
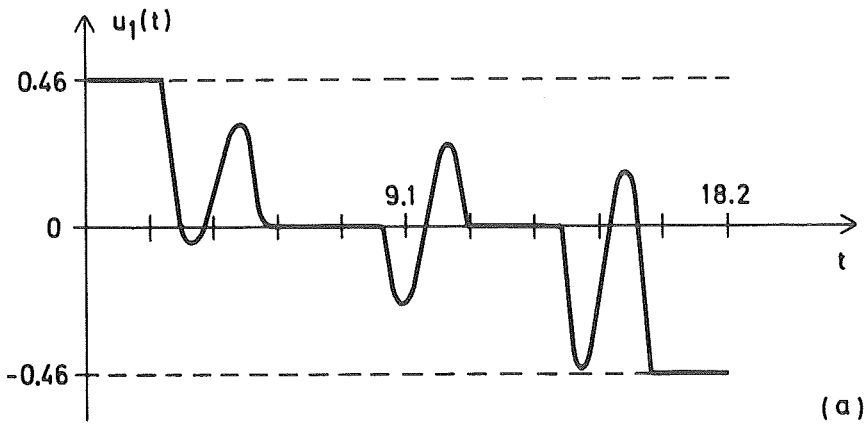


Fig 6. - Computed optimal control variables  $u_1(t)$  for  
 a)  $t_f = 18.2$ , b)  $t_f = 18.6$  and c)  $t_f = 19.0$ .

Compared with other numerical methods, the Differential Dynamic Programming algorithm used here has a characteristic quality, namely the minimization of the Hamiltonian  $H$  in the backward loop (see part 3). For some problems, this may cause much trouble, and it may be necessary to use a separate subroutine for this finite-dimensional minimization. It may thus be worth while to consider this problem already when the cost functional is formulated. In this case the admissible region  $\Omega$  in the control space will always constitute a rectangular region, and with our choice of the cost functional (4.10), the contour levels of  $H$  will be circles. Thus the minimization of  $H$  becomes very simple, and may be performed analytically. In Fig 7 two typical situations are illustrated. In Fig 7a it is assumed that the velocity limits are not reached, while Fig 7b illustrates the feasible region  $\Omega$  when the velocities of the trolley and the winch have reached the limits 1.4 meters/sec respectively  $-0.5$  meters/sec.

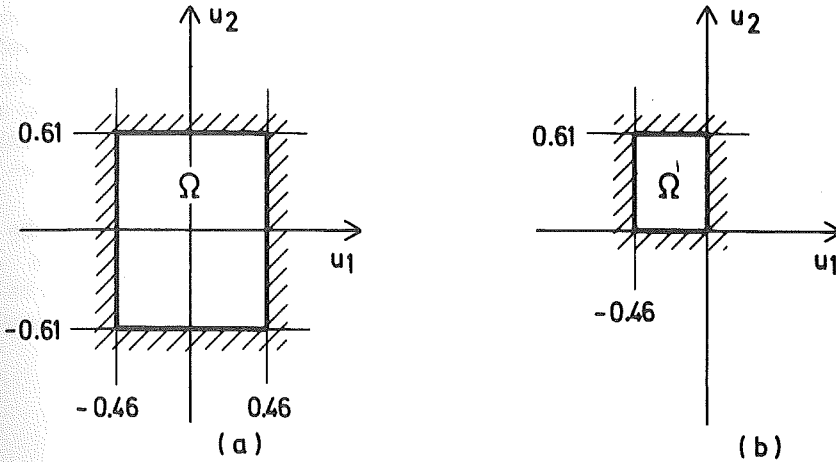


Fig 7. - Feasible region  $\Omega$  in the control space for  
 a)  $-1.4 < x_2 < 1.4$  and  $-0.5 < x_6 < 0.5$   
 b)  $x_2 = 1.4$  and  $x_6 = -0.5$ .

With the inclusion of the quadratic term  $c\psi^T\psi$  in the cost functional (4.10), the algorithm converged fairly fast provided that the initial guess of  $u(t)$  was sufficiently close to the optimal solution. A significant property is that the number of iterations depend heavily on the terminal time  $t_f$ .

Typical figures are 3-4 ( $t_f = 20.0$ ) to 8-9 ( $t_f = 18.12$ ). The corresponding

execution times varied from about 1 min. to about 7-8 min. However, for bad initial guesses of  $u(t)$  the Riccati equation (4.9) still lacked bounded solutions, and the algorithm then rapidly diverged. There are different possibilities to avoid this numerical difficulty. One way is to completely reformulate the cost functional, and for example punish the deviation of the state variables from a suitably chosen smooth trajectory. Another way, and the one chosen here, is to put  $H_{\dot{x}\dot{x}} = 0$  in (4.9). This will not affect the optimal solution, but only the second order terms and the local linear feedback around the optimal solution. This change of the Riccati equation can be interpreted as if we in reality consider the cost functional

$$\bar{J} = c\psi^T\psi + b^T\psi + \int_0^{t_f} \{u_1^2 + u_2^2 - (\bar{x}-\hat{x})^T H_{\dot{x}\dot{x}} (\bar{x}-\hat{x})\}$$

where the nominal trajectory  $\bar{x}(t)$  is successively updated, and where  $H_{\dot{x}\dot{x}}$  is evaluated at  $\bar{x}, \hat{u}$  (see part 3). With this modification, the convergence rate was slightly decreased in the vicinity of the optimal solution, but the stability was considerably improved, and even very bad initial guesses of the control variables could be accepted. Notice however, that since there are no general results established concerning the preservation of convergence when (4.9) is modified, this can only be verified through numerical experiments. However, this possibility to change the second order terms has proved useful for many different problems. It is thus thought that a thorough study of this computational trick could be very profitable and might contribute to further knowledge about computational methods.

#### 4.5 Stack profile constraints

So far we have neglected the possibility that the stack profile must be taken into account. In this section we will briefly consider this problem, and a straightforward way to include the stack profile constraints is illustrated. The analysis is restricted to the particular situation indicated in Fig 2, and we will for simplicity assume that this stack profile can be approximated with the smooth curve

$$-\eta - \frac{7}{25} (\xi - 10)^2 - 5 = 0$$

Thus the motion of the load must satisfy



$$-\eta_2 - \frac{7}{25} (\xi_2 - 10)^2 - 5 \leq 0 \quad (4.11)$$

It should be pointed out that the solution technique used here allows for an arbitrary number of constraints. At the price of increased programming work, it is thus possible to characterize the constraints very accurately with a large number of inequalities. However, to simplify the notations, we will also make the further approximation that the angular deviations are small enough to substitute  $\xi_2$  by  $x_1$  (the position of the trolley) and  $-\eta_2$  by  $x_5$  (the length of the cable). The additional state variable constraint then is

$$S_5(x;t) = x_5 - \frac{7}{25} (x_1 - 10)^2 - 5 \leq 0 \quad (4.12)$$

(4.12) is a second order constraint, and in analogy with the velocity constraints it could be transformed into a constraining hyperplane. However, since (4.12) is a rather crude approximation, it makes no sense to strive for very accurate solutions. We have thus chosen to include  $S_5$  with the Kelley penalty function technique [10]. The cost functional is then modified to

$$\bar{J} = c\psi^T \psi + b^T \psi + \int_0^{t_f} \{u_1^2 + u_2^2 + r_k \cdot h[S_5] \cdot S_5^2\} dt$$

where

$$h[S_5] = \begin{cases} 1 & S_5 \geq 0 \\ 0 & S_5 < 0 \end{cases}$$

and to satisfy  $S_5 \leq 0$ ,  $\bar{J}$  should be successively minimized for a monotone increasing sequence  $\{r_k\}$  with  $\lim_{k \rightarrow \infty} r_k = \infty$ . However,

since we also want to determine the minimum time necessary for the transfer, we must iterate on both  $r_k$  and  $t_f$ . In this case we chose  $t_f$  sufficiently large (25 sec.), and with  $t_f$  fixed, the problem was solved for increasing values of  $r_k$ . It was found that  $r_k = 10$  was sufficient to satisfy  $S_5 \leq 0$  with the specified accuracy. The optimal solution for  $r_k = 10$  was then used as the initial nominal solution for a somewhat smaller value of  $t_f$ . In this way  $t_f$  was successively reduced until the corresponding minimum values of the cost began to rapidly increase. This increase occurred around  $t_f = 23.2$  secs., and thus  $t_f = 23.2$  is

considered as the minimum time. As a comparison, the time required to decrease the cable length to  $x_5 = 5$ , and then, with maximum velocity at  $x_1 = 10$ , bring the trolley to final position is 23.0 secs.

The computed optimal control strategies and the movements of the trolley and the load are illustrated in Fig 8, and the corresponding state variables  $x_2$  (trolley velocity),  $x_3$  (angular deviation) and  $x_6$  (winching velocity) in Fig 9. It can be seen that the winching velocity is the limiting factor before the point  $x_1 = 10$  is reached ( $t = 14.4$ ), and that the trolley velocity is the determining factor after this point has been reached. It can also be noticed that the computed optimal solution agrees fairly well with the intuitive strategy considered above.

#### 4.6 Optimal control in the presence of disturbances

The load transfer is generally exposed to different kinds of disturbances e.g. heavy wind disturbances, and since the accuracy of the terminal state is critical, it must be possible to compensate for these.

One possibility is then to linearize the nonlinear system equations (Model 1) around the optimal solution, and design a stabilizing regulator for this linearized system, for example with linear-quadratic optimal control theory. The corrections  $\delta u$  to the optimal control strategy will then be given by a linear feedback  $\delta u(t) = -L(t)\delta x(t)$  from the deviations  $\delta x$  from the optimal state variables. With a suitable choice of the quadratic criterion, the deviations  $\delta x$  may then be effectively damped out. However, the actual performance of this local regulator will depend on both the control variable constraints  $g(x,u;t) \leq 0$ , which may prevent the corrections  $\delta u$  to be realized, and on the properties of the disturbances. In practice, the efficiency of the local regulator can thus only be verified through simulations.

The second order Differential Dynamic Programming algorithm provides a similar local linear feedback (3.5.24)

$$\delta u = \beta_1 \delta x + \beta_2 \delta b \quad (4.13)$$

for neighbouring optimal solutions, and with a slight modification this may sometimes be used to handle the disturbances. Recalling that (4.13) is obtained when  $\psi$  is directly adjoined to the cost functional by means of Lagrange multipliers  $b$ , (4.13) thus determines the neigh-

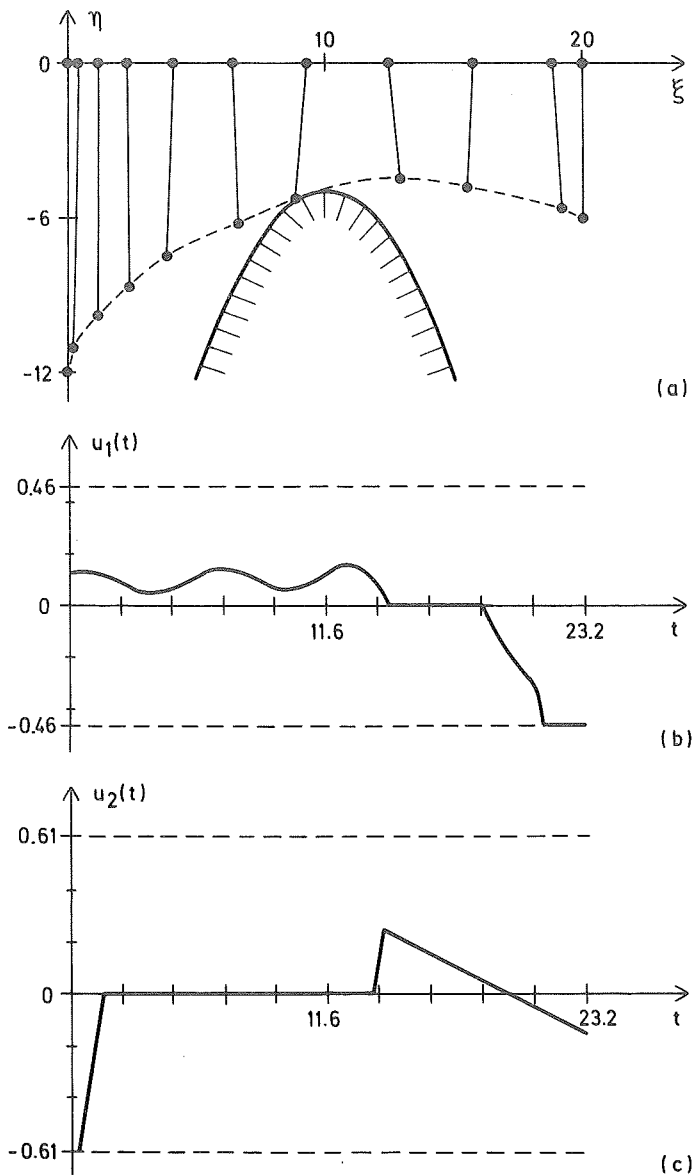


Fig 8. - Computed minimum time solution when the stack profile constraint is considered.

a) - movements of the trolley and the load. The distance between two successive points is  $t = 2.32$  secs.

b) - acceleration of the trolley ( $u_1(t)$ ).

c) - acceleration of the winch ( $u_2(t)$ ).

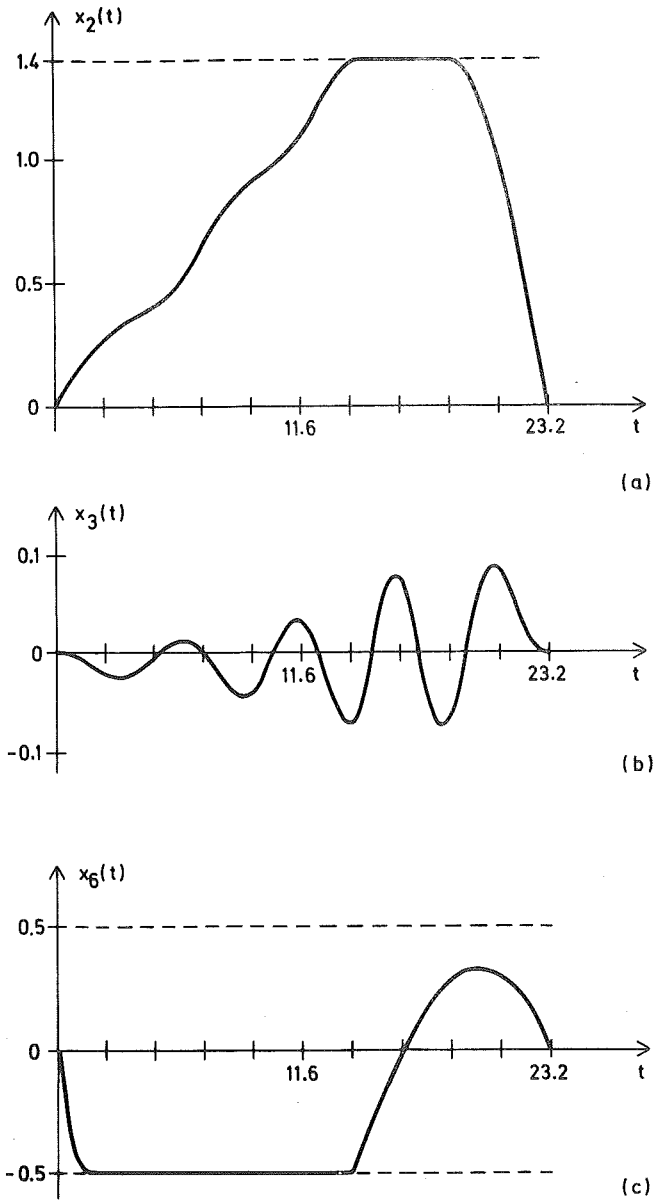


Fig 9. - Computed minimum time solution when the stack profile constraint is considered.

- a) - trolley velocity ( $x_2(t)$ ).
- b) - angular deviation ( $x_3(t)$ ).
- c) - winching velocity ( $x_6(t)$ ).

houring solutions for a free end-point problem. However, from (3.5.46) follows that  $\delta b$  and  $\delta x$  are related (to first order) by

$$\delta b = - V_{bb}^{-1} V_{xb}^T \delta x$$

provided that  $V_{bb}$  is nonsingular. For the terminal constrained problem, the relevant feedback thus is

$$\delta u = (\beta_1 - \beta_2 V_{bb}^{-1} V_{xb}^T) \delta x \quad (4.14)$$

where all quantities are evaluated along the optimal solution. Notice however, that the properties of  $\beta_1$  and  $\beta_2$  (3.5.40) exclude corrections (except for  $Q\hat{g}_x dx$ ) of the optimal control strategy  $\hat{u}$  when the control variable constraints are active, that is, when  $g(\hat{u}) = 0$ . If these constraints are active most of the time, which was shown to be the case for the minimum time problems considered above, the feedback (4.14) will consequently be active only for a small part of the time interval. In practice, this kind of feedback is probably useless. However, the problem may be overcome by increasing the terminal time  $t_f$ . In Section 4.4 it was shown that the bang-bang structure of the minimum time solution is rapidly lost when  $t_f$  increases, and that the influence of the control variable constraints in a corresponding degree is reduced. At the price of increased transfer times it may thus be possible to increase the activity of the local regulator. However, the full efficiency can still only be verified by simulations where the transfer is exposed to realistic disturbances.

#### 4.7 Some aspects on the possibility to realize the optimal control strategies

The analysis in the preceding sections has clearly indicated that the container handling problem is very complex in case a lot of different stack profiles and terminal states must be considered. Due to the many different situations that may occur, off-line computation and storage of the optimal strategies in a mass memory of a computer attached to the process is probably excluded. To realize the full optimal strategies, it then remains to compute the optimal solution in real time. However, even with an extremely capacitive computer like the UNIVAC 1108, these computations are too time-consuming to be done between two

successive transfers. In Section 4.5 it was reported that the execution times are as large as 1 minute in the most advantageous cases and 7-8 minutes in case the initial guess of  $u(t)$  is bad and the terminal time  $t_f$  is close to the minimum time. Since the transfers take about 20 secs., it is thus clear that the execution times will be much too long even if the computer program is specially adapted to this particular problem. With the present capacity of digital computers, it can then be concluded that the implementation of full optimal strategies probably is restricted to cases where the crane either has a rather limited repertoire, or where there are sufficient time available between two successive operations to compute new strategies.

However, the knowledge of the full optimal solutions may be valuable and it may be possible to exploit this knowledge to design some kind of suboptimal control strategies. For example, it can be seen in both Sections 4.4 and 4.5 that the strategy to damp out the oscillations at the terminal time is the same. The load is swung ahead of the trolley, and then the trolley retardates as much as possible. Although this principle seems very natural, and generally is used in manual operation too, it might be worthwhile to further investigate this structure to design a control strategy for only the last part of the transfer. Other possibilities might be to operate with a limited number of control strategies or to divide the total transfer into a number of characteristic movements and then utilize the local feedback. However, it should again be emphasized that the full efficiency of the suboptimal as well as the full optimal control strategies can only be verified through simulations where the transfer is exposed to adequate disturbances.

## 5. OPTIMAL CONTROL OF MODEL 2

Most of the problems considered in the preceding section apply equally well for a torque controlled crane (Model 2). We will thus restrict the study for Model 2 to the simplest case, that is, minimum time transfer when the stack profile can be neglected. It will then turn out that the structure of the optimal strategy is rather similar to Model 1, and thus a detailed study will probably not contribute with any new results. However, a verification of the similarities between the two models is valuable, since it indicates that if the crane is designed to operate with advanced control strategies, it may as well be constructed without the expensive acceleration control equipment.

Similar to Section 4, we will thus briefly account for the state and control variable constraints, and specify the cost functional. The same iterative technique as in Section 4 was used to determine the minimum time, and the computed optimal solution is illustrated. Finally the minimization of the Hamiltonian is briefly discussed, since it proves that this operation is significantly more involved in this case than for Model 1.

### 5.1 State and control variable constraints. Choice of cost functional

In Section 3 it was shown that the operation of the torque controlled crane can be described by the following model:

$$\frac{dx_1}{dt} = x_2 \quad x_1(0) = 0$$

$$\frac{dx_2}{dt} = u_1 + u_2 \delta \cdot \sin x_3 \quad x_2(0) = 0$$

$$\frac{dx_3}{dt} = x_4 \quad x_3(0) = 0$$

$$\frac{dx_4}{dt} = -\frac{g \cdot \sin x_3}{x_5} - \frac{2 x_4 x_6}{x_5} - \frac{u_1 \cdot \cos x_3}{x_5} - \frac{u_2 \delta \cdot \cos x_3 \cdot \sin x_3}{x_5} \quad x_4(0) = 0$$

$$\frac{dx_5}{dt} = x_6 \quad x_5(0) = 12$$

$$\frac{dx_6}{dt} = g \cdot \cos x_3 + x_5 x_4^2 - u_1 \cdot \sin x_3 - u_2 \delta \cdot \sin^2 x_3 - u_2 \quad x_6(0) = 0$$

The state variables are the same as for Model 1, but in this case the control variables are  $u_1 = \frac{F_1}{M}$  and  $u_2 = \frac{F_2}{m}$ .  $\delta$  stands for the

quotient between the mass of the load and the mass of the trolley, that is,  $\delta = \frac{m}{M}$ . Since both  $u_2$  and  $\delta$  depend on  $m$ , the optimal control strategies will be different for different masses of the load. To simplify the analysis, we have thus assumed that  $\delta = 0.4$ . In practice, this roughly corresponds to a container of medium weight.

To make the properties of the models as equal as possible, we have further assumed that  $u_2$  is restricted by

$$g - 0.61 \leq u_2 \leq g + 0.61$$

This implies that the maximum acceleration and retardation of the cable length is the same for both models when the angular deviation and the angular deviation velocity of the load are zero. Similarly, it is assumed that

$$-0.46 \leq u_1 \leq 0.46$$

that is, the maximum acceleration and retardation of the trolley are the same for both models when the angular deviation of the load is zero.

The terminal constraints  $\psi$  are independent of the model, and thus

$$\psi(x(t_f); t_f) = \begin{pmatrix} x_1(t_f) - 20 \\ x_2(t_f) \\ x_3(t_f) \\ x_4(t_f) \\ x_5(t_f) - 6 \\ x_6(t_f) \end{pmatrix}$$

The velocity constraints are also assumed to be the same as for Model 1. However, the transformation of  $S(x;t) \leq 0$  into constraining hyperplanes is dependent on the system dynamics, and in this case we get



$$\begin{aligned}
u_1 + u_2 \delta \cdot \sin x_3 + a_1 (x_2 - 1.4) &\leq 0 \\
-u_1 - u_2 \delta \cdot \sin x_3 + b_1 (-x_2 - 1.4) &\leq 0 \\
g \cdot \cos x_3 + x_5 x_4^2 - u_1 \cdot \sin x_3 - u_2 \delta \cdot \sin^2 x_3 - u_2 + c_1 (x_6 - 0.5) &\leq 0 \\
-g \cdot \cos x_3 - x_5 x_4^2 + u_1 \cdot \sin x_3 + u_2 \delta \cdot \sin^2 x_3 + u_2 + d_1 (-x_6 - 0.5) &\leq 0
\end{aligned}$$

Similar to Model 1, the slopes  $a_1, \dots, d_1$  were all set equal to 10.

For the different reasons discussed in Section 4, the adjoined cost functional was chosen as

$$\bar{J} = 2 \psi^T \psi + b \psi + \int_0^{t_f} \{u_1^2 + (u_2 - g)^2\} dt$$

and for stability reasons,  $H_{xx}$  was set equal to zero in the Riccati equation (4.9).

## 5.2 Numerical solution

To determine the minimum time required for the transfer, the iterative procedure described in Section 4 was used, that is,  $t_f$  was successively decreased until the terminal constraints  $\psi$  could not be satisfied. However, in this case it turned out that the structure of the optimal control strategy was extremely sensitive to the terminal time, and the accuracy reached for Model 1 was impossible to reproduce. We can thus only conclude that the minimum time is about 18.4 secs. In Fig 10 the computed optimal control variables are shown for  $t_f = 18.5$ , and the corresponding state variables  $x_2$  (trolley velocity),  $x_3$  (angular deviation) and  $x_6$  (winching velocity) are illustrated in Fig. 11. It can be seen that the basic structure of the control and state variables are the same as for Model 1. The additional time required for the transfer (compared with Model 1) is explained by the negative angular deviation during the first five seconds. The trolley acceleration is thus decreased in this interval, and since the deviation does not become positive until the trolley velocity is reached, it is not possible to compensate this through an increased acceleration. It should also be noticed, that the strategy to damp out the oscillations at the terminal time is the same as for

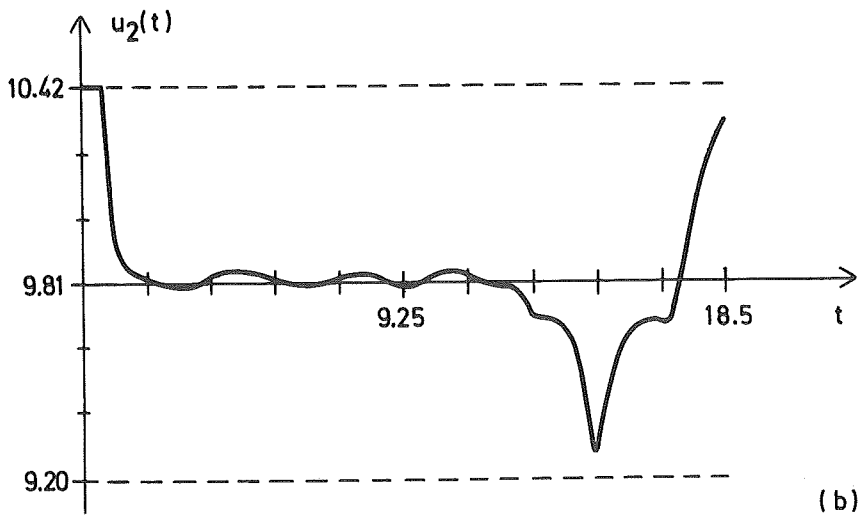
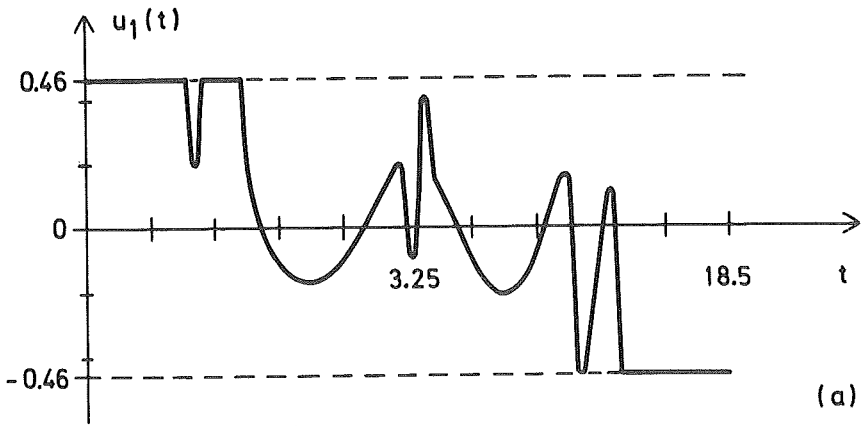


Fig 10. - Computed optimal control variables  $u_1(t)$  and  $u_2(t)$  for Model 2 with  $t_f = 18.5$ .

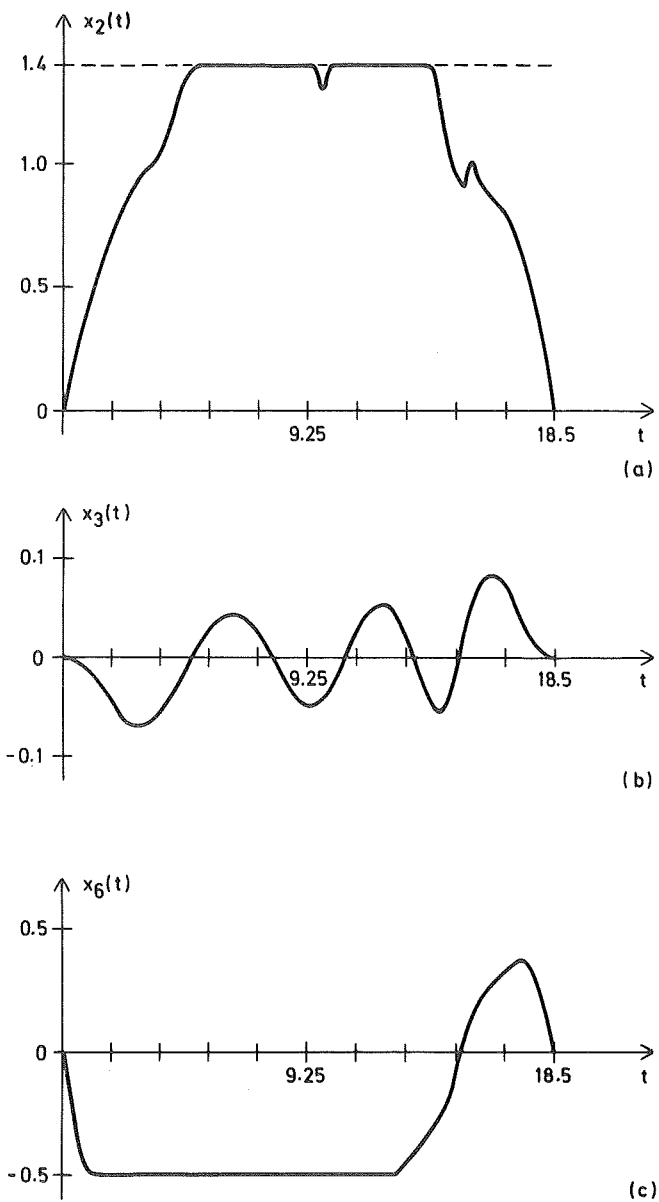


Fig 11. - Computed optimal solution for Model 2 with  $t_f = 18.5$   
 a) - trolley velocity ( $x_2(t)$  ).  
 b) - angular deviation ( $x_3(t)$  ).  
 c) - winching velocity ( $x_6(t)$  ).

Model 1, namely to swing the load ahead, and at the same time retardate the trolley as much as possible.

Model 2 also well illustrates that the complete minimization of the Hamiltonian may be a drawback of the Differential Dynamic Programming algorithm (see part 3). In Fig 12 the feasible region  $\Omega$  in the control space is illustrated for the case when both the trolley velocity and the winching velocity limits are reached. Fig 12a roughly illustrates the character of  $\Omega$  for negative angular deviations  $x_3$ , and in Fig 12b,  $\Omega$  is outlined for a positive angular deviation. It can be seen, that

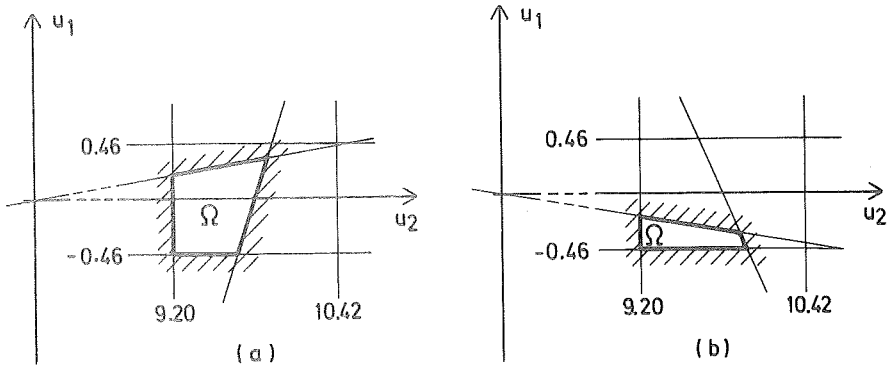


Fig 12. - Typical feasible regions  $\Omega$  in the control space for  
a) negative, and b) positive angular deviations  $x_3$ .

although the contour levels of the Hamiltonian are circles, the minimization is not straightforward, and the analytical minimization thus required a great deal of programming work. An alternative would then be to use a separate algorithm for finite-dimensional optimization in each step in the backwards integration loop. However, this was found to heavily increase the execution time, which already was rather large (about 6 min.), and to decrease the accuracy below an acceptable level.

## 6. REFERENCES

- [1] E. Anselmino and T. M. Liebling, "Zeitoptimale Regelung der Bewegung einer hängenden Last zwischen zwei beliebigen Randpunkten", Proc. International Analogue Computational Meetings, Lausanne, vol. I, September 1967, 482-492.
- [2] A. E. Bryson, W. F. Denham and S. E. Dreyfus, "Optimal Programming Problems with Inequality Constraints I: Necessary Conditions for Extremal Solutions", AIAA J., vol. 1, 1963, 2544-2550.
- [3] A. E. Bryson and Y. C. Ho, "Applied Optimal Control", Blaisdell, Waltham, 1969.
- [4] W. R. Dodds, "Optimization of Ore Unloading Systems Using the Parameter Sweep Technique", Proc. International Analogue Computational Meetings, Lausanne, vol. I, September 1967, 476-481.
- [5] P. Dyer and S. R. McReynolds, "The computation and Theory of Optimal Control", Academic Press, New York and London, 1970.
- [6] R. E. Foerster and I. Flügge-Lotz, "A Neighbouring Optimal Feedback Control Scheme for Systems Using Discontinuous Control", Journal of Optimization Theory and Applications, vol. 8, 1971, 367-395.
- [7] B. Friedland and P. E. Sarachik, "A Unified Approach to Suboptimum Control", Proc. Third Congress of the International Federation of Automatic Control, London, 1966, vol. 1.
- [8] P. Hippe, "Zeitoptimale Steuerung eines Erzentladlers", Regelungstechnik und Prozess-Datenverarbeitung, vol. 18, 1970, 346-350.
- [9] D. H. Jacobson and D. Q. Mayne, "Differential Dynamic Programming", American Elsevier Publishing Comp., New York, 1970.

- [10] H. J. Kelley, "Methods of Gradients", in G. Leitman, Ed., "Optimization Techniques", Academic Press, New York and London, 1962.
- [11] L. S. Lasdon, S. K. Mitter and A. D. Waren, "The Conjugate Gradient Method for Optimal Control Problems", IEEE Trans. Automatic Control, vol. 12, 1967, 132-138.
- [12] K. Mårtensson, "A Constraining Hyperplane Technique for State Variable Constrained Optimal Control Problems", Research Report, Division of Automatic Control, Lund Institute of Technology, to appear. (Part 3 of this thesis).
- [13] K. Mårtensson, "Computational Methods for Optimal Control Problems", Research Report, Division of Automatic Control, Lund Institute of Technology, to appear.
- [14] K. Mårtensson, "A New Approach to Constrained Function Optimization", Research Report 7112, Division of Automatic Control, Lund Institute of Technology, March 1971. (Part 2 of this thesis).
- [15] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mischenko, "The Mathematical Theory of Optimal Processes", Wiley, New York, 1962.

