On the Matrix Riccati Equation

Mårtensson, Krister

1970

Document Version:
Publisher's PDF, also known as Version of record

ON THE MATRIX RICCATI EQUATION

K. MÅRTENSSON

THESIS FOR THE DEGREE
OF TEKNOLOGIE LICENTIAT

REPORT 7002 APRIL 1970
LUND INSTITUTE OF TECHNOLOGY
DIVISION OF AUTOMATIC CONTROL

ON THE MATRIX RICCATI EQUATION †

K. Mårtensson

ABSTRACT

Properties of the algebraic equation

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

are studied for arbitrary nonnegative definite and positive definite matrices Q_1 and Q_2. The possible number of stationary solutions of the Riccati equation is established. The theory for linear systems with quadratic loss is then generalized, and numerical consequences are studied.

ACKNOWLEDGEMENTS

The author would like to thank Professor K.J. Åström for his valuable comments and suggestions, and Mrs. G. Christensen, who typed all the manuscripts.

---

† This work was supported by the Swedish Board for Technical Development (Contract 69-631/U489).

TABLE OF CONTENTS

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 on optimal control of linear time invariant systems with quadratic loss. Consider the system

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

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

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

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

$$u(t) = L(t)x(t) \tag{1.3}$$

where

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

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

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

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

$$S(t_f) = Q_0 \tag{1.6}$$

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

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

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

$$u(t) = -Lx(t) \tag{1.8}$$

where

$$L = Q_2^{-1}B^TS \tag{1.9}$$

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

$$A^TS + SA - SBQ_2^{-1}B^TS + Q_1 = 0 \tag{1.10}$$

If an observability criteria is imposed on the pair [C, A], where $Q_1 = C^TC$ and rank C = rank Q_1, the unique solution S of (1.10) is positive definite, and the optimal system

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

is asymptotic stable [5], [6]. If [C, A] is just detectable, that is unstable modes are observable, the optimal system is still asymptotic stable, but the unique nonnegative definite solution of (1.11) 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 algebraic 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 was assumed to have distinct eigenvalues, while our results hold even for multiple eigenvalues. This was found necessary since distinct eigenvalues restricted the possible choices of the criteria matrices Q_1 and Q_2. 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 7 and 8 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 à 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 to give 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 $A^TX + XA - XBQ_2^{-1}B^TX + Q_1 = 0$

2.1. GENERAL FORM OF THE SOLUTIONS

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

$$A^TX + XA - XBQ_2^{-1}B^TX + Q_1 = 0 \tag{2.1}$$

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

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

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 λ_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 λ_i is an eigenvalue of multiplicity k, the corresponding eigenvectors are defined as the nontrivial solutions of

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

$a_1, a_2, \ldots, a_k$ will be called the generalized eigenvectors, and a_i is the eigenvector of rank i corresponding to the multiple eigenvalue λ_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

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

will bring E on Jordan block form.

Following [1] we then have

Theorem 1:

Each solution of (2.1) can be expressed as

$$X = [c_1 \ldots c_n][b_1 \ldots b_n]^{-1} \tag{2.4}$$

where the inverse is assumed to exist for certain combinations of eigenvectors. Conversely, if $[b_1 \ldots b_n]$ is nonsingular, then

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

satisfies (2.1).

Proof:

Suppose X is a solution of (2.1) and introduce

$$G = A - BQ_2^{-1}B^TX \tag{2.5}$$

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

$$XG = XA - XBQ_2^{-1}B^TX \tag{2.6}$$

and substitute in (2.1)

$$XG = -A^TX - Q_1 \tag{2.7}$$

Let S be a transformation that brings G on Jordan form. Then

$$S^{-1}G\,S = J$$

Further, let

$$R = XS$$

Then

$$\begin{aligned} G &= SJS^{-1} \\ X &= RS^{-1} \end{aligned} \tag{2.8}$$

Substitute into (2.6) and (2.7).

$$SJ = AS - BQ_2^{-1}B^TR$$

$$RJ = -A^TR - Q_1S$$

or

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

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

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

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

$$a_i\lambda_i = Ea_i$$

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

$$\begin{aligned} a_i\lambda_i &= Ea_i \\ a_i + \lambda_i a_{i+1} &= Ea_{i+1} \\ &\vdots \\ a_{i+k-2} + \lambda_i a_{i+k-1} &= Ea_{i+k-1} \end{aligned}$$

or

$$\begin{aligned} (E - \lambda_i I)a_i &= 0 \\ (E - \lambda_i I)a_{i+1} &= a_i \\ &\vdots \\ (E - \lambda_i I)a_{i+k-1} &= a_{i+k-2} \end{aligned} \qquad (2.10)$$

Since S is assumed nonsingular, a_i, $i = 1 \ldots n$, cannot be identical to the null vector, and thus the system (2.10) must have nontrivial solutions. But this holds if and only if λ_i is an eigenvalue of multiplicity k to E, and then $a_i, \ldots, a_{i+k-1}$ 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

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

The extension to non-diagonal Jordan forms obviously restricts the possibilities to compose a solution out of $2n$ arbitrary eigenvectors. Suppose λ_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 $a_i, \ldots, a_{i+k-2}$ with rank $1, \ldots, k-1$ must also be columns in

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

Consequently the à priori upper limit for the possible number of solution 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

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

and assume that

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

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

It is easy to verify that

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

and the solutions will then be the same.

$$\left[c_1 \cdots c_i c_j \cdots c_n\right]\left[b_1 \cdots b_i b_j \cdots b_n\right]^{-1} =$$

$$= \left[c_1 \cdots c_j c_i \cdots c_n\right]\left[b_1 \cdots b_j b_i \cdots b_n\right]^{-1}$$

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 restrictions imposed by a non-diagonal Jordan form is illustrated in the following example.

Let

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

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

$$a^1_{\lambda=1} = \begin{pmatrix} 1 \\ 2 \\ 2 \\ -2 \end{pmatrix} ; \quad a^2_{\lambda=1} = \begin{pmatrix} -1 \\ -3/2 \\ 1 \\ 0 \end{pmatrix} ; \quad a^1_{\lambda=-1} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} ; \quad a^2_{\lambda=-1} = \begin{pmatrix} 1 \\ 3/2 \\ 0 \\ 0 \end{pmatrix}$$

Suppose $a^1_{\lambda=1}$ and $a^2_{\lambda=-1}$ 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.

Corrollary:

Let

$$a_i = \begin{pmatrix} b_i \\ c_i \end{pmatrix} \qquad i = 1 \ldots n$$

be eigenvectors of

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

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

Proof:

The corrollary follows immediately from the fact that J is the Jordan form of $A - BQ_2^{-1}B^TX$ and $S = [b_1 \ \ldots \ b_n]$ 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 λ is an eigenvalue of E, $\bar{\lambda}$ ($\bar{\lambda}$ is the complex conjugate of λ), $-\lambda$ 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 $[b_1 \ \ldots \ b_n]^{-1}$ exists, it is possible to find a solution X of (2.1) such that the closed loop system matrix $A - BQ_2^{-1}B^TX$ is asymptotic stable.

2.2. HERMITIAN AND REAL SYMMETRIC SOLUTIONS

Next we concentrate upon those solutions X of (2.1) which has the property that they are hermitian. From [1] we have the follcwing theorem.

Theorem 2:

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

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

is hermitian.

Proof:

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

$$P = \left[b_1 \cdots b_n\right]^* \left[c_1 \cdots c_n\right] \tag{2.11}$$

where $\left[b_1 \cdots b_n\right]^*$ is the adjoint of $\left[b_1 \cdots b_n\right]$. Then

$$X = \left\{\left[b_1 \cdots b_n\right]^{-1}\right\}^* P\left\{\left[b_1 \cdots b_n\right]^{-1}\right\}$$

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

$$T = \begin{pmatrix} 0_n & I_n \\ -I_n & 0_n \end{pmatrix}$$

Since E is Hamiltonian [4] it is then easily verified that

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

From (2.11) we have

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

and

$$P_{jk} - \tilde{P}_{kj} = b_j^* c_k - c_j^* b_k = a_j^* T a_k$$

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

$$P_{jk} - \tilde{P}_{kj} = (\bar{\lambda}_j + \lambda_k)^{-1}(\bar{\lambda}_j a_j^* T a_k + \lambda_k a_j^* T a_k) \tag{2.12}$$

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 $Ea_j = \lambda_j a_j$, $Ea_k = \lambda_k a_k$.

Then

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

$$= (\bar{\lambda}_j + \lambda_k)^{-1} a_j^* (E^T T + TE) a_k = 0$$

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

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

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

$$= (\bar{\lambda}_j + \lambda_k)^{-1}(a_j^*(E^T T + TE)a_k - a_j^* T a_{k-1}) =$$

$$= -(\bar{\lambda}_j + \lambda_k)^{-1} a_j^* T a_{k-1}$$

Analoguous to (2.12) this is equivalent to

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

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

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

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 λ_j and λ_k are assumed to be multiple eigenvalues, and a_j, a_k are generalized eigenvectors both of rank larger than one. Then

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

which yields

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

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.

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

$$+ a_{j-1}^* Ta_{k-1} + a_j^* Ta_{k-1})$$

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

Now let λ_r be an eigenvalue of multiplicity r, and $a_1, \ldots, a_r$ the corresponding eigenvectors. Then any attempt to include $a_1, \ldots, a_k$ but not $a_{k+1}, \ldots, a_r$, $1 \leq k < r$, in the solution will violate the condition $\bar{\lambda}_j \neq -\lambda_k$. The reason for this is as follows:

If we have selected $a_1, \ldots, a_k$ we cannot make use of any of the r eigenvectors corresponding to $-\bar{\lambda}_r$. From the remaining $2n - 2r$ eigenvectors we must either choose the one corresponding to λ_i or the one corresponding to $-\bar{\lambda}_i$, 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 possibilities to satisfy the sufficient condition for X to be hermitian, is to include all eigenvectors corresponding to λ_r or all eigenvectors corresponding to $-\bar{\lambda}_r$.

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

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

$$\begin{pmatrix} 2n \\ n \end{pmatrix}$$

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:

Necessary and sufficient conditions for a solution

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

to be real are

i) all eigenvectors $a_1, \dots, a_n$ are real,

or

ii) if a_i of rank k corresponding to the eigenvalue λ_i, $\mathrm{Im}(\lambda_i) \neq 0$, is used to construct the solution X, then $\bar{a}_i$ of rank k corresponding to $\bar{\lambda}_i$ must also be included in the solution.

Proof:

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

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

and

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

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

$$G = A - BQ_2^{-1}B^TX$$

G is real if X is real, and then the eigenvalues of G are real or complex conjugated. But according to the corrollary 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 $Q_1 = C^TC$, is completely observable, the stationary solution 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 Q_1 to be an arbitrary symmetric nonnegative definite matrix. Thus A is allowed to have one or more unstable modes not detectable in [C, A]. (In the sequel we will use the notation $[Q_1, A]$.) It will further be assumed that the eigenvalues of A are different, and none of the undetectable modes are pure imaginary or zero.

Properties of the solution X due to nonobservable modes of $[Q_1, A]$ will be considered in this section. These results will later be used in 2.4 and in section 3 to establish properties of the optimal control problem.

Since A is assumed to have distinct eigenvalues we will use 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 $\lambda_1, \ldots, \lambda_n$ are distinct eigenvalues of A. The mode λ_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 $Q_1 = C^TC$, has at least one element not identical zero.

Let λ_i be a nonobservable mode of the pair $[Q_1, A]$, and let x_i be the corresponding eigenvector. Then from the definition, $Cx_i = 0$ and $Q_1x_i = C^TCx_i = 0$.

Theorem 4:

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

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

and the corresponding eigenvector is

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

Proof:

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

$$\begin{pmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{pmatrix} \begin{pmatrix} x_i \\ 0_n \end{pmatrix} = \begin{pmatrix} Ax_i \\ -Q_1x_i \end{pmatrix} = \lambda_i \begin{pmatrix} x_i \\ 0_n \end{pmatrix}$$

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 $\lambda_1, \ldots, \lambda_n$ are distinct eigenvalues of A. The mode λ_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 λ_i is a noncontrollable mode of [A, B] and y_i^T the corresponding left hand eigenvector of A, then, analoguous 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 λ_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 is

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

Proof:

$$\begin{pmatrix} A & -BQ_2^{-1}B^T \\ -Q_1 & -A^T \end{pmatrix} \begin{pmatrix} 0_n \\ y_i \end{pmatrix} = \begin{pmatrix} -BQ_2^{-1}B^T y_i \\ -A^T y_i \end{pmatrix} = -\lambda_i \begin{pmatrix} 0_n \\ y_i \end{pmatrix}$$

It is now possible to justify the demand for stabilizability of the pair $[A, B]$, that is controllability of unstable modes [8], with pure algebraic considerations. Suppose there exists a non-controllable mode $\lambda_i > 0$. Then according to theorem 5, $-\lambda_i$ is an eigenvalue of E, and the corresponding eigenvector has the structure

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

Among all possible solutions of (2.1), there is one and only one solution X such that the closed loop system matrix $A - BQ_2^{-1}B^TX$ is asymptotic stable. This solution X consists of the eigenvectors corresponding to the eigenvalues with negative real parts. Thus the eigenvector

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

must be included. But $b_i = 0_n$ and then the matrix $[b_1 \dots b_n]$ will be singular. This contradicts the assumption that there exists a solution X of (2.1) such that $A - BQ_2^{-1}B^TX$ is stable.

Since one of the main purposes of optimal control theory is to yield an asymptotic stable closed loop system, we will then from now on assume that the pair $[A, B]$ is stabilizable.

The conditions for symmetry proved in section 2.1, can now be extended to cover situations where nonobservable modes of $[Q_1, A]$ appear. 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 $p_{jk} = \bar{p}_{kj}$. Now assume $\lambda_j = -\bar{\lambda}_k$, where both λ_j and λ_k are non-observable modes of $[Q_1, A]$. Then

$$P_{jk} - \bar{P}_{kj} = b_j^* c_k - c_j^* b_k = 0$$

since $c_j = c_k = 0_n$, and thus P is still hermitian. Finally the situation may occur that λ_j is a nonobservable mode of the pair $[Q_1, A]$, and that the criteria matrices Q_1 and Q_2 are chosen so that $\lambda_k = -\lambda_j$ is an eigenvalue of E not due to the symmetry properties of E.

For simplicity assume that λ_j is real. As the eigenvalues of a matrix are continuous functions of the matrix elements, it follows that both λ_j and λ_k then will be multiple. λ_j being nonobservable implies that $c_j = 0_n$ and

$$P_{jk} - \bar{P}_{kj} = b_j^* c_k - c_j^* b_k = b_j^* c_k$$

It then remains to prove that $b_j^* c_k = b_j^T c_k = 0$.

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

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

Introduce the $2n \times 2n$ matrix

$$V = \begin{pmatrix} T & 0_n \\ 0_n & (T^{-1})^T \end{pmatrix}$$

where 0_n denotes the $n \times n$ null matrix. As V is nonsingular the eigenvalues of

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

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

$$\hat{a}_i = Va_i \qquad (2.14)$$

This holds for generalized eigenvectors too.

Carrying out the transformation VEV^{-1} we get

$$\hat{E} = \begin{pmatrix} TAT^{-1} & -TBQ_2^{-1}B^TT^T \\ -(T^{-1})^TQ_1T^{-1} & -(T^{-1})^TA^TT^T \end{pmatrix}$$

which reduces to

$$\hat{E} = \left(\begin{array}{cc|c} \lambda_1 & & \\ & \ddots & \\ & \lambda_j & \\ & & \ddots \\ & & \lambda_n \end{array} \quad \begin{array}{c} \\ \\ R \\ \\ \\ \end{array} \right) \qquad (2.15)$$

$$\left(\begin{array}{c} \\ \\ P \\ \\ \\ \end{array} \quad \begin{array}{cc} & \\ & \\ -\lambda_1 & \\ & \ddots \\ & -\lambda_j \\ & \ddots \\ & -\lambda_n \end{array}\right)$$

where $R = -TBQ_2^{-1}B^TT^T$ and $P = -(T^{-1})^TQ_1T^{-1} = -(CT^{-1})^T(CT^{-1})$. As λ_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 equals zero.

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

$$\hat{A}_k^1 = \begin{pmatrix} \hat{B}_1^1 \\ \vdots \\ \hat{B}_n^1 \\ \hat{C}_1^1 \\ \vdots \\ \hat{C}_j^1 \\ \vdots \\ \hat{C}_n^1 \end{pmatrix}$$

as the corresponding eigenvector of rank one.

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

$$(\hat{E} - \lambda_k I)\tilde{a}_k^2 = (\hat{E} + \lambda_j I)\tilde{a}_k^2 = \tilde{a}_k^1$$

But

$$(\hat{E} + \lambda_j I) = \left(\begin{array}{ccc|ccc} \lambda_1+\lambda_j & & & & & \\ & \ddots & & & R & \\ & & \lambda_n+\lambda_j & & & \\ \hline & & & -\lambda_1+\lambda_j & & \\ & P & & & \ddots & \\ & & & & 0 & \ddots \\ & & & & & -\lambda_n+\lambda_j \end{array}\right) \qquad (2.16)$$

which, since the j:th row of P equals zero, implies

$$\tilde{c}_j^1 = 0$$

From (2.14)

$$\tilde{a}_k^1 = \begin{pmatrix} T & 0_n \\ 0_n & (T^{-1})^T \end{pmatrix}\begin{pmatrix} b_k^1 \\ c_k^1 \end{pmatrix}$$

and since $(T^{-1})^T = (x_1 \dots b_j \dots x_n)^T$ we get

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

Then

$$P_{jk} - \bar{P}_{kj} = 0$$

which completes the proof. If $\lambda_k = -\lambda_j$ is an eigenvalue of multiplicity m, the proof still holds for the generalized eigenvectors corresponding to λ_k up to rank m-1. This follows from

$$[E - (-\lambda_j)I]\tilde{a}_k^{\ell} = \tilde{a}_k^{\ell-1} \qquad 2 \leq \ell \leq m$$

and then

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

The results are summarized in the following theorem.

Theorem 6:

Suppose λ_j is a nonobservable distinct mode of $[Q_1, A]$. Further let $a_1, \ldots, a_n$ be eigenvectors of E corresponding to $\lambda_1, \ldots, \lambda_j, \lambda_k, \ldots, \lambda_n$ and assume that $[b_1 \ldots b_n]^{-1}$ exists. If $\lambda_j \neq \lambda_i$, $1 \leq i \leq n$, and $\lambda_k = -\bar{\lambda}_j$, then

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

is hermitian if either

i) λ_k is a nonobservable mode of $[Q_1, A]$

or

ii) the number of generalized eigenvectors corresponding to λ_k included in the solution is less than or equal to m-1, where m is the multiplicity of λ_k.

The theorem is illustrated in the following example.

$$A = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad Q_1 = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix} \qquad Q_2 = (1)$$

The eigenvalues of E are

$\lambda_1 = 2$ 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. λ_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 λ_1 and λ_2 are

$$a_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad a_2 = \begin{pmatrix} 1 \\ 4 \\ 0 \\ 4 \end{pmatrix}$$

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 and 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 $a_1, \ldots, a_n$ will cause $X = [c_1 \ldots c_n][b_1 \ldots b_n]^{-1}$ to be nonnegative definite?

Theorem 7:

Let $a_1, \ldots, a_n$ be eigenvectors corresponding to $\lambda_1, \ldots, \lambda_n$. Assume that $\mathrm{Re}\{\lambda_i\} \neq 0$, $i = 1 \ldots n$. If

$$X = \begin{bmatrix} c_1 \cdots c_n \end{bmatrix} \begin{bmatrix} b_1 \cdots b_n \end{bmatrix}^{-1}$$

is symmetric and positive definite, then $\mathrm{Re}\{\lambda_i\} < 0$, $i = 1 \ldots n$.

Proof:

Consider the closed loop system matrix

$$G = A - BQ_2^{-1}B^TX$$

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

$$G^TX + XG = -(Q_1 + XBQ_2^{-1}B^TX)$$

The asymptotic stability of G then follows immediately from Lyapunov stability theory.

Since there is only one way to select n eigenvalues of E with $\mathrm{Re}\{\lambda\} < 0$, theorem 7 implies that (2.1) can never have more than one positive definite solution. Following [1], the reversal of theorem 7 is

Theorem 8:

Suppose Q_1 is a nonnegative definite and Q_2 a positive definite symmetric matrix, and let $a_1, \ldots, a_n$ be eigenvectors corresponding to $\lambda_1, \ldots, \lambda_n$. If $\mathrm{Re}\{\lambda_i\} < 0$, $i = 1 \ldots n$, and $\begin{bmatrix} b_1 \cdots b_n \end{bmatrix}$ is nonsingular, then

$$X = \begin{bmatrix} c_1 \cdots c_n \end{bmatrix} \begin{bmatrix} b_1 \cdots b_n \end{bmatrix}^{-1}$$

is symmetric and nonnegative definite.

Proof:

The symmetry immediately follows from theorems 2 and 3. To prove that X is nonnegative definite, let

$$X = \left\{\left[b_1 \cdots b_n\right]^{-1}\right\}^* P\left\{\left[b_1 \cdots b_n\right]^{-1}\right\}$$

where

$$P = \left[b_1 \cdots b_n\right]^* \left[c_1 \cdots c_n\right]$$

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

$$U(t) = \left[e^{\lambda_1 t} a_1, \ldots, e^{\lambda_n t} a_n\right]$$

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

$$U(t) = \left[e^{\lambda_1 t} a_1, \ldots, e^{\lambda_k t} a_k, e^{\lambda_k t}(a_{k+1} + a_k t), \ldots\right.$$

$$\ldots, e^{\lambda_k t}\left(a_{k+r} + a_{k+r-1} \cdot t + \ldots + \frac{a_k \cdot t^{r-1}}{(r-1)!}\right), \ldots$$

$$\left.\ldots, e^{\lambda_n t} a_n\right] \qquad (2.17)$$

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

$$\frac{dU(t)}{dt} = EU(t) \qquad U(0) = \left[a_1, \ldots, a_n\right]$$

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

$$L = \begin{bmatrix} 0_n & I_n \\ 0_n & 0_n \end{bmatrix}$$

where 0_n is the null matrix of order $n \times n$.

Then

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

Further introduce

$$S(t) = - U^*(t)LU(t) + U^*(0)LU(0) \qquad (2.18)$$

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

$$\lim_{t \to \infty} U(t) = 0$$

and thus

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

(2.18) is equivalent to

$$S(t) = - \int_0^t \frac{d}{ds} [U^*(s)LU(s)]ds =$$

$$= - \int_0^t [U^*(s)E^TLU(s) + U^*(s)LEU(s)]ds$$

$$= - \int_0^t U^*(s)[E^TL + LE]U(s)ds$$

But

$$E^TL + LE = \begin{bmatrix} -Q_1 & 0_n \\ 0_n & -BQ_2^{-1}B^T \end{bmatrix}$$

and then $S(t) \geq 0$ $\forall t$. As $t \to \infty$, $S(t) \to P$, and thus P is nonnegative definite.

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. In theorem 9 conditions for the nonexistence of positive definite solutions are given, and in theorem 10 it is proved that in some cases there may be several nonnegative definite solutions of (2.1).

Theorem 9:

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

Proof:

The theorem is proved by contradiction. Assume there is a positive definite solution of (2.1). Then theorem 7 implies that the eigenvector a_i corresponding to $\lambda_i < 0$ must be included in the solution. But since λ_i is nonobservable, a_i has the structure

$$a_i = \begin{bmatrix} b_i \\ 0_n \end{bmatrix}$$

according to theorem 4. Then

$$X = \begin{bmatrix} c_1 & \cdots & 0_n & \cdots & c_n \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_i & \cdots & b_n \end{bmatrix}^{-1}$$

is singular, which contradicts the assumption that X is positive definite.

Theorem 10:

Let $\lambda_i > 0$ be a distinct nonobservable mode of $[Q_1, A]$. Then there are at least two nonnegative definite solutions of (2.1).

Proof:

Since $\lambda_i > 0$ is a nonobservable mode of $[Q_1, A]$, both λ_i and $-\lambda_i$ are eigenvalues of E. Suppose X_1 is constructed from the eigenvectors corresponding to $-\lambda_i$ and the remaining n-1 eigenvalues with $\mathrm{Re}\{\lambda\} < 0$. Then X_1 is nonnegative definite according to theorem 8. Now let $-\lambda_i$ be replaced by λ_i, and let X_2 be the corresponding solution. To simplify the proof, we assume that the eigenvalues of E are distinct. Connecting to the proof of theorem 8, it remains to prove that $U^*(t)LU(t) \to 0$ as $t \to \infty$. From the definition of U follows

$$U^*(t)LU(t) = \begin{bmatrix} b_1^* e^{\bar{\lambda}_1 t} & c_1^* e^{\bar{\lambda}_1 t} \\ \vdots & \vdots \\ b_i^* e^{\bar{\lambda}_i t} & c_i^* e^{\bar{\lambda}_i t} \\ \vdots & \vdots \\ b_n^* e^{\bar{\lambda}_n t} & c_n^* e^{\bar{\lambda}_n t} \end{bmatrix} \begin{bmatrix} 0_n & I_n \\ 0_n & 0_n \end{bmatrix} \begin{bmatrix} b_1 e^{\lambda_1 t} & \cdots & b_i e^{\lambda_i t} & \cdots & b_n e^{\lambda_n t} \\ c_1 e^{\lambda_1 t} & \cdots & c_i e^{\lambda_i t} & \cdots & c_n e^{\lambda_n t} \end{bmatrix}$$

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

$$[U^*(t)LU(t)]_{k\ell} = b_k^* c_\ell e^{\bar{\lambda}_k t} e^{\lambda_\ell t}$$

Since

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

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

$$b_k^* c_\ell e^{\bar{\lambda}_k t} e^{\lambda_\ell t} \to 0$$

as $t \to \infty$ as $\mathrm{Re}(\lambda_k) < 0$ and $\mathrm{Re}(\lambda_\ell) < 0$. But the i:th column of $U^*(t)LU(t)$ is identical to the zero column vector since $c_i = 0_n$. The symmetry then implies that the i:th row equals zero too. Thus $U^*(t)LU(t) \to 0$ as $t \to \infty$, and $S(t) = -U^*(t)LU(t) + U^*(0)LU(0) + P$. It then follows from theorem 8 that X_2 is nonnegative definite. The solutions X_1 and X_2 are not identical, since the eigenvalues of $G_1 = A - BQ_2^{-1}B^TX_1$ and $G_2 = A - BQ_2^{-1}B^TX_2$ are different. This completes the proof of two different nonnegative definite solutions of (2.1).

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

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

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

$\lambda_6 = 3$ $\lambda_6 = -\lambda_5$

As $\lambda_5 = -3$ is nonobservable, there is no positive definite solution of (2.1) according to theorem 9. The corresponding eigenvectors are

$$a_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad a_2 = \begin{pmatrix} 3 \\ 2 \\ -3 \\ 6 \\ 0 \\ 0 \end{pmatrix} \quad a_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad a_4 = \begin{pmatrix} 4 \\ 8 \\ -12 \\ 0 \\ 12 \\ 0 \end{pmatrix} \quad a_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad a_6 = \begin{pmatrix} -3 \\ -6 \\ -1 \\ 0 \\ 0 \\ 6 \end{pmatrix}$$

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

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

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

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

iv) a_2, a_4, a_5; $X_4 = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 & 4 & 0 \\ 2 & 3 & 0 \\ -3 & -12 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 28 & -24 & 0 \\ -24 & 36 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

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 $\lambda_1, \ldots, \lambda_m$ are m distinct nonobservable modes of $[Q_1, A]$ such that $\mathrm{Re}(\lambda_i) \geq 0$, $i = 1 \ldots n$. Using the result of theorem 10 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 $[b_1 \ \ldots \ b_n]^{-1}$ 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 11:

Let $\lambda_1, \ldots, \lambda_m$ be distinct nonobservable modes of $[Q_1, A]$ such that $\mathrm{Re}\{\lambda_i\} > 0$, $i = 1, \ldots, m$. Assume that X_1 is the nonnegative definite solution obtained by the eigenvectors corresponding to the eigenvalues of E with $\mathrm{Re}\{\lambda\} < 0$. 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

$$(A - B Q_2^{-1} B^T X_1)^T (X_1 - X_2) + (X_1 - X_2)(A - B Q_2^{-1} B^T X_1) =$$

$$= - (X_1 - X_2) B Q_2^{-1} B^T (X_1 - X_2)$$

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

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

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

In the previous example X_4 is the largest solution. Using 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 $\lambda_1, \ldots, \lambda_m$ all are included. In the example above X_1 is the smallest solution.

3. THE RICCATI EQUATION IN OPTIMAL CONTROL PROBLEMS

3.1. THE OPTIMAL CONTROL PROBLEM

Consider the linear time-invariant system

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

with the criteria

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

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

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

where $S(t_0)$ is a nonnegative definite symmetric solution of the matrix Riccati equation

$$-\frac{dS}{dt} = A^TS + SA - SBQ_2^{-1}B^TS + Q_1 \qquad S(t_f) = Q_0 \tag{3.3}$$

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

$$L(t) = Q_2^{-1}B^TS(t)$$

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

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

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

$$A^TS + SA - SBQ_2^{-1}B^TS + Q_1 = 0 \qquad (3.5)$$

Then S yields the solution to the optimal regulator problem. It is also shown that $S(t)$, $t \leq t_f$, is a continuous function of the boundary condition Q_0.

In [8] Wonham generalizes to $[A, B]$ being stabilizable and $[Q_1, A]$ being detectable. For an explanation of these concepts we refer to [8]. Then $S(t)$ converges towards a unique nonnegative definite solution S of (3.5), and $S(t)$, $t \leq t_f$, is still a continuous function of the boundary condition Q_0. In both cases the optimal closed loop system $A - BQ_2^{-1}B^TS$ will be asymptotic stable. In this section we will make a further generalization, and assume 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 holds [5], but according to section 2 there may be more than one nonnegative definite solution of the stationary Riccati equation (3.5).

We will then prove that the boundary condition Q_0 determines the stationary solution of $S(t)$ as $t \to -\infty$, and thus $S(t)$ is no longer a continuous function of Q_0.

For the numerical computation this implies that a straightforward integration of the Riccati equation in reversed time may be an unstable procedure.

The asymptotic dependence on Q_0 has a nice physical interpretation and this finally leads to a generalization of optimal control theory for linear systems with quadratic loss functions.

3.2. UPPER AND LOWER A PRIORI BOUNDS

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

$$u(t) = \hat{L}x(t)$$

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

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

$$\hat{\Psi}(t;t) = I$$

The corresponding cost is

$$\mathcal{J} = x^T(t)\hat{\Psi}^T(t_f;t)Q_0\hat{\Psi}(t_f;t)x(t) +$$

$$+ \int_t^{t_f} x^T(t)\hat{\Psi}^T(s;t)\left\{Q_1 + \hat{L}^T Q_2 \hat{L}\right\}\hat{\Psi}(s;t)x(t)ds$$

or

$$\mathcal{J} = x^T(t)\hat{S}(t)x(t)$$

where

$$\hat{S}(t) = \hat{\Psi}^T(t_f;t)Q_0\hat{\Psi}(t_f;t) + \int_t^{t_f} \hat{\Psi}^T(s;t)\left(Q_1 + \hat{L}^TQ_2\hat{L}\right)\hat{\Psi}(s;t)ds \qquad (3.6)$$

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

$$(A - B\hat{L})^T\hat{S} + \hat{S}(A - B\hat{L}) + Q_1 + \hat{L}^TQ_2\hat{L} = 0 \qquad (3.7)$$

Obviously $J^0 \leq \mathcal{J}$, and then $S(t) \leq \hat{S}(t)$, $t \leq t_f$. Then any linear feedback $\hat{L}$ such that $A - B\hat{L}$ 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 $Re(\lambda_i) < 0$, $i = 1 \ldots n$. Then

$$A^TS_1 + S_1A - S_1BQ_2^{-1}B^TS_1 + Q_1 = 0$$

and the closed loop system matrix $(A - BQ_2^{-1}B^TS_1)$ is asymptotic stable. Further, assume $S_2(t)$ is the solution of

$$-\frac{dS_2}{dt} = A^TS_2 + S_2A - S_2BQ_2^{-1}B^TS_2 + Q_1$$

with boundary condition

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

I is the identity matrix and α is a scalar.

$(S_2 - S_1)$ satisfy the differential equation

$$-\frac{d}{dt}(S_2 - S_1) = (A - BQ_2^{-1}B^TS_1)^T(S_2 - S_1) + (S_2 - S_1) \cdot$$

$$\cdot (A - BQ_2^{-1}B^TS_1) - (S_2 - S_1)BQ_2^{-1}B^T(S_2 - S_1) \quad (3.8)$$

with boundary condition

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

Now choose $\alpha > \| S_1 \|$ ($\alpha > \max_i \lambda_i$, where λ_i 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 $(A - BQ_2^{-1}B^TS_1)$. Then

$$\frac{\partial}{\partial t}\Psi(t;s) = (A - BQ_2^{-1}B^TS_1)\Psi(t;s)$$

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

and

$$\frac{\partial}{\partial s}\Psi(t;s) = -\Psi(t;s)(A - BQ_2^{-1}B^TS_1)$$

(3.8) is equivalent to the integral equation

$$(S_2 - S_1)(t) = \Psi^T(t_f;t)\Big\{(\alpha I - S_1)^{-1} +$$

$$+ \int_t^{t_f} \Psi^T(t_f;s)BQ_2^{-1}B^T\Psi(t_f;s)ds\Big\}^{-1}\Psi(t_f;t) \quad (3.9)$$

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

$$\left\{(\alpha I - S_1)^{-1} + \int_t^{t_f} \Psi^T(t_f;s)BQ_2^{-1}B^T\Psi(t_f;s)ds\right\}^{-1}$$

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

$$\left\{(\alpha I - S_1)^{-1} + \int_t^{t_f} \Psi^T(t_f;s)BQ_2^{-1}B^T\Psi(t_f;s)ds\right\}^{-1} \leq (\alpha I - S_1)$$

and

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

The fundamental matrix $\Psi(t_f;t) \to 0$ as $t \to -\infty$ since $(A - BQ_2^{-1}B^TS_1)$ is asymptotic stable, and then $(S_2 - S_1)(t) \to 0$ as $t \to -\infty$.

The solution of (3.3) with boundary condition

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

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

$$-\frac{dS_1}{dt} = A^TS_1 + S_1A - S_1BQ_2^{-1}B^TS_1 + Q_1$$

$$S_1(t_f) = Q_0$$

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

$$-\frac{dS_2}{dt} = A^TS_2 + S_2A - S_2BQ_2^{-1}B^TS_2 + Q_1$$

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

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

$$-\frac{d}{dt}(S_2 - S_1) = (A - BQ_2^{-1}B^TS_1)(S_2 - S_1) + (S_2 - S_1)(A - BQ_2^{-1}B^TS_1) -$$

$$- (S_2 - S_1)BQ_2^{-1}B^T(S_2 - S_1) \qquad (3.10)$$

$$(S_2 - S_1)(t_f) = \beta I - Q_0$$

With $\Psi(t;s)$ being the fundamental matrix associated with $(A - BQ_2^{-1}B^TS_1(t))$, (3.10) is equivalent to

$$(S_2 - S_1)(t) = \Psi^T(t_f;t)(\beta I - Q_0)\Psi(t_f;t) +$$

$$+ \int_t^{t_f} \Psi^T(s;t)(S_2(s) - S_1(s))BQ_2^{-1}B^T \cdot$$

$$\cdot (S_2(s) - S_1(s))\Psi(s;t)ds \qquad (3.11)$$

$(S_2 - S_1)(t)$ 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 à priori bound $\bar{S}(t)$, such that $S(t) \leq \bar{S}(t)$, $t \leq t_f$. $\bar{S}(t)$ 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 à 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) then follows that $S_2(t) \geq S_1(t)$, $t \leq t_f$. The smallest solution $\underline{S}(t)$ of (3.3), will then correspond to the boundary condition $\underline{S}(t_f) = 0$. $\underline{S}(t)$ is the solution of the integral equation

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

where

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

From (3.12) follows that $\underline{S}(t)$ is monotonic non-decreasing as $t \to -\infty$, and since the solutions are bounded, $\underline{S}(t)$ converges towards a solution of the stationary Riccati equation (3.5). This obviously is the smallest solution S', because assume that S(t) converges towards 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 à 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) towards a stationary solution of (3.5) as $t \to -\infty$, for arbitrary nonnegative definite boundary conditions Q_0.

3.3. CONVERGENCE PROPERTIES

The convergence of S(t) towards 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 case when all modes of A are unstable and nondetectable, that is $Q_1 = 0$. It will further be assumed that $Q_0 > 0$. These results may be combined with those of [6] and [8] to prove convergence in the general case.

As before, the differential equation

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

is transformed into an integral equation

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

where

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

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

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 $S^{-1}(t)$.

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

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

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 $\phi(t_f, t) \to 0$ as $t \to -\infty$, and

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

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 towards the unique positive definite solution of

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

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

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

as $t \to -\infty$. 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 $t \to -\infty$. Consider the equivalent integral equation

$$S(t) = \Psi^T(t_f;t)Q_0\Psi(t_f;t) +$$

$$+ \int_t^{t_f} \Psi^T(t_f;s)\left[Q_1 + S(s)BQ_2^{-1}B^TS(s)\right]\Psi(t_f;s)ds$$

where $\Psi(t;s)$ is the fundamental matrix associated with the closed loop system matrix $[A - BQ_2^{-1}B^TS(t)]$. When $t,s \to -\infty$

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

and

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

The latter condition holds since $S(t)$ is bounded, and means that the optimal system $A - BQ_2^{-1}B^TS$ cannot have unstable modes observable in Q_1.

Now let $\lambda_i > 0$ be a nonobservable mode of $[Q_1, A]$. Then stationary solutions S_1 and S_2 exist, such that $\lambda_i > 0$ is an eigenvalue of the closed loop system $A - BQ_2^{-1}B^TS_1$ and $-\lambda_i < 0$ of the system $A - BQ_2^{-1}B^TS_2$. From section 2.4 $S_2 \geq S_1$. If λ_i is a nonobservable mode of $[Q_0, A]$, there is no need to stabilize this mode since it will not affect the criteria. Then S_1 is the optimal stationary solution, and the optimal system will contain an unstable mode. However, if λ_i is observable of $[Q_0, A]$, S_1 cannot be the solution, since this could yield an infinite cost due to the term $x^T(t_f)Q_0x(t_f)$. Then $S(t) \to S_2$ as $t \to -\infty$.

The boundary condition Q_0 thus plays the same role as Q_1 to determine what stationary solution S(t) converges at.

In the general case, assume that A has r unstable eigenvalues $\lambda_1, \ldots, \lambda_r$, nonobservable in $[Q_1, A]$. If $\lambda_1, \ldots, \lambda_k$, $k < r$, are observable in $[Q_0, A]$, S(t) must converge towards a stationary solution S of (3.5), such that the optimal system $A - BQ_2^{-1}B^TS$ has eigenvalues $-\lambda_1, \ldots, -\lambda_k, \lambda_{k+1}, \ldots, \lambda_r$.

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 backwards in time. However, the existence of several stationary solutions may cause even the backwards integration to be an unstable process. This is illustrated in the following example [7].

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad Q_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad Q_2 = [1]$$

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

$$S_1 = \begin{bmatrix} 3+\sqrt{2} & 1+\sqrt{2} \\ 1+\sqrt{2} & 1+\sqrt{2} \end{bmatrix} \qquad S_2 = \begin{bmatrix} \sqrt{2}-1 & -\sqrt{2}+1 \\ -\sqrt{2}+1 & \sqrt{2}-1 \end{bmatrix}.$$

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 $S_1 \geq S_2$.

To make the solution S(t) converge towards S_2, the boundary condition $S(t_f) = 0$ is chosen according to section 3.2.

From (3.3) then follows that S(t) will have the structure

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

where $a(t) > 0$, $t < t_f$. 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

$$S(t_1) = \begin{pmatrix} a(t) + \varepsilon & -a(t) \\ -a(t) & a(t) \end{pmatrix}$$

where $\varepsilon > 0$ is a small quantity. $S(t_1)$ then is positive definite, and can be considered as boundary condition for further computation of $S(t)$, $t < t_1 < t_f$. But $[S(t_1), A]$ is completely observable and the solution will converge towards the largest solution S_1. This is illustrated in fig. 1, where the S_{11} element is plotted versus time. The disturbance 10^{-7} is introduced in the 1-1 element of Q_0, and a fourth order Runge-Kutta method is used for the integration [11].

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

$$S(t_1) = \begin{pmatrix} a(t) + \varepsilon & -a(t) + \varepsilon \\ -a(t) + \varepsilon & a(t) + \varepsilon \end{pmatrix}$$

For $a(t) > 0$ and $\varepsilon > 0$, $S(t_1)$ is positive definite, and $S(t)$ will converge towards S_1 as $t \to -\infty$. Another way to compute $S(t)$ is the fundamental matrix approach [5], [11]. With the computing method proposed in [11], the errors entered in the following manner.

$$S(t_1) = \begin{pmatrix} a(t) & -a(t) - \varepsilon \\ -a(t) - \varepsilon & a(t) \end{pmatrix}$$

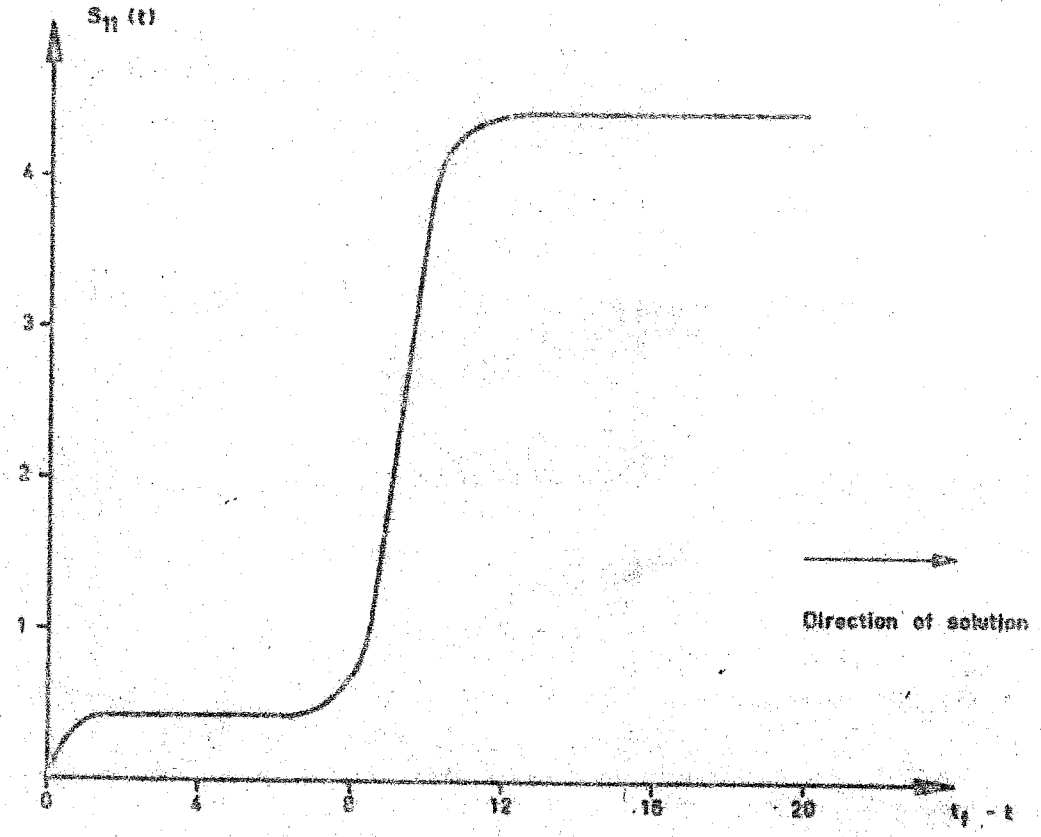

Fig. 1 - $S_{11}(t)$ computed with fourth order Runge-Kutta method.

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

For $\alpha > 0$, $S(t_f)$ is indefinite, and can no longer be considered as new boundary condition for further computation. However, computational experiments show that $S(t)$ still converges towards S_1, and the fundamental matrix method then can be considered as a stable method. The 1-1 element of the computed solution $S(t)$ is shown in fig. 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 occured, and the stationary solution S_1 was never reached.

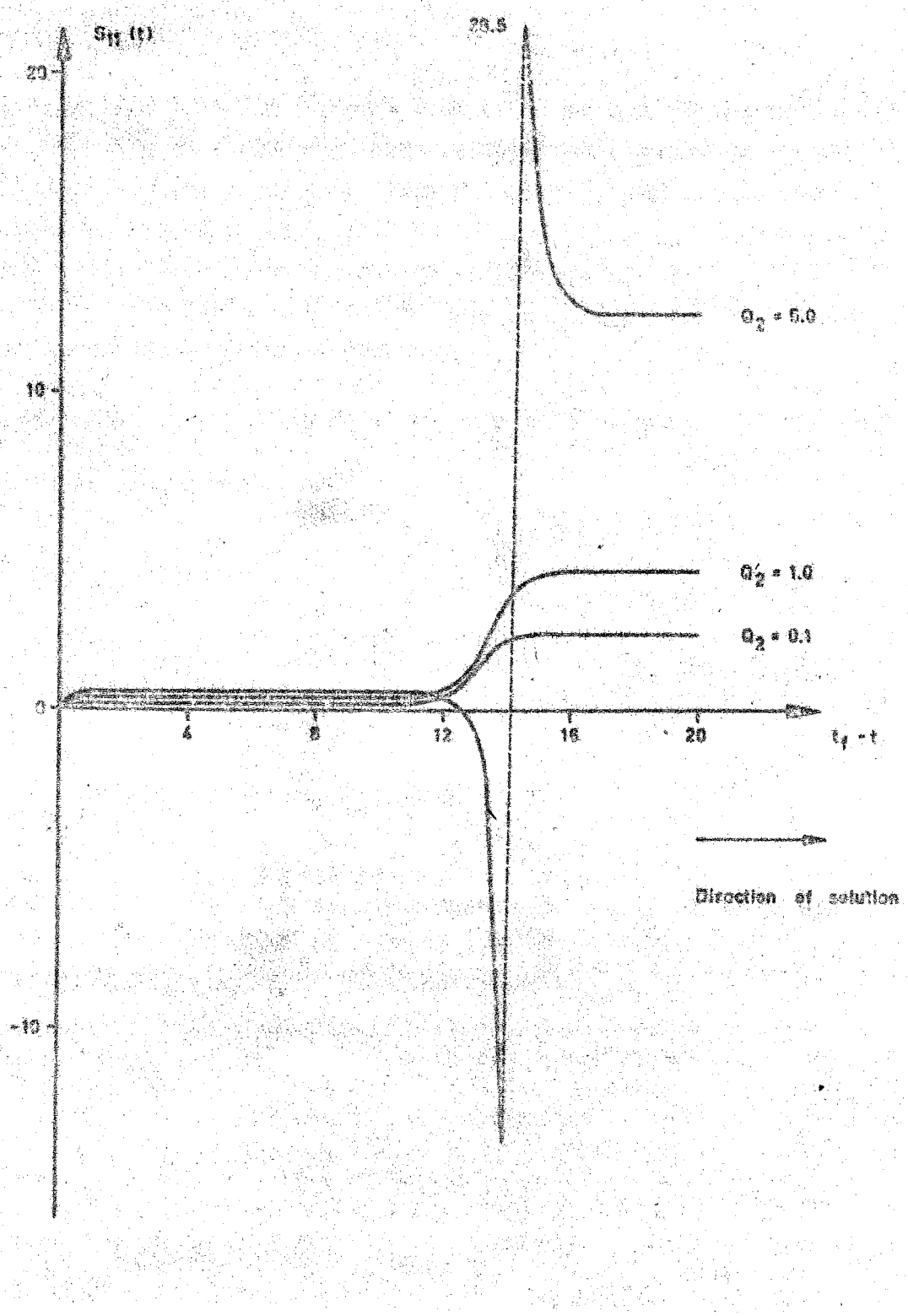

Fig. 2 - $S_{11}(t)$ computed with fundamental matrix method for various Q_2. $S_{11}(t)$ is plotted versus the time difference $t_f - t$.

3.5. GENERALIZATION OF OPTIMAL CONTROL THEORY FOR LINEAR SYSTEMS WITH QUADRATIC LOSS

The previous sections indicate a possible generalization of the optimal control theory for linear systems with quadratic loss. We then drop the requirement that the pair $[Q_1, A]$ should be observable, but it is still assumed that $[A, B]$ is stabilizable. Since asymptotic stability of the optimal system is a desired property, the minimizing control will be searched for in the class of stable linear feedback controls.

Theorem 12:

Consider the stabilizable system

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

with the loss function

$$V = \int_0^\infty (x^T(s)Q_1x(s) + u^T(s)Q_2u(s))ds$$

where Q_1 is nonnegative definite symmetric, and Q_2 positive definite symmetric. In the class of asymptotic stable linear feedback controls, the minimizing control is given by

$$u = -Q_2^{-1}B^TSx$$

where

$$S = \lim_{t \to \infty} S^*(t)$$

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

$$-\frac{dS^*}{dt} = A^TS^* + S^*A - S^*BQ_2^{-1}B^TS^* + Q_1$$

with boundary condition

$$S^*(t_f) = I$$

(t_f is arbitrary)

Proof:

From 3.3 follows that $S^*(t)$ converges towards 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 $u = -Lx$, $L_m = Q_2^{-1}B^TS_m$ yields the minimum value of the loss function V.

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

$$V = x^T(0)S_1x(0)$$

where

$$S_1 = \int_0^\infty e^{(A-BL_1)^Ts}\left\{Q_1 + L_1^TQ_2L_1\right\}e^{(A-BL_1)s}\,ds \qquad (3.16)$$

is nonnegative definite symmetric.

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

$$(A - BL_1)^TS_1 + S_1(A - BL_1) + Q_1 + L_1^TQ_2L_1 = 0 \qquad (3.17)$$

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

$$(A - BL_m)^TS_m + S_m(A - BL_m) + Q_1 + L_m^TQ_2L_m = 0 \qquad (3.18)$$

This is equivalent to

$$(A - BL_1)^T S_m + S_m(A - BL_1) + Q_1 + L_1^T B^T S_m + S_m BL_1 -$$

$$- L_m^T B^T S_m - S_m BL_m + L_m^T Q_2 L_m = 0 \qquad (3.19)$$

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

$$(A - BL_1)^T (S_1 - S_m) + (S_1 - S_m)(A - BL_1) + L_1^T Q_2 L_1 -$$

$$- L_1^T Q_2 L_m - L_m^T Q_2 L_1 + L_m^T Q_2 L_m = 0 \qquad (3.20)$$

or

$$(A - BL_1)^T (S_1 - S_m) + (S_1 - S_m)(A - BL_1) + (L_1 - L_m)^T Q_2 \cdot$$

$$\cdot (L_1 - L_m) = 0$$

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

$$\int_0^\infty u^T(s)Q_2u(s)ds$$

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 to find an asymptotic stable linear feedback $u = -Lx$ for the system

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

which minimizes

$$V = \int_0^\infty u^TQ_2u$$

Q_2 is positive definite symmetric, and V 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 $\lambda_1, \ldots, \lambda_k$ such that $Re(\lambda) > 0$, and $\lambda_{k+1}, \ldots, \lambda_n$ with $Re(\lambda) < 0$. Since $Q_1 = 0$, $\lambda_1, \ldots, \lambda_k$ are nondetectable, and then E has the eigenvalues $\pm\lambda_1, \ldots, \pm\lambda_k, \pm\lambda_{k+1}, \ldots, \pm\lambda_n$, independent of Q_2. The optimal stable system thus has the

eigenvalues $-\lambda_1, \ldots, -\lambda_k, \lambda_{k+1}, \ldots, \lambda_n$. This can be formulated as some kind of minimum energy principle.

Theorem 13:

Consider the system

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

where A has eigenvalues $\lambda_1, \ldots, \lambda_k$ such that $\mathrm{Re}\{\lambda\} > 0$, and $\lambda_{k+1}, \ldots, \lambda_n$ with $\mathrm{Re}\{\lambda\} < 0$. The minimum energy regulator $u = -Lx$ then has the property that the eigenvalues of the closed loop system are $-\lambda_1, \ldots, -\lambda_k, \lambda_{k+1}, \ldots, \lambda_n$.

Notice that the feedback L is independent of any specific choice of the positive definite criteria matrix Q_2.

4. REFERENCES

[1] Potter: Matrix Quadratic Solutions, J. SIAM Appl. Math., Vol. 14, No. 3, May, 1966.

[2] Friedman: Principles and Techniques of Applied Mathematics, John Wiley, New York, 1965.

[3] Åström: Lecture Notes on Optimal Control Theory, Lund Inst. of Techn., Div. of Automatic Control, 1965.

[4] Kalman, Englar: A User's Manual for the Automatic Synthesis Program, NASA report CR-475, 1966.

[5] Kalman: Contributions to the Theory of Optimal Control, Bol. Soc. Mat. Mex., 1960.

[6] Bucy: Global Theory of the Riccati Equation, J. of Computer and System Sciences, Vol. 1, No. 4, Dec., 1967.

[7] Kalman: When is a Linear Control System Optimal?, JACC, 1963.

[8] Wonham: On Matrix Quadratic Equations and Matrix Riccati Equations, Techn. rep. 67-5, Div. of Appl. Math., Brown University.

[9] Wonham: On Pole Assignment in Multi-Input Controllable Linear Systems, Techn. rep. 67-2, Div. of Appl. Math., Brown University.

[10] Beckenbach, Bellman: Inequalities, Springer Verlag, New York, 1965.

[11] Mårtensson: Linear Quadratic Control Package, Part I - The Continuous Problem, Res. rep. 6802, Lund Inst. of Techn., Div. of Automatic Control.