Vladimír Kučera

A review of the matrix Riccati equation

Kybernetika, Vol. 9 (1973), No. 1, (42)--61

Persistent URL: http://dml.cz/dmlcz/124651

KYBERNETIKA — VOLUME 9 (1973), NUMBER 1

A Review of the Matrix Riccati Equation

VLADIMÍR KUČERA

This paper reviews some basic results regarding the matrix Riccati equation of the optimal control and filtering theory. The theoretical exposition is divided into three parts dealing respectively with the steady-state algebraic equation, the differential equation, and the asymptotic properties of the solution. At the end a survey of existing computational techniques is given.

INTRODUCTION

As usual, $\boldsymbol{R}$ denotes the field of real numbers, $\boldsymbol{R}^n$ stands for the n-dimensional vector space over $\boldsymbol{R}$, a prime denotes the transpose of a matrix, an asterisk denotes the complex conjugate transpose of a matrix, and $P \geqq Q$ means that $P - Q$ is hermitian or real symmetric nonnegative matrix. Square brackets represent matrices composed of the symbols inside.

In order to get a better motivation for the problems to be discussed we first pose the underlying physical problem.

Given the linear, continuous-time, constant system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = A\,x(t) + B\,u(t)\,, \quad x(t_0) = x_0\,, \tag{1}$$

$$y(t) = H\,x(t)\,, \tag{2}$$

where $x \in \boldsymbol{R}^n$, $u \in \boldsymbol{R}^r$, and $y \in \boldsymbol{R}^p$ are the state, the input, and the output of the system respectively and A, B, H are constant matrices over $\boldsymbol{R}$ of appropriate dimensions, find a control $u(t)$ over $t_0 \leqq t \leqq t_f$ which for any $x_0 \in \boldsymbol{R}^n$ minimizes the cost functional

$$\mathscr{J} = \tfrac{1}{2}x'(t_f)\,S\,x(t_f) + \tfrac{1}{2}\int_{t_0}^{t_f}(x'Qx + u'u)\,\mathrm{d}t\,. \tag{3}$$

with $S \geqq 0$, $Q \geqq 0$.

This problem is referred to as the *least squares* optimal control problem and it can be solved by the minimum principle of Pontryagin [1], [19], [29], [32], by the dynamic programming of Bellman [1], [3], [7], [15], [19], [32] or by the second method of Lyapunov [33].

The minimum value $\mathscr{J}_0$ of (3) is given as

$$\mathscr{J}_0 = \tfrac{1}{2}x'(t_0)\, P(t_0)\, x(t_0) \tag{4}$$

and it is attained if and only if the control

$$u(t) = -B'\, P(t)\, x(t) \tag{5}$$

is used. Here P is an $n \times n$ matrix solution of the Riccati differential equation

$$-\frac{dP}{dt} = -P(t)\, BB'\, P(t) + P(t)\, A + A'\, P(t) + Q\,, \tag{6}$$

$$P(t_f) = S\,.$$

Note that this equation must be solved *backward* from t_f to t_0 in order to obtain the optimal control.

One special case is frequent in applications, namely $t_f \to \infty$, the so called *regulator problem*. In this particular case it may happen that $P(t)$ approaches a finite constant, P_∞, as $t_f \to \infty$, or, equivalently, as $t \to -\infty$ in (6). Then

$$\mathscr{J}_0 = \tfrac{1}{2}x'(t_0)\, P_\infty\, x(t_0)\,; \tag{7}$$

the control law

$$u(t) = -B'P_\infty\, x(t) \tag{8}$$

is independent of time and P_∞ satisfies the quadratic *algebraic* equation

$$-PBB'P + PA + A'P + Q = 0\,. \tag{9}$$

In the sections to follow we first investigate the algebraic equation (9), then the differential equation (6) and the asymptotic behaviour of the solution of (6) as $t \to -\infty$. Finally some computing techniques for both equation (6) and (9) are surveyed.

THE QUADRATIC EQUATION

The matrix equation (9) has been extensively studied [6], [16], [22], [23], [26], [30], [36]. It is well-known that it can possess a variety of solutions. First of all (9) may have no solution at all. If it does have one, there can be both real and complex solutions, some of them being hermitian or symmetric. There can be even infinitely many solutions. Due to the underlying physical problem, however, only nonnegative

solutions are of interest to us. Therefore, we are mainly concerned with the existence and uniqueness of such a solution.

In this section we summarize some long-standing as well as recent results [22], [23], [26], [30] on (9) which will prove useful later. First of all, write

$$Q = C'C\,, \quad S = D'D\,.$$

Then λ is said to be an *uncontrollable eigenvalue* [13], [22] of the pair (A, B) if there exists a row vector $w \neq 0$ such that $wA = \lambda w$ and $wB = 0$. Similarly, λ is an *unobservable eigenvalue* of the pair (C, A) if there exists a vector $z \neq 0$ such that $Az = \lambda z$ and $Cz = 0$.

The pair (A, B) is said to be *stabilizable* [35] if a matrix L over $\boldsymbol{R}$ exists such that $A + BL$ is stable (i.e., all its eigenvalues have negative real parts), or, equivalently, if the unstable eigenvalues of (A, B) are controllable [13], [35].

Analogically, the pair (C, A) is defined to be *detectable* [35] if a matrix F over $\boldsymbol{R}$ exists such that $FC + A$ is stable, or, if the unstable eigenvalues of (C, A) are observable [13], [35].

A nonnegative solution of (9) is said to be an *optimizing solution* [28] if it yields the optimal control (8); it is called a *stabilizing solution* [28] if the control (8) is stable. We shall denote these solutions P_o and P_s, respectively.

Further we introduce the $2n \times 2n$ matrix

$$M = \begin{bmatrix} A, & -BB' \\ -C'C, & -A' \end{bmatrix}. \tag{10}$$

Unless otherwise stated we shall henceforth assume that the M matrix is diagonalizable, that is, it has $2n$ eigenvectors. This assumption is made for the sake of simplicity and is by no means essential.

Let

$$Ma_i = \lambda_i a_i\,, \quad r_i M = \lambda_i r_i\,, \quad i = 1, 2, \ldots, 2n\,,$$

and write

$$a_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad r_i' = \begin{bmatrix} u_i \\ v_i \end{bmatrix},$$

where $x_i \in \boldsymbol{R}^n$, $y_i \in \boldsymbol{R}^n$, $u_i \in \boldsymbol{R}^n$ and $v_i \in \boldsymbol{R}^n$.

Thus the a_i is a column vector whereas the r_i is a row vector. They are sometimes called the right and the left eigenvectors of M, respectively.

It is well-known that the eigenvectors can be chosen so that

$$\begin{aligned} r_i a_j &= 0\,, \quad i \neq j\,, \\ &\neq 0\,, \quad i = j\,. \end{aligned} \tag{11}$$

The following seems to have been proved first in [10], [26] and [30].

Theorem 1. *Each solution P of* (9) *has the form*

$$P = YX^{-1}, \tag{12}$$

where

$$X = [x_1, x_2, \ldots, x_n],$$
$$Y = [y_1, y_2, \ldots, y_n]$$

correspond to such a choice of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of M that X^{-1} exists.

Conversely, all solutions are generated in this way.

Proof. Let P satisfies (9) and set

$$K = A - BB'P,$$

the closed-loop system matrix. Then we infer from (9) that

$$PK = -Q - A'P$$

and hence

$$M\begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} K. \tag{13}$$

Let

$$J = X^{-1}KX = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

be the Jordan canonical form of K and set $PX = Y$. Then (13) yields

$$M\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} J. \tag{14}$$

Since J is diagonal, the columns of $\begin{bmatrix} X \\ Y \end{bmatrix}$ constitute the eigenvectors of M associated with $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $P = YX^{-1}$.

The converse can be proved by reversing the arguments. □

Corollary. *The matrix $K = A - BB'P$ given by the solution* (12) *has the eigenvalues λ_i associated with the eigenvectors x_i, $i = 1, 2, \ldots, n$.*

Proof. The J matrix is the Jordan form of K and X is the associated transformation matrix. □

Theorem 2. *Let λ_i be an eigenvalue of M and $\begin{bmatrix} x_i \\ y_i \end{bmatrix}$ the corresponding right eigenvector. Then $-\lambda_i$ is an eigenvalue of M and $\begin{bmatrix} -y_i \\ x_i \end{bmatrix}'$ the corresponding left eigenvector.*

Proof. By direct verification, making use of the identity

$$M' = \begin{bmatrix} 0, & -I \\ I, & 0 \end{bmatrix} M \begin{bmatrix} 0, & -I \\ I, & 0 \end{bmatrix}$$

and the fact that a left eigenvector of M associated with λ is a transposed right eigenvector of M' associated with λ. □

Note that M being a real matrix, its eigenvalues occur in quadruples $(\lambda, \lambda^*, -\lambda, -\lambda^*)$.

The procedure described in Theorem 1 is the time-domain counterpart of the spectral factorization in the frequency domain.

We also point out that there can exist at most one stabilizing solution due to the symmetry of the eigenvalues of M. If $P_s = YX^{-1}$ is such a solution, then the matrix X^*Y is hermitian [30].

The following fundamental theorems originated in [22], [23], [24].

Theorem 3. *The stabilizing solution of* (9) *exists if and only if* (A, B) *is stabilizable and* $\operatorname{Re} \lambda \neq 0$ *for all eigenvalues* λ *of* M.

Proof. Necessity: Suppose the stabilizing solution P_s exists. It is associated with n stable eigenvalues of M and hence no eigenvalue can have zero real part due to their symmetrical distribution about the imaginary axis.

Moreover, the matrix $L = -B'P_s$ stabilizes $A + BL$, i.e., the pair (A, B) is stabilizable.

Sufficiency: Suppose the hypothesis holds and let no stabilizing solution exist. Then either (i) we cannot choose n stable eigenvalues of M, a contradiction, or (ii) we can do so but the X matrix in (14) is singular.

If (ii) is the case, write z for any nonzero vector of $\mathcal{N}(X)$, the null space of X. Since $X^*Y = Y^*X$, we have

$$0 = Y^*Xz = X^*Yz\,. \tag{15}$$

By virtue of (10) and (14),

$$\begin{aligned} AX - BB'Y &= XJ\,, \\ -C'CX - A'Y &= YJ\,. \end{aligned} \tag{16}$$

The first equation yields

$$z^*Y^*AXz - z^*Y^*BB'Yz = z^*Y^*XJz$$

and hence by (15) and by the definition of z

$$B'Yz = 0\,. \tag{17}$$

But

$$0 = AXz - BB'Yz = XJz\,,$$

which means that $\mathcal{N}(X)$ is a J-invariant subspace of $\boldsymbol{R}^n$. Hence there exists at least one nonzero vector $\hat{z} \in \mathcal{N}(X)$ such that

$$J\hat{z} = \mu\hat{z} \tag{18}$$

where μ coincides with one of the stable eigenvalues of M.

The second equation (16) postmultiplied by $\hat{z}$ yields

$$-A'Y\hat{z} = YJ\hat{z}\,. \tag{19}$$

Collecting (17) through (19) gives us

$$(Y\hat{z})'\,A = -\mu(Y\hat{z})', \quad \operatorname{Re}(-\mu) > 0\,,$$
$$(Y\hat{z})'\,B = 0\,.$$

Thus (A, B) is not stabilizable, contradicting our hypothesis. □

Theorem 4. *The stabilizing solution is the only nonnegative solution of* (9) *if and only if* (C, A) *is detectable.*

Proof. Sufficiency: Assuming (C, A) detectable we shall demonstrate that any solution P of (9) is the stabilizing solution. Suppose to the contrary that a λ exists such that

$$Kz = \lambda z\,, \quad \operatorname{Re}\lambda \geqq 0\,, \quad K = A - BB'P\,.$$

On rearranging equation (9) reads

$$PBB'P + PK + K'P + C'C = 0$$

and hence

$$(\lambda + \lambda^*)\,z^*Pz = -z^*PBB'Pz - z^*C'Cz\,.$$

Since $\lambda + \lambda^* = 2\operatorname{Re}\lambda \geqq 0$, the left hand side of this equation is nonnegative, while the right hand side is nonpositive. Therefore either is zero and

$$B'Pz = 0\,,$$
$$Cz = 0\,,$$

which in turn implies

$$Az = \lambda z\,, \quad \operatorname{Re}\lambda \geqq 0\,,$$
$$Cz = 0\,.$$

Thus (C, A) is not detectable contradicting our hypothesis. Hence K is stable.

However, there is at most one stabilizing solution, i.e., the solution P is unique. For another proof refer to [36].

Necessity: By contradiction, suppose an undetectable eigenvalue λ_1 of (C, A)

exists. We shall show the existence of at least two nonnegative solutions of equation (9). One of them is the stabilizing solution $P_s = YX^{-1}$ by hypothesis.

To form another solution $P_1 = Y_1X_1^{-1}$ we substitute the eigenvector $\begin{bmatrix} z_1 \\ 0 \end{bmatrix}$ of M associated with λ_1 for the eigenvector $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ associated with $-\lambda_1$, thus obtaining

$$X_1 = [z_1, x_2, \ldots, x_n], \quad Y_1 = [0, y_2, \ldots, y_n].$$

Also set

$$\hat{X} = [x_2, \ldots, x_n], \quad \hat{Y} = [y_2, \ldots, y_n].$$

Now Theorem 2 along with (11) implies that

$$[z_1^*, 0^*]\begin{bmatrix} -y_i \\ x_i \end{bmatrix} = 0, \quad i = 2, 3, \ldots, n.$$

Hence $z_1^*\hat{Y} = 0$ and

$$X_1^*Y_1 = \begin{bmatrix} 0, & z_1^*\hat{Y} \\ 0, & \hat{X}^*\hat{Y} \end{bmatrix} = \begin{bmatrix} 0, & 0 \\ 0, & \hat{X}^*\hat{Y} \end{bmatrix} \geqq 0 \tag{20}$$

because $X^*Y \geqq 0$.

To prove that X_1 is nonsingular, suppose the contrary is true. Then a vector $v \neq 0$ exists such that $z_1 = \hat{X}v$ and, consequently,

$$0 = z_1^*\hat{Y} = v^*\hat{X}^*\hat{Y};$$

that is, $\det \hat{X}^*\hat{Y} = 0$. Observing that $\det \hat{X}^*\hat{Y}$ is a principal minor of the nonnegative matrix X^*Y, it is easy to see that

$$[x_1, \hat{X}]^*\hat{Y}v = X^*(\hat{Y}v) = 0.$$

This is a contradiction, however, as X is nonsingular and $\hat{Y}v \neq 0$.

Thus P_1 does exist and is different from P_s because it corresponds to a different $n-$tuple of eigenvalues of M. □

Theorem 5. *Stabilizability of (A, B) and detectability of (C, A) is necessary and sufficient for equation (9) to have a unique nonnegative solution which stabilizes the closed-loop system.*

Proof. Sufficiency part is a well-established result in [36]; it can also be inferred from Theorems 3 and 4.

Necessity part is a simple consequence of Theorems 3 and 4. □

We note that optimality does not necessarily imply stability. Indeed, the optimizing solution minimizes the cost functional whereas the stabilizing solution makes the closed-lopp stable, and this is quite a different property. However, in control applications an optimal system which is stable as well is desirable. Conditions for this case to hold have been stated in Theorem 5, first published in [22].

Further we turn our attention to the situation when these conditions are not in force. If (C, A) is detectable and (A, B) is not stabilizable, no nonnegative solution of (9) exists. Indeed, (C, A) detectable implies by Theorem 4 that all solutions are the stabilizing solutions. But such a solution does not exist by Theorem 3.

Further let (A, B) be stabilizable and (C, A) be not necessarily detectable. In this case we characterize all nonnegative solutions of (9), see [23], [24].

Let $\lambda_1, \lambda_2, \ldots, \lambda_p$, $p \geqq 0$, be those eigenvalues of M that are also the undetectable eigenvalues of (C, A), i.e.,

$$\begin{aligned} &Az_i = \lambda_i z_i\,, \\ &Cz_i = 0\,. \\ &\operatorname{Re}\lambda_i \geqq 0\,, \quad i = 1, 2, \ldots, p\,. \end{aligned}$$

An eigenvalue λ of A is called *cyclic* if any two eigenvectors of A associated with λ are linearly dependent. We restrict ourselves to the cyclic λ_i, $i = 1, 2, \ldots, p$. Under this condition equation (9) has only a finite number of nonnegative solutions [23], [24]. Otherwise nondenumerable many nonnegative solutions exist as the example $A = B = I$, $C = 0$, $n > 1$ shows, see [23], [24] for details.

Define

$$\mathscr{C} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \tag{21}$$

and write $\mathscr{C}_k$, $k = 1, 2, \ldots$ for the subsets of $\mathscr{C}$. Note that all $-\lambda_i$, $i = 1, 2, \ldots, p$ must be used to form the stabilizing solution if it exists, since $\operatorname{Re}(-\lambda_i) \leqq 0$.

Write P_k for the solution of (9) which is generated from the stabilizing solution by replacing all $-\lambda_i$ with λ_i, $\lambda_i \in \mathscr{C}_k$.

The theorems below have originally been proven in [23], [24].

Theorem 6. *Suppose the stabilizing solution of equation* (9) *exists. Then the set of all P_k form the class of all nonnegative solutions of* (9), *and there are exactly 2^p such solutions.*

Proof. Existence of all P_k can be proved in an identical manner as the existence of P_1 in the "if" part of the proof of Theorem 4.

To see that no other nonnegative solution of (9) exists, firstly substitute an eigenvalue $\lambda \in \mathscr{C}$ for the eigenvalue $-\lambda$. Arguing as in [23] we conclude that either $X^*Y \ngeqq 0$ or X is singular and hence the corresponding solution cannot be nonnegative.

Secondly make the substitution of an eigenvalue λ, $\operatorname{Re}\lambda \geqq 0$, for an eigenvalue μ, $\operatorname{Re}\mu < 0$, $-\mu \neq \lambda$. Applying (11) and Theorem 2 we conclude that $X^*Y \neq Y^*X$ and such a solution cannot be nonnegative, either.

The number of solutions follows from the fact that there are 2^p subsets of a set consisting of p elements. □

Remark. If the M matrix is not diagonalizable, a more refined analysis in [23]

shows that, in general, there exist

$$\prod_{i=1}^{p}(1+s_i)$$

nonnegative solutions, where

$$\begin{aligned} Cz_{i+j} &= 0\,, \quad j = 0, 1, \ldots, s_i - 1\,, \\ &\neq 0\,, \quad j = s_i \end{aligned}$$

and

$$\begin{aligned} Az_{i+j} &= \lambda_i z_i\,, && j = 0\,, \\ &= \lambda_i z_{i+j} + z_{i+j-1}\,, && j = 1, 2, \ldots, s_i - 1\,. \end{aligned}$$

Theorem 7. *Any two nonnegative solutions P_k, P_l of* (9) *satisfy*

$$P_k \geqq P_l \quad \textit{if and only if} \quad \mathscr{C}_k \subseteq \mathscr{C}_l\,. \tag{22}$$

Otherwise speaking, the set of all nonnegative solutions constitute a distributive lattice with respect to the partial ordering $\geqq$. P_o is the smallest nonnegative solution (the zero element of the latice) and P_s is the largest nonnegative solution (the identity element of the lattice).

Proof. The claim hinges on the observation that the family of subsets $\mathscr{C}_k$ constitute a distributive lattice with respect to the partial ordering by inclusion and that (22) is an isomorphism.

In particular, $\mathscr{C}_k = \emptyset$ implies $P_k = P_s$ whereas $\mathscr{C}_k = \mathscr{C}$ implies $P_k = P_o$.

The relation (22) itself follows from (20). □

Note that if (C, A) is detectable then $\mathscr{C} = \emptyset$ and hence $P_o = P_s$. However, (C, A) undetectable implies that $\mathscr{C} \neq \emptyset$ and hence $P_o \neq P_s$; the two solutions never coincide:

The physical interpretation of different real nonnegative solutions is as follows [26]. Each nonnegative solution is a conditionally optimizing solution of (9), the condition being a certain degree of stability. Specifically, P_k stabilizes the undetectable eigenvalues of (C, A) included in $\mathscr{C}_k$ and no others. Equation (9) thus contains the optimal solutions for all degrees of stability [23], [24], [26]. The idea that the more undetectable eigenvalues is stabilized the higher is the cost $\mathscr{J}$ is made rigorous via the concept of lattice.

THE DIFFERENTIAL EQUATION

In this section we summarize some fundamental results on the Riccati equation (6) which are scattered in the literature [1], [6], [8], [9], [15], [17], [19], [20], [32], [36].

The well-known theorem on differential equations guarantees only *local* existence and uniqueness for the solution of (6); without further analysis we cannot conclude existence over $t_0 \leqq t \leqq t_f$ because of the phenomenon of finite escape time.

Nonetheless, the following is proved in [6], [8], [9], [15], [36].

Theorem 8. *Let $P(t, S, t_f)$ be the solution of* (6) *which passes through S at $t = t_f$.*
Then given any t_0, $P(t, S, t_f)$ exists and is unique on $t_0 \leqq t \leqq t_f$ regardless of S.

Proof. In view of the local existence the $P(t, S, t_f)$ exists for some $t \leqq t_f$.
Let $\Psi(t, t_f)$ be the transition matrix of A. Then

$$P(t, S, t_f) = \Psi'(t, t_f)\, S\, \Psi(t, t_f) + \\ + \int_t^{t_f} \Psi'(t, \tau) \left[Q - P(\tau, S, t_f)\, BB'\, P(\tau, S, t_f)\right] \Psi(t, \tau)\, \mathrm{d}\tau \leqq \\ \leqq \Psi'(t, t_f)\, S\, \Psi(t, t_f) + \int_t^{t_f} \Psi'(t, \tau)\, Q\, \Psi(t, \tau)\, \mathrm{d}\tau\,,$$

which can be verified by elementary differentiation. The a priori upper bound above provides a Lipschitz constant on any finite interval $t_0 \leqq t \leqq t_f$ no matter how large. Hence $P(t, S, t_f)$ exists and is unique *globally* for any S. □

Theorem 9. *For any $t_0 \leqq t \leqq t_f$ and any $S \geqq 0$,*

$$P(t, S, t_f) \geqq 0\,. \tag{23}$$

Proof. If $P(t, S, t_f)$ satisfies (6), then $P^*(t, S, t_f)$ does so. Hence $P(t, S, t_f)$ is a hermitian matrix. Moreover, using (4), we obtain (23) since

$$\mathscr{J}_o = \min_u \mathscr{J} \geqq 0\,.$$ □

Theorem 10. *The solution P of* (6) *enjoys the following properties*:

$$P(t_0, 0, t_1) \leqq P(t_0, 0, t_2) \tag{24}$$

for any $t_0 \leqq t_1 \leqq t_2$,

$$P(t_2, 0, t_f) \leqq P(t_1, 0, t_f) \tag{25}$$

for any $t_1 \leqq t_2 \leqq t_f$,

$$P(t, S_1, t_f) \leqq P(t, S_2, t_f) \tag{26}$$

for any nonnegative $S_1 \leqq S_2$ and $t \leqq t_f$.

Proof. Regarding (24), we have

$$x_0'\, P(t_0, 0, t_1)\, x_0 = \min_u \int_{t_0}^{t_1} (x'Qx + u'u)\, \mathrm{d}t \leqq \\ \leqq \min_u \int_{t_0}^{t_2} (x'Qx + u'u)\, \mathrm{d}t = \\ = x_0'\, P(t_0, 0, t_2)\, x_0$$

for any $x_0 \in \boldsymbol{R}^n$.

Regarding (25), the substitution $t_f - t = \tau$ in (3), gives us

$$P(t_0, 0, t_f) = P(0, 0, t_f - t_0)$$

and hence (25) follows from (24) on identifying t_0 with t_1 and t_2 in turn.

Finally, (26) is obvious since increasing the terminal penalty results in increased cost functional (3).

We conclude this section by emphasizing that we are not concerned with stability of the optimal closed-loop system since the control problem is defined over a finite interval.

ASYMPTOTIC BEHAVIOUR OF THE SOLUTION

Now we let $t_f \to \infty$ and investigate the properties of $P(t)$ over the half-line $t_0 \leqq t < \infty$. Since the Riccati equation coefficients are independent of time, $P(t_0, S, t_f) = P(0, S, t_f - t_0)$ and it makes no difference to consider $t \to -\infty$ and the half-line $-\infty < t \leqq t_f$ instead. The first arrangement reflects the time evolution of the system, whereas the other is convenient from the computational point of view.

It is easy to see that additional assumptions will be required to keep $P(t)$ from passing off to infinity.

Theorem 11. *If (A, B) is stabilizable, then the solution $P(t, S, t_f)$ of (6) is bounded on $-\infty < t \leqq t_f$ regardless of S and t_f.*

Proof. Consider a control

$$\hat{u}(t) = L\hat{x}(t), \quad t_0 \leqq t,$$

such that the closed-loop system matrix $A + BL$ is stable. Then

$$x_0' P(t_0, S, t_f) x_0 = \min_u \left[x'(t_f) S x(t_f) + \int_{t_0}^{t_f} (x'Qx + u'u)\, dt\right] \leqq$$

$$\leqq \int_{t_0}^{\infty} (\hat{x}'Q\hat{x} + \hat{u}'\hat{u})\, dt = x_0' W(t_0, t) x_0$$

since $\hat{x}(t_f) \to 0$ as $t_f \to \infty$, and $x_0 = \hat{x}_0$.

Here $W(t_0, t)$ is a finite matrix over $\boldsymbol{R}$ for any $t \leqq t_f$. Hence $P(t_0, S, t)$ is bounded from above for any t_0, any $t_0 \leqq t < \infty$, and any S.

Equivalently, $P(t, S, t_f)$ is bounded for any t_f, any $-\infty < t \leqq t_f$, and any S. □

Theorem 11 is proved in [36] without any appeal to variational ideas. Our proof, however, is considerably simpler and more intuitive.

First the asymptotic properties of $P(t, 0, t_f)$ will be discussed and existence of an equilibrium solution of (6) deduced, see [15], [36].

Theorem 12. *If* (A, B) *is stabilizable, then*

$$\lim_{t \to -\infty} P(t, 0, t_f) = P_\infty ,$$

a finite nonnegative matrix.

Moreover, P_∞ *is a fixed point solution of* (6),

$$P_\infty = P(t, P_\infty, t_f) .$$

Proof. By Theorem 10, (25), $P(t, 0, t_f)$ is monotone nondecreasing in t in the ordering of nonnegative matrices. Further Theorem 11 implies that $P(t, 0, t_f)$ is bounded from above and consequently there exists a finite matrix $P_\infty \geqq 0$ so that $P(t, 0, t_f) \to P_\infty$ as $t \to -\infty$.

Since P_∞ is independent of t, it satisfies both (6) and (9), that is, $P_\infty = P(t, P_\infty, t_f)$. □

We remark that even if the stabilizing solution P_s of (9) exists, $P_\infty = P_s$ need not be true. Naturally, there arises the question: To which nonnegative solution of (9) does $P(t, 0, t_f)$ converge? The answer is contained in the two theorems below [26], [36].

Theorem 13. *If* (A, B) *is stabilizable, then*

$$P_\infty = P_o .$$

Proof. Since $S \geqq 0$, we infer from Theorem 10, (26) that $P(t, S, t_f) \geqq P(t, 0, t_f)$ for any S. Thus, as $t \to -\infty$, $P(t, 0, t_f) \to P_o$, the smallest nonnegative solution of (9). □

Theorem 14. *If* (A, B) *is stabilizable and* (C, A) *is detectable, then*

$$P_\infty = P_o = P_s .$$

Proof. By hypothesis Theorem 5 implies that equation (9) possesses the unique nonnegative solution $P_o = P_s$. Thus the claim follows from Theorem 13. □

Now we generalize and consider the asymptotic behaviour of $P(t, S, t_f)$ rather than of $P(t, 0, t_f)$. Theorem 12 is not known to hold for arbitrary nonzero $S \geqq 0$ as yet, even though some conjectures so that effect has been made [26]. The difficulty stems from the fact that $P(t, S, t_f)$ is not monotone nondecreasing in t.

We avoid the impasse by assuming that not only (A, B) is stabilizable but also $\operatorname{Re} \lambda \neq 0$ for all eigenvalues λ of M. This is equivalent to the assumption that P_s exists.

Thus we are ready to prove the original results of this paper.

Theorem 15. *If P_s exists then*

$$\lim_{t \to -\infty} P(t, S, t_f) = P_\infty ,$$

a finite nonnegative matrix.

Moreover, P_∞ is a fixed point solution of (6),

$$P_\infty = P(t, P_\infty, t_f) .$$

Proof. For $S = 0$ the theorem reduces to Theorem 12. Thus it will suffice to show that the term $x'(t_f)\, S\, x(t_f)$ approaches zero as $t_f \to \infty$.

By the assumption of $\operatorname{Re} \lambda \neq 0$ this term goes either to zero or to infinity. The latter case is not possible, however, since by Theorem 11 the cost is bounded from above. □

Observe again that, in general, P_∞ need not coincide with either P_s or P_o. And again we ask: To which nonnegative solution of (9) does $P(t, S, t_f)$ converge?

We have to take $S = D'D$ into account. Let $\mu_1, \mu_2, \ldots, \mu_q$, $q \geqq 0$, be those eigenvalues of M that are also the undetectable eigenvalues of (D, A), i.e.,

$$Az_i = \mu_i z_i ,$$
$$Dz_i = 0 ,$$
$$\operatorname{Re} \mu_i \geqq 0 , \quad i = 1, 2, \ldots, q .$$

Define

$$\mathscr{D} = \{\mu_1, \mu_2, \ldots, \mu_q\}$$

in accordance with (21).

Theorem 16. *If P_s exists then*

$$P_\infty = P_k$$

if and only if

$$\mathscr{C}_k = \mathscr{C} \cap \mathscr{D} .$$

Thus $P(t)$ can be made to approach any nonnegative solution P_k of (9) *by taking an appropriate S.*

Proof. To demonstrate the interaction of S and Q assume first that all undetectable eigenvalues λ_i of (C, A) belong to $\mathscr{D}$, where $\mathscr{D}$ corresponds to an arbitrary $S \geqq 0$. Then $Dz_i = 0$ and hence

$$\lim_{t \to -\infty} P(t, S, t_f) = \lim_{t \to -\infty} P(t, 0, t_f) = P_o$$

by Theorem 13. Note that P_o is generated by the set $\mathscr{C} = \mathscr{C} \cap \mathscr{D}$.

Second, let one undetectable eigenvalue λ_1 of (C, A) does not belong to $\mathscr{D}$. Then $Dz_1 \neq 0$ and

$$\lim_{t \to -\infty} P(t, S, t_f) = P_1$$

where P_1 is generated by the set $\mathscr{C}_1 = \mathscr{C} - \{\lambda_1\} = \mathscr{C} \cap \mathscr{D}$. It is so since otherwise the cost would be infinite due to the term $x'(t_f)\, D'D\, x(t_f)$.

Consequently, $P_\infty = P_k$ if and only if $\mathscr{C}_k = \mathscr{C} \cap \mathscr{D}$. □

Observe that P_∞ is the smallest nonnegative solution of the lattice generated by the set $\mathscr{C} \cap \mathscr{D}$ rather than by the $\mathscr{C}$ itself.

Theorem 17. *If* (A, B) *is stabilizable and* (C, A) *is detectable, then*

$$P_\infty = P_o = P_s$$

for any $S \geqq 0$.

Proof. By Theorem 5, equation (9) has the unique nonnegative solution $P_o = P_s$. Thus our claim is proved by referring to Theorem 16. □

From the computational viewpoint Theorem 16 renders it possible to find an S such that $P(t, S, t_f)$ will approach a desired equilibrium solution P_∞. In particular,

$$P_\infty = P_s \quad \text{if and only if} \quad \mathscr{C} \cap \mathscr{D} = \emptyset$$

and

$$P_\infty = P_o \quad \text{if and only if} \quad \mathscr{C} \cap \mathscr{D} = \mathscr{C}\,.$$

The relation

$$\lim_{t \to -\infty} P(t, S, t_f) = P_k \Leftrightarrow \mathscr{C} \cap \mathscr{D} = \mathscr{C}_k$$

is an equivalence and, therefore, it induces a decomposition of the cone of nonnegative matrices S into equivalence classes. Each class contains those S that make $P(t, S, t_f)$ converge towards a particular P_k. Thus we see that P_∞ is *not a continuous function* of S [26]. This is an observation of crucial importance in computations.

COMPUTATIONAL TECHNIQUES

This section is concerned with the computational aspects. Various techniques of finding a solution to equations (6) and (9) are reviewed. It is very difficult to label a particular method as being superior to the others. Each method may prove better in one application but it may fail in another. Our aim is to bring the most important methods to the reader's attention and indicate their applicability, advantages and objections.

Computing the solutions of the algebraic equation

(*i*) *Eigenvector Solution*

This method has been reported in [2], [10], [12], [23], [26], [27], [30] and its essentials are stated in Theorem 1. Apart from its theoretical importance, it is computationally promissing as efficient and accurate algorithms to compute eigenvalues of a matrix are now available.

This is the only method for computing any solution of (9), no matter whether it is nonnegative or not.

(ii) Iterative Solution

This technique has been first reported in [4], [20], [21] and is believed to be one of the best methods to find the stabilizing solution of (9). It is assumed that (9) possesses a unique nonnegative (stabilizing) solution P_s, which is being found by successive linear approximations.

Let P_j, $j = 0, 1, \ldots$ be the unique nonnegative solution of the linear algebraic equation

$$P_j K_j + K_j' P_j + C'C + L_j' L_j = 0 \tag{27}$$

where, recursively,

$$L_j = -B' P_{j-1}, \quad j = 1, 2, \ldots,$$
$$K_j = A + BL_j$$

and where L_0 is chosen such that the matrix $K_0 = A + BL_0$ is stable. Then

$$P_s \leqq P_{j+1} \leqq P_j \leqq \ldots, \quad j = 0, 1, \ldots,$$

and

$$\lim_{j \to \infty} P_j = P_s. \tag{28}$$

Indeed, K_0 being stable, (27) has a unique nonnegative solution P_0. This yields a K_1 and, in turn, a P_1. A little manipulation with the associated cost functionals reveals that $P_s \leqq P_1 \leqq P_0$. P_1 being bounded, K_1 is stable etc. A theorem on monotonic convergence of nonnegative matrices guarantees the existence of a limit. Then (27) is identical to (9) when $j \to \infty$ and (28) holds by uniqueness of P_s.

The method provides monotonic and quadratic convergence. In fact, it is an ingenious modification of Newton's method reported in [5].

(iii) Solution Using the Sign Function

This method is described in [31] and can be viewed as a simplification of (i) to find the stabilizing solution.

The matrix function sign Z is defined in [31] by

$$\operatorname{sign} Z = \lim_{k \to \infty} Z_{k+1},$$

where

$$Z_{k+1} = \tfrac{1}{2}(Z_k + Z_k^{-1}), \quad k = 0, 1, \ldots,$$
$$Z_0 = Z.$$

Define also

$$\operatorname{sign}^+ Z = \tfrac{1}{2}(I + \operatorname{sign} Z).$$

Then assuming $K = A - BB'P$ is stable and defining a matrix V by

$$KV + VK' + BB' = 0\,,$$

it is easy to check that

$$M = \begin{bmatrix} I, & -V \\ P, & I - PV \end{bmatrix} \begin{bmatrix} K, & 0 \\ 0, & -K' \end{bmatrix} \begin{bmatrix} I - VP, & V \\ -P, & I \end{bmatrix}.$$

Hence

$$\operatorname{sign}^+ M = \begin{bmatrix} I, & -V \\ P, & I - PV \end{bmatrix} \begin{bmatrix} 0, & 0 \\ 0, & I \end{bmatrix} \begin{bmatrix} I - VP, & V \\ -P, & I \end{bmatrix} =$$

$$= \begin{bmatrix} VP, & -V \\ -(I - PV)\,P, & I - PV \end{bmatrix}$$

and P follows immediately if V^{-1} exists.

Note that M is not taken through a complete spectral factorization, but into two parts only.

Computing the solution of the differential equation

(iv) Transition Matrix Solution

This old method was rediscovered in [8], [9], [15], [25]. It is important from the theoretical standpoint since it enables us to solve analytically equation (6) even in the time-varying case. However, analytic solutions can be found only in exceptional cases.

Denote

$$p(t) = P(t)\,x(t)\,. \tag{29}$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} x \\ p \end{bmatrix} = M \begin{bmatrix} x(t) \\ p(t) \end{bmatrix} \tag{30}$$

by virture of (9), and the transversality condition becomes

$$p(t_f) = S\,x(t_f)\,. \tag{31}$$

Let $\Phi(t, t_f)$ be the transition matrix for (30), i.e.,

$$\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} \Phi_{11}(t, t_f), & \Phi_{12}(t, t_f) \\ \Phi_{21}(t, t_f), & \Phi_{22}(t, t_f) \end{bmatrix} \begin{bmatrix} x(t_f) \\ p(t_f) \end{bmatrix} \tag{32}$$

where the $\Phi_{ij}(t, t_f)$ are $n \times n$ submatrices formed by partitioning Φ. Then (29), (31) and (32) gives

$$P(t, S, t_f) = [\Phi_{21}(t, t_f) + \Phi_{22}(t, t_f)\,S]\,[\Phi_{11}(t, t_f) + \Phi_{12}(t, t_f)\,S]^{-1}\,.$$

(v) Negative Exponential Solution

This technique, reported first in [34], bypasses one difficulty arising in (iv), namely, that $\Phi(t, t_f)$ contains stable as well as unstable modes. As time proceeds, the unstable modes tend to dominate and accuracy is at stake.

The basic trick is to recast the equations so that only negative exponentials occur in computations.

Write

$$W = \begin{bmatrix} W_{11}, & W_{12} \\ W_{21}, & W_{22} \end{bmatrix}$$

for a $2n \times 2n$ matrix that diagonalizes M,

$$M = W \begin{bmatrix} J, & 0 \\ 0, & -J \end{bmatrix} W^{-1},$$

and set

$$\begin{bmatrix} x \\ p \end{bmatrix} = W \begin{bmatrix} \hat{x} \\ \hat{p} \end{bmatrix}. \tag{33}$$

Then

$$\begin{bmatrix} \hat{x}(t_f) \\ \hat{p}(t) \end{bmatrix} = \begin{bmatrix} e^{-J(t_f - t)}, & 0 \\ 0, & e^{-J(t_f - t)} \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ \hat{p}(t_f) \end{bmatrix}$$

and by (33),

$$\hat{p}(t_f) = R\,\hat{x}(t_f)$$

where

$$R = -[W_{22} - SW_{12}]^{-1}[W_{21} - SW_{11}].$$

Denoting

$$G(t, t_f) = e^{-J(t_f - t)} R e^{-J(t_f - t)},$$

we finally get by (29)

$$P(t, S, t_f) = [W_{21} + W_{22}\,G(t, t_f)][W_{11} + W_{12}\,G(t, t_f)]^{-1}.$$

The steady-state solution P_∞ is obtained as

$$P_\infty = \lim_{t \to -\infty} P(t, S, t_f) = W_{21} W_{11}^{-1},$$

which is identical to equation (12).

(vi) Solution Via Numerical Integration

This is perhaps the most natural numerical technique to obtain the transient solution of (6). Usually the Runge-Kutta method is applied [18]. The method can also be applied to compute a desired steady-state solution as a function of $P(t_f) = S$, see Theorem 16. However, for this purpose the numerical integration is time consuming and not very accurate.

Table 1.

Nonnegative Solutions of Algebraic Equation (9)

	Re $\lambda \neq 0$ for all λ	Re $\lambda = 0$ for some λ
(A, B) stabilizable (C, A) detectable	P_s exists and it is the unique nonnegative solution, $P_s = P_o$	
(A, B) stabilizable (C, A) not detectable	P_s exists, there are other nonnegative solutions, $P_s \neq P_o$	P_s does not exist, there are other nonnegative solutions
(A, B) not stabilizable (C, A) detectable	No nonnegative solution exists including P_s and P_o	
(A, B) not stabilizable (C, A) not detectable	P_s does not exist, other nonnegative solutions may or may not exist	

Table 2.

Asymptotic Properties of Solutions to Riccati Equation (6)

	$t_f \to \infty$			t_f finite
	$S = 0$	$S \neq 0$		
		Re $\lambda \neq 0$ for all λ	Re $\lambda = 0$ for some λ	
(A, B) stabilizable (C, A) detectable	P_∞ exists, $P_\infty = P_o = P_s$			There is always a unique solution $P(t)$ and it is nonnegative
(A, B) stabilizable (C, A) not detectable	P_∞ exists, $P_\infty = P_o \neq P_s$	P_∞ exists, P_∞ = any nonnegative solution depending on S	?	
(A, B) not stabilizable (C, A) detectable	P_∞ does not exist			
(A, B) not stabilizable (C, A) not detectable	P_∞ may or may not exist			

60 To conclude this section we mention yet another group of methods to solve (9). The common feature is that equation (9) is transformed into a recurrent algebraic equation [14] whose solution coincides with the unique nonnegative solution of (9). The resulting matrix recurrent equation is then straightforward to solve.

CONCLUSIONS

The reader will have noticed that the interrelations of different nonnegative solutions of (9) are quite complicated. Letting $t_f \to \infty$ and studying the asymptotic behaviour of (6) makes the matters even worse. Therefore the following Tables 1 and 2 are provided to summarize the results for equations (9) and (6), respectively.

We also stress that the aim of the paper was to study the matrix Riccati equation in general and not to restrict ourselves to common engineering applications. This approach in turn provides a deep insight into the problem of least squares control and the engineer is well prepared to cope with ill-behaved solutions.

(Received June 5, 1972.)

REFERENCES

[1] M. Athans, P. L. Falb: Optimal Control. McGraw-Hill, New York 1966.
[2] R. W. Bass: Machine Solution of High-Order Matrix Riccati Equations. Douglas paper 4538.
[3] R. Bellman: Dynamic Programming. Princeton University Press, Princeton, N. Y. 1957.
[4] S. P. Bingulac, M. R. Stojić, N. Ćuk: On an Iterative Solution of Time-Invariant Riccati Equation. In: Prepr. JACC, St. Louis, Miss., 1971, 178—182.
[5] T. R. Blackburn: Solution of the Algebraic Matrix Equation Via Newton-Raphson Iteration. In: Prepr. JACC, Ann Arbor, Mich., 1968, 940—945.
[6] R. W. Brockett: Finite Dimensional Linear Systems. John Wiley, New York 1970.
[7] A. E. Bryson, Yu-Chi Ho: Applied Optimal Control. Bleisdell, Waltham, Mass. 1969.
[8] R. S. Bucy: Global Theory of the Riccati Equation. J. Comp. System Sci. *1* (December 1967), 4, 349—361.
[9] R. S. Bucy, P. D. Joseph: Filtering for Stochastic Processes with Applications to Guidance. Interscience, New York 1968.
[10] J. J. O'Donnell: Asymptotic Solution of the Matrix Riccati Equation of Optimal Control. In: Proc. 4th Ann. Allerton Conf. Circuit and Syst. Theory, Urbana, Ill., 1966, 577—586.
[11] S. E. Dreyfus: Dynamic Programming and the Calculus of Variations. Academic Press, New York 1965.
[12] A. F. Fath: Computational Aspects of the Linear Optimal Regular Problem. Trans. IEEE *AC-14* (October 1968), 547—550.
[13] M. L. J. Hautus: Stabilization, Controllability and Observability of Linear Autonomous Systems. Nederl. Akad. Wetensch., Proc. Ser. *A73* (1970), 448—455.
[14] K. L. Hitz: Relations Between Continuous-Time and Discrete-Time Quadratic Minimization. Ph. D. Thesis, The University of Newcastle, Australia, 1970.
[15] R. E. Kalman: Contributions to the Theory of Optimal Control. Bol. Soc. Mat. Max. *5* (1961), 102—119.
[16] R. E. Kalman: When Is a Linear Control System Optimal? Trans. ASME, J. Basic Engr. *86D* (March 1964), 51—60.

[17] R. E. Kalman, R. S. Bucy: New Results in Linear Prediction and Filtering Theory. Trans. ASME, J. Basic Engr. *83D* (1961), 95—100.
[18] R. E. Kalman, T. S. Englar: A Users Manual for the Automatic Synthesis Program. NASA Rept. CR-475, June 1966.
[19] R. E. Kalman, P. L. Falb, M. A. Arbib: Topics in Mathematical System Theory. McGraw-Hill, New York, 1969.
[20] D. L. Kleinman: On the Linear Regulator Problem and the Matrix Riccati Equation. MIT Electronic Systems Lab., Cambridge, Mass, Rept. ESL-R-271, June 1966.
[21] D. L. Kleinman: On an Iterative Technique for Riccati Equation Computations. Trans. IEEE *AC-13* (February 1968), 114—115.
[22] V. Kučera: A Contribution to Matrix Quadratic Equations. Trans. IEEE *AC-17* (June 1972), 344—347.
[23] V. Kučera: On Nonnegative Definite Solutions to Matrix Quadratic Equations. In: Proc. 5th IFAC Congress Vol. 4, Paris, 1972. Also Automatica 7 (July 1972), 413—423.
[24] V. Kučera: The Discrete Riccati Equation of Optimal Control. Kybernetika *8* (1972), 430—447.
[25] J. J. Levin: On the Matrix Riccati Equation. Trans. Am. Math. Soc. *10* (1959), 519—524.
[26] K. Mårtensson: On the Matrix Riccati Equation. Information Sci. *3* (1971), 1, 17—49.
[27] A. G. J. McFarlane: An Eigenvector Solution of the Linear Optimal Regulator Problem. J. Electron. Contr. *14* (June 1963).
[28] B. P. Molinari: The Stabilizing Solution of the Matrix Quadratic Equation. To appear in SIAM J. Control *10* (1972).
[29] L. S. Pontryagin at al.: The Mathematical Theory of Optimal Processes. Interscience, New York 1961.
[30] J. E. Potter: Matrix Quadratic Solutions. SIAM J. Appl. Math. *14* (May 1966), 3, 496—501.
[31] J. D. Roberts: Linear Model Reduction and Solution of the Algebraic Riccati Equation by Use of the Sign Function. To appear in Trans. IEEE *AC-17* (1972).
[32] A. P. Sage: Optimum Systems Control. Prentice-Hall, Englewood Cliffs, N. J. 1968.
[33] V. Strejc: State Space Synthesis of Discrete Linear Systems. Kybernetika *8* (1972), 83—113.
[34] D. R. Vaughan: A Negative Exponential Solution for the Matrix Riccati Equation. Trans. IEEE *AC-14* (February 1969), 72—75.
[35] W. M. Wonham: On Pole Assignment in Multi-Input Controllable Linear Systems. Trans. IEEE *AC-12* (December 1967), 660—665.
[36] W. M. Wonham: On a Matrix Riccati Equation of Stochastic Control. SIAM J. Control *6* (1968), 4, 681—698.

Ing. Vladimír Kučera, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Vyšehradská 49, 128 48 Praha 2. Czechoslovakia.